SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

EXISTENCE THEOREMS FOR HEREDITARY LAGRANGE AND MAYER PROBLEMS OF OPTIMAL CONTROL*

THOMAS S. ANGELL†

Abstract. In this paper, we prove existence theorems for optimal solutions in control systems governed by functional differential equations. We use a model for abstract hereditary systems, formulated by Hale and Cruz, which subsumes functional differential equations of finitely retarded type, equations of neutral type which are linear in $\dot{x}_t$, as well as a large class of Volterra integral equations. Using this model, we define an abstract hereditary control system, and then prove several existence theorems for optimal control problems of the types of Lagrange and Mayer in the line of previous work by Cesari.

1. Introduction. In a previous paper [2], we proved existence theorems for optimal control problems for systems whose dynamics are described by functional differential equations of finitely retarded type. Here, we present existence theorems for a much broader class of systems, namely the hereditary structures discussed by Hale and Cruz [16]. In this latter paper, the authors prove theorems of existence, uniqueness and continuous dependence of solutions for this general class of equations which includes, as special cases, functional differential equations of finitely retarded type, Volterra integral equations, difference equations, as well as those functional differential equations of neutral type in which the derivative $\dot{x}_t$ appears linearly. The control systems which we discuss here, will include, as special cases, those of [2] as well as a large class of neutral functional differential equations and a class of Volterra integral equations with kernels of the form $[K(t, s, x(s)) + F(s, u(s))]$.

Our specific goal is to extend the results of [2], in the line of previous work by Cesari, to this new and more general class of hereditary systems and to present new theorems for a class of hereditary Mayer problems. The principal results are the existence theorems, Theorem 6.2 relating to Lagrange problems of optimal control and Theorem 7.1 applying to problems in the form of Mayer.

We should also point out the relationship between the present results and the work of Warga [22, Chap. VII] pertaining to functional-integral equations. On the one hand, the theorems of [22] give existence results for optimal control including, as do the present results, the case of ordinary and retarded functional differential equations. In addition, [22] covers a much broader class of integral equations than do our results. On the other hand, there seems to be no natural way to include the class of neutral functional differential equations discussed here in the class of functional integral equations discussed in [22] without making additional differentiability assumptions which the approach of [16] is designed to avoid. Moreover, the approach used here utilizing closure and lower closure theorems differs from the approach in [22] which uses the concept of relaxed solutions.

The primary difficulty encountered in applying the techniques of [2] to the present class of hereditary systems is that the usual boundedness assumptions

* Received by the editors February 22, 1973, and in revised form November 11, 1974.

† Department of Mathematics, University of Delaware, Newark, Delaware 19711. This research was supported in part by a University of Delaware Research Foundation grant.

or growth conditions imposed on the right member of the differential equation do not alone yield equicontinuity of the minimizing sequence. As the reader will see in the sequel, this phenomenon is due to the appearance of terms involving the past history of the process in the left member of the equation.

Finally, we point out that since we are dealing with a functional differential system whose state space is a function space, the usual condition $|u| \leqq M$ is not enough to guarantee the various compactness conditions needed for the proof of the existence theorems.

2. Description of hereditary control systems and examples. Let r be a positive real number and consider the set $C([-r, 0], E^n)$, the class of all n-vector-valued continuous functions with domain $[-r, 0]$, equipped with the topology of uniform convergence. When there is no chance of confusion, we will write simply $C([-r, 0])$ for $C([-r, 0], E^n)$. Let A be a closed bounded subset of $E^1 \times C([-r, 0])$, let $U(t, \phi)$ be a closed subset of E^m, and let $g: A \to E^n$ be continuous. We will be concerned here with a differential equation of the form

$$d/dt[x(t) - g(t, x_t)] = F(t, x_t, u(t)), \tag{2.1}$$

where the function u, with values in E^m, is the control function x with values in E^n and defined on an interval of the form $[t_1 - r, t_2]$, is the trajectory, and for any $t \in [t_1, t_2]$, $x_t(\theta) = x(t + \theta)$, $-r \leqq \theta \leqq 0$. We shall refer to (2.1) as an hereditary control system with control function u.

Since the set A is assumed bounded, its projection onto the t-axis is contained in some compact interval $[t_1^*, t_2^*]$, and the initial value problem

$$d/dt[x(t) - g(t, x_t)] = h(t, x_t), \qquad t \in [t_1^*, t_2^*],$$
$$x_{t_1^*}(\theta) = \phi$$

is equivalent to the initial value problem posed by Hale and Cruz in [16] as can be verified easily by the reader. In particular, let $M = \{(t, \phi, u): (t, \phi) \in A$ and $u \in U(t, \phi)\}$ and let F_i, $i = 1, \cdots, n$, be given real-valued functions, defined and continuous on the set M. We write $F = (F_1, \cdots, F_n)$. If the function $u: [t_1, t_2] \to E^m$, $t_1^* \leqq t_1 \leqq t_2 \leqq t_2^*$, is chosen to be measurable, then (2.1) is an hereditary differential system of the type considered in [16, § 7]. The results of [16] show that under hypotheses, among which is that the function g is nonatomic at zero, the initial value problem will have a solution provided one specifies, at time t_1, a continuous function $\phi \in C([-r, 0])$. Essentially, the property of being nonatomic at zero insures that the function g does not depend very strongly on the value $\phi(0)$. The reader is referred to [16] for the precise definition as well as for examples of this behavior.

We now consider pairs of functions $\{x, u\}$, each pair consisting of a measurable function $u: [t_1, t_2] \to E^m$ and a corresponding continuous function $x: [t_1 - r, t_2] \to E^n$, which satisfies, almost everywhere, the hereditary differential equation

$$d/dt[x(t) - g(t, x_t)] = F(t, x_t, u(t)), \qquad t_1 \leqq t \leqq t_2, \tag{2.2a}$$

subject to boundary conditions

$$(t_1, x_{t_1}, t_2, x_{t_2}) \in B, \tag{2.2b}$$

where B is a given closed subset of $E^1 \times C([-r, 0]) \times E^1 \times C([-r, 0])$, as well as constraints

$$(t, x_t) \in A \quad \text{for all } t \in [t_1, t_2], \tag{2.2c}$$

$$u(t) \in U(t, x_t) \quad \text{for almost all } t \in [t_1, t_2]. \tag{2.2d}$$

Such pairs of functions are called admissible, and for such a pair, the function x is called a trajectory of the control system while the function u is called the control generating the trajectory x. Note that, implicit in the definition of a trajectory is the condition that the function $x(t) - g(t, x_t), t_1 \leqq t \leqq t_2$, is absolutely continuous so that $d/dt[x(t) - g(t, x_t)]$ exists almost everywhere in $[t_1, t_2]$.

A control system which admits at least one admissible pair is called a controllable system. In what follows, we will always assume that the system is controllable. Some remarks on the question of sufficient conditions for the controllability of systems of retarded type have been made elsewhere by the author [1].

Our purpose in this paper is to discuss existence theorems for optimization problems of the types of Mayer and Lagrange. In the first case, we assume as given a continuous function $h: B \to E^1$, and we seek the minimum of the functional $I[x, u] = h(t_1, x_{t_1}, t_2, x_{t_2})$ over some nonempty class Ω of admissible pairs $\{x, u\}$; that is, we seek a pair $\{x_0, u_0\} \in \Omega$ such that $I[x_0, u_0] \leqq I[x, u]$ for all $\{x, u\} \in \Omega$.

In the case of Lagrange problems, we wish to minimize an integral of the form

$$I[x, u] = \int_{t_1}^{t_2} F_0(t, x_t, u(t))\, dt$$

under the side conditions (2.2abcd). Here F_0 is a real-valued function defined on M and we shall denote by Ω any class of pairs $\{x, u\}$ admissible in the sense stated above and for which $F_0(t, x_t, u(t))$ is L-integrable in $[t_1, t_2]$. Briefly, we shall say that Ω is a class of admissible pairs for the Lagrange problem under consideration.

We shall refer to this latter problem as a Lagrange problem with unilateral constraints. If we take the sets $U(t, \phi)$ to be compact, the resulting problem may be referred to as a Pontryagin problem, and if F_0 is taken as $F_0 = 1$, the resulting problem is one of time-optimal control.

We now present three examples, similar to those given in [16], to show that the control systems described here include a number of control systems which have been studied previously.

Example 1. Setting the function $g(t, \phi) = 0$, equation (2.1) becomes

$$d/dt\, x(t) = F(t, x_t, u). \tag{2.3}$$

This is the type of equation we discussed in detail in our previous paper [2], a functional differential equation of finitely retarded type.

Example 2. Let the function g have sufficiently smooth derivatives. Then equation (2.1) becomes

$$\dot{x}(t) - g'_\phi(t, x_t)\dot{x}_t - g'_t(t, x_t) = F(t, x_t, u), \tag{2.4}$$

where $\dot{x}_t(\theta) = \dot{x}(t + \theta)$, $\theta \in [-r, 0]$. This equation includes equations of neutral type in which the derivative $\dot{x}_t$ occurs linearly. Such equations are discussed by Driver [14] and by Hale and Meyer [17].

Example 3. Let $K:[0, T] \times [0, T] \times E^n \to E^n$, $f:[0, T] \to E^n$, $0 < T < \infty$, be given functions. Assume that F is independent of ϕ and that the function g has the form

$$g(t, \phi) = \int_{-t}^{0} K(t, t + \theta, x(t + \theta))\, d\theta + f(t), \qquad t \in [0, T], \quad \phi \in C([-r, 0]).$$

Then, (2.1) becomes

$$\frac{d}{dt}\left[x(t) - \int_{-t}^{0} K(t, t + \theta, x(t + \theta))\, d\theta - f(t)\right] = F(t, u)$$

or

$$x(t) - \int_{-t}^{0} K(t, t + \theta, x(t + \theta))\, d\theta - f(t) = \int_{0}^{t} F(s, u(s))\, ds.$$

Making the change of variables $s = t + \theta$, the first integral becomes $\int_0^t K(t, s, x(s))\, ds$ and so, equation (2.1) becomes

$$x(t) = f(t) + \int_{0}^{t} [K(t, s, x(s)) + F(s, u(s))]\, ds, \tag{2.5}$$

which is a Volterra equation for x once a control function u has been specified. The initial value problem for (2.1) and the solution of (2.5) are equivalent problems provided that the initial functions of (2.1) satisfy the condition $\phi(0) = f(0)$. We refer the reader to [3] and [4] where, in the same spirit as the present paper, we treat directly systems governed by Volterra integral equations. In addition, the book of Warga [22] treats such models in a very general context.

Remark. We also wish to point out that certain problems involving hyperbolic partial differential equations have been shown to be equivalent to problems involving neutral functional differential equations of the form (2.4). The interested reader is referred to [5] for details.

3. The orientor field problem and property (Q). Control systems of the type described in the previous section can be written in terms of orientor field (or contingent) equations. In other words, we consider the orientor field problem

$$d/dt[x(t) - g(t, x_t)] \in Q(t, x_t), \qquad (t, x_t) \in A, \tag{3.1}$$

where $Q:A \to 2^{E^n}$ is given by

$$Q(t, \phi) = \{z = F(t, \phi, u) : u \in U(t, \phi)\}.$$

A solution $x(t)$ of (3.1) is a continuous function x, defined on an interval of the form $[t_1 - r, t_2]$ such that (a) $x(t) - g(t, x_t)$ is absolutely continuous on $[t_1, t_2]$; (b) $(t, x_t) \in A$ for all $t \in [t_1, t_2]$; and (c) $d/dt[x(t) - g(t, x_t)] \in Q(t, x_t)$ almost everywhere in $[t_1, t_2]$. In what follows, we will assume that the set $Q(t, \phi)$ is convex for each $(t, \phi) \in A$.

Clearly, any solution of the original control problem gives rise to a solution of this orientor field problem. The question of whether every solution of (3.1) which, in addition, satisfies the boundary condition (2.2b) can be viewed as a trajectory of the control system (2.2abcd) is answered by a standard argument

involving the McShane–Warfield extension of Filippov's implicit function lemma [19].

Remark. One remark needs to be made concerning the use of this result in the present context of hereditary systems. For control systems involving ordinary differential equations, the set M is a subset of a Euclidean space and is the union of countably many compact metrizable subsets. In our problem, the set M does not have this property, but it is a separable space. McShane and Warfield have shown that their theorem remains true in the separable case provided one is willing to invoke the continuum hypothesis. The reader is referred to [19] for details.

In what follows, it will be convenient to consider the space X of all continuous vector functions x defined on arbitrary intervals of the projection I_A of the set A onto the t-axis, $x:[a, b] \to E^n$. We now introduce the structure of a metric space on X by introducing, as usual, a metric function $\rho: X \times X \to E^{1+}$. For this purpose, let $x, y \in X$ with x defined on an interval $[a, b]$ and y defined on $[c, d]$. We may extend x and y to all of I_A by taking $x(t) = x(a)$ for all $t < a$, and $x(t) = x(b)$ for all $t > b$, and similarly for y. We then define the distance function $\rho(x, y)$ by

$$\rho(x, y) = |a - c| + |b + d| + \sup |x(t) - y(t)|,$$

where the supremum is taken over all $t \in I_A$. With this metric structure, the space $\{X, \rho\}$ is complete. When all functions x_n of a sequence $\{x_n\}$ are defined on a fixed interval, the convergence of the sequence to an element x in the metric topology is just the uniform convergence on that interval.

We now introduce the concept of a closed class of admissible pairs for the problems under consideration. Before doing so, we remind the reader that a trajectory x of the control system is a continuous vector function $x(t) = (x^1, \cdots, x^n)$, $t_1 \leqq t \leqq t_2$, such that x is generated by some measurable control function u, satisfying the constraints $u(t) \in U(t, x_t)$ almost everywhere, and such that the function $x(t) - g(t, x_t)$ is absolutely continuous in $[t_1, t_2]$.

For Mayer problems, we may introduce the following.

DEFINITION 3.1. A class Ω of admissible pairs is said to be closed provided, for every sequence $\{x_k, u_k\}$, $k = 1, 2, \cdots$, of pairs in Ω such that $x_k \to x$ in the ρ-metric, where x is a trajectory of the control system, among all measurable functions u which make the pair $\{x, u\}$ admissible, there exists one u such that $\{x, u\} \in \Omega$.

For Lagrange problems with functional $I[x, u] = \int_{t_1}^{t_2} F_0(t, x_t, u(t))\, dt$, we have a corresponding notion, differing slightly from the above as is to be expected since the notion of admissible pair is slightly different.

DEFINITION 3.2. A class Ω of admissible pairs for the Lagrange problem is said to be closed provided, for every sequence $\{x_k, u_k\}$, $k = 1, 2, \cdots$, of pairs in Ω such that $x_k \to x$ in the ρ-metric where x is a trajectory of the control system and for which $j = \lim I[x_k, u_k] < +\infty$, among all measurable functions u which make $\{x, u\}$ admissible and for which $I[x, u] \leqq j$, there exists one u such that $\{x, u\} \in \Omega$.

Clearly, the class of all admissible pairs is closed.

Given any point $(\bar{t}, \bar{\phi}) \in A$ and number $\delta > 0$, we denote by $N_\delta(\bar{t}, \bar{\phi})$ the set of all $(t, \phi) \in A$ such that $|t - \bar{t}| \leqq \delta$, $\|\phi - \bar{\phi}\| \leqq \delta$. Thus $N_\delta(\bar{t}, \bar{\phi})$ is a neighborhood, in the relative topology on A, of the element (t, ϕ).

Since the sets $Q(t, \phi)$ with which we will be dealing will be closed but not necessarily compact, we will need a concept of metric upper semicontinuity for set-valued mappings. To this end, we introduce the following definition which is a restatement of a definition introduced by Cesari [10].

DEFINITION 3.3. A set-valued function $Q: A \to 2^{E^n}$ is said to have property (Q) at a point $(\bar{t}, \bar{\phi}) \in A$ if

$$Q(\bar{t}, \bar{\phi}) = \bigcap_{\delta > 0} \operatorname{cl} \operatorname{co} \left[\bigcup_{(t,\phi) \in N_\delta(\bar{t}, \bar{\phi})} Q(t, \phi) \right]$$

$$= \bigcap_{\delta > 0} Q(t, \phi; \delta),$$

where we denote by $Q(t, \phi; \delta)$ the subset of E^n defined by

$$Q(t, \phi; \delta) = \operatorname{cl} \operatorname{co} \left[\bigcup_{(t,\phi) \in N_\delta(\bar{t}, \bar{\phi})} Q(t, \phi) \right].$$

The function Q is said to have the property (Q) with respect to (t, ϕ) in A if it has property (Q) with respect to (t, ϕ) at each point of A.

Property (Q) is a generalization of the more familiar concept of metric upper semicontinuity for closed and convex sets (see [10, p. 377]).

4. A closure theorem and conditions for lower closure. In this section, we formulate both a closure theorem and a theorem giving sufficient conditions for lower closure of functionals in integral form. As we shall see, the concept of lower closure is an extension of the concept of lower semicontinuity for free problems of the calculus of variations. The first theorem of this section is useful when we must deal with singular components and have no information concerning the convergence of the derivatives $d/dt[x^{(k)}(t) - g(t, x_t^{(k)})]$ along a minimizing sequence $\{x^{(k)}\}$. For the Lagrange problem, however, the presence of a growth condition involving F and F_0 will be enough to guarantee the weak convergence of this sequence of derivatives and it will be possible to establish the needed closure theorem assuming only a weakened form of property (Q), namely property (Q) with respect to ϕ only. This form of property (Q) has been used by Cesari [13], Olech [20], M. F. Bidaut [8] and Berkovitz [6] to establish existence theorems for Lagrange problems involving ordinary differential equations.

Let I be any interval of the real line and let $C(I, E^k)$ denote the set of all continuous functions mapping I into E^k. As in the previous sections, we will continue to write simply $C(I)$ for $C(I, E^n)$.

Denote by $y = (x^1, \cdots, x^s)$ the s-vector made up of components $x^1, \cdots, x^s$, $1 \leqq s \leqq n$, of the vector $x = (x^1, \cdots, x^n)$, and let z be the complementary $(n - s)$-vector $z = (x^{s+1}, \cdots, x^n)$. Thus, we may write $x = (y, z)$. Assume that the domain of F is contained in $E^1 \times C([-r, 0], E^s) \times E^m$ rather than in $E^1 \times C([-r, 0]) \times E^m$ and let A_0 be a subset of $E^1 \times C([-r, 0], E^s)$. Set $A = A_0 \times C([-r, 0], E^{n-s})$. We will assume that the orientor field in A depends only on t and y_t and not on z_t, and that $g(t, \phi) = (g_1, \cdots, g_n)$ has the form $g_{s+1} = g_{s+2} = \cdots = g_n = 0$. Then a solution to the orientor field problem

$$d/dt[x(t) - g(t, x_t)] \in Q(t, y_t), \qquad (t, x_t) \in A, \tag{4.1}$$

is an element $x \in C([t_1 - r, t_2])$ such that $x(t) - g(t, x_t)$ is absolutely continuous on $[t_1, t_2]$, $x(t) = (y(t), z(t))$ with $(t, y_t) \in A_0$ (and hence $(t, x_t) \in A$) for every $t \in [t_1, t_2]$, and $d/dt[x(t) - g(t, x_t)] \in Q(t, y_t)$ almost everywhere in $[t_1, t_2]$.

THEOREM 4.1 (Closure Theorem). *Denote by A_0 a closed subset of $E^1 \times C([-r, 0], E^s)$ and let $A = A_0 \times C([-r, 0], E^{n-s})$. Let $Q: A_0 \to 2^{E^n}$, and assume that the map Q satisfies the property* (Q) *with respect to (t, ϕ).*

Suppose that $\{x^{(k)}\}$ is a sequence of solutions to the orientor field problem (4.1), *$x^{(k)} = (y^{(k)}, z^{(k)})$ defined on $[t_{1_k} - r, t_{2_k}]$, and such that* (i) *the $y^{(k)}$ converge in the ρ-metric to a continuous function y defined on $[t_1 - r, t_2]$ with the property that $x_i(t) - g_i(t, x_t)$, $i = 1, \cdots, \alpha$, is absolutely continuous*; (ii) *the $z^{(k)}$ converge pointwise to a function z almost everywhere on $[t_1 - r, t_2]$ and z admits a decomposition $z = Z + S$ where Z is absolutely continuous on $[t_1, t_2]$ and $S' = 0$ almost everywhere on $[t_1, t_2]$*; *and* (iii) *the function $g(t, \phi) = (g_1, \cdots, g_n)$ has the form $g_{s+1} = g_{s+2} = \cdots = g_n = 0$. Then the continuous vector function $X = (y, Z)$, defined on $[t_1 - r, t_2]$ is a solution of* (4.1).

The proof of Theorem 4.1 follows closely the proof of Theorem 3.1 in [2] which treats the retarded case. We refer the reader to that paper for details and mention here only that the proof involves the use of the Ascoli theorem on components governed by a growth condition and Helly's selection process on the other "singular components". The appearance of such singular components is typical in Mayer problems (see § 7). References [6] and [13] show that, under hypotheses, the use of this closure theorem can be avoided, and the weakened form of property (Q) used.

Specifically, in the case that some condition guarantees the weak convergence of the derivatives of a minimizing sequence, e.g., in the case that the functions F and F_0 are related by the growth condition to be described later (see Def. 6.1), it is possible to establish a lower closure theorem under the following weakened version of property (Q) (see Cesari [13]).

DEFINITION 4.1. *A set-valued function $Q: A \to 2^{E^n}$ is said to have property* (Q) *with respect to ϕ at a point $(\bar{t}, \bar{\phi}) \in A$ provided*

$$Q(\bar{t}, \bar{\phi}) = \bigcap_{\delta > 0} \operatorname{cl} \operatorname{co} \left[\bigcup_{(\bar{t}, \phi) \in N_{\delta, \bar{t}}(\bar{\phi})} Q(\bar{t}, \phi) \right],$$

where $N_{\delta, \bar{t}}(\bar{\phi}) = \{(\bar{t}, \phi) \in A : \|\phi - \bar{\phi}\| \leqq \delta\}$.

Note that this definition differs from Definition 3.3 only in that $N_{\delta, \bar{t}}(\bar{\phi})$ replaces $N_\delta(\bar{t}, \bar{\phi})$.

We now introduce the concept of lower closure (see [9], [12]) for functionals in integral form. This concept reduces to the familiar concept of lower semicontinuity for free problems.

DEFINITION 4.2. Let $x \in C([t_1 - r, t_2])$ be such that the function $[x(t) - g(t, x_t)]$ is absolutely continuous on $[t_1, t_2]$ and $(t, x_t) \in A$ for all $t \in [t_1, t_2]$. A functional I has the property of lower closure at x if, for any sequence $\{x^{(k)}, u^{(k)}\}$, $k = 1, 2, \cdots$, of admissible pairs such that $x^{(k)}$ converges pointwise almost everywhere in I_A to x, such that the derivatives $d/dt[x^{(k)}(t) - g(t, x_t^{(k)})]$ converge weakly in L_1 to $d/dt[x(t) - g(t, x_t)]$, and such that $\underline{\lim}\, I[x^{(k)}, u^{(k)}] < +\infty$ as $k \to \infty$, there is some

measurable function $u:[t_1, t_2] \to E^m$ for which $\{x, u\}$ is admissible and

$$I[x, u] \leqq \underline{\lim}\, I[x^{(k)}, u^{(k)}]. \tag{4.2}$$

We remark that, if M is closed and if the sets $Q(t, \phi)$ are closed and convex, then the closure theorem and the McShane–Warfield lemma guarantee the existence of some measurable function u such that the pair $\{x, u\}$ is admissible. However, it is possible that (4.2) does not hold for the pair $\{x, u\}$ (see [9, p. 90]).

In order to give sufficient condition for lower closure, we must introduce a new map $\tilde{\tilde{Q}}:A \to 2^{E^{n+1}}$ given by

$$\tilde{\tilde{Q}}(t, \phi) = \{\tilde{z} = (z^0, z) \in E^{n+1} : z^0 \geqq F_0(t, \phi, u), z = F(t, \phi, u), u \in U(t, \phi)\}.$$

We may now state the following theorem on lower closure which is the analogue for Lagrange problems of the preceding closure theorem for Mayer problems. It is a corollary of the more general lower closure theorem, Theorem 5.1, of Cesari [13].

THEOREM 4.2. *Let A be a closed subset of $E^1 \times C([-r, 0])$, and for every $(t, \phi) \in A$, suppose that $U(t, \phi) \subset E^m$. Assume that the set $M = \{(t, \phi, u) : (t, \phi) \in A, u \in U(t, \phi)\}$ is closed and let $\tilde{F}(t, \phi, u) = (F_0, F_1, \cdots, F_n) = (F_0, F)$ be continuous on M. Assume that the sets $\tilde{\tilde{Q}}(t, \phi)$ are closed and convex and that $\tilde{\tilde{Q}}$ satisfies property* (Q) *with respect to ϕ in A. Also, suppose that the set $B \subset E^1 \times C([-r, 0]) \times E^1 \times C([-r, 0])$ is closed and that, for some locally integrable function ψ, we have $F_0(t, \phi, u) \geqq \psi(t)$ for all $(t, \phi, u) \in M$. Then the integral*

$$I[x, u] = \int_{t_1}^{t_2} F_0(t, x_t, u(t))\, dt$$

has the property of lower closure at every $x \in C([t_1 - r, t_2])$ such that the function $x(t) - g(t, x_t)$ is absolutely continuous on $[t_1, t_2]$ and $(t, x_t) \in A$ for all $t \in [t_1, t_2]$.

We will not repeat the proof here but refer the reader to [13]. The proof proceeds by applying the Banach–Saks–Mazur theorem to the weakly convergent sequence of derivatives to obtain a new sequence of convex combinations which converges strongly to $d/dt[x(t) - g(t, x_t)]$. Property (Q) is then invoked to prove that this latter derivative satisfies the orientor field relation (4.1). This technique was used by Cesari in [13], by Berkovitz in [6] and more recently in [7] which treats multidimensional problems. We emphasize that, for the present theorem, we need only property (Q) with respect to ϕ and not with respect to (t, ϕ). This weakening is not in general possible under the sole assumption that the sequence $\{x^{(k)}\}$ converges to x uniformly. Indeed, uniform convergence of the $x^{(k)}$ to x does not imply weak convergence of the derivatives $x^{(k)'}$ to x' in L_1, and there are examples (see [13]) which show that property (Q) with respect to both variables is essential in this case.

5. General remarks. Concerning the proof of existence theorems for optimal control, we wish to make the following preliminary remarks.

As was shown in our previous paper [2], it is convenient to require the compactness of the initial data, and this restriction will be in force throughout this paper also. Indeed, the growth condition is enough to insure that the functions f_n generated by a minimizing sequence, are equiabsolutely continuous on the

intervals $[t_{1_n}, t_{2_n}]$, but gives no information on the behavior of the functions on sets $[t_{1_n} - r, t_{1_n}]$. Even for the retarded case, it is possible to give examples of equations for which a sequence of solutions may converge on the intervals $[t_{1_n}, t_{2_n}]$ but which diverge on the intervals $[t_{1_n} - r, t_{1_n}]$. We recall here an example of [15, p. 41] of a functional differential equation of finitely retarded type whose trajectories after a suitable time are zero regardless of the initial function in the unit ball of $C([-1, 0])$.

Example. Consider the equation

$$\dot{x}(t) = -\psi(t)x(t-1), \qquad t \geqq 0,$$

where

$$\psi(t) = \begin{cases} 2\sin^2 \pi t, & t \in [2n, 2n+1], \quad n = 1, 2, \cdots, \\ 0, & t \in (2n-1, 2n), \quad n = 1, 2, \cdots. \end{cases}$$

We show that for $t \geqq 3$, $x(t) = 0$. In fact, for $t \in [1, 2]$, $x(t) = x(1)$, and

$$\dot{x}(t) = -\psi(t)x(1), \qquad t \in [2, 3].$$

Thus

$$x(3) = x(1)\left[1 - 2\int_2^3 \sin^2 \pi s \, ds\right] = 0$$

and so, $x(t) = 0$ for $t \in [3, 4]$ and indeed, $x(t) = 0$ for $t \geqq 3$. For $t \in [1, 2]$ the equation is just $\dot{x}(t) = 0$ and the solution corresponding to the initial function ϕ is $x(t) = \phi(1)$, $t \in [1, 2]$. Since the initial functions are taken to lie in the unit ball of $C([-1, 0])$, the trajectories on $[1, 2]$ form an equicontinuous equibounded family. Moreover, since on $[2, 3]$ we have $|\dot{x}(t)| \leqq 2$, the trajectories on $[1, 3]$ are equicontinuous and equibounded.

Hence, the trajectories restricted to $[1, +\infty)$ are compact while the initial conditions are not.

It may, however, be the case that for some classes of equations, the compactness of the initial data guarantee the compactness of the trajectories. We discuss this possibility at the end of this section.

Moreover, as usual in direct methods of the calculus of variations, we shall need conditions guranteeing that, from a minimizing sequence $\{x^{(n)}\}$ of admissible trajectories, we can extract a subsequence $\{x^{(n_k)}\}$ which converges, in some suitable topology, to an element which actually gives the minimum of the functional I. For example, in Lagrange problems, it will be natural to assume first a growth condition involving F and F_0. This will immediately guarantee that the sequence of absolutely continuous functions $f_n(t) = x^{(n)}(t) - g(t, x_t^{(n)})$ is equiabsolutely continuous and thus that we can extract from it a subsequence $\{f_{n_k}\}$ whose elements converge uniformly (that is, converge in the ρ-metric) to an absolutely continuous function f as $k \to \infty$ (see the proof of Theorem 6.2 below). This will, of course, affect the behavior of the corresponding sequence $\{x^{(n_k)}\}$.

In particular, we need to show that this latter sequence, or at least a subsequence, will converge in the ρ-metric to a trajectory of the system. Since we have assumed that the set of initial conditions is compact and (§ 2) that the function g

is nonatomic at zero, any condition guaranteeing the convergence of the f_n, e.g., the growth conditions mentioned above, is sufficient to guarantee the convergence of a minimizing sequence. The requirement that g be nonatomic at zero implies that the nonlinear operator defined by g is a contraction on a suitably chosen closed set of continuous functions. Convergence of the trajectories in the ρ-metric to a trajectory of the system is established by straightforward modification of the arguments establishing continuous dependence in [16, § 6] (see in particular Theorem 6.4 of [16]).

Finally, as mentioned earlier (Introduction), the usual condition $|u(t)| \leqq M$ is not by itself enough to guarantee the compactness of the minimizing sequence. We point out that in some specific cases this condition may suffice. For systems described by integral equations of the form (2.5) we can see that the set of trajectories is, in fact, a relatively compact set under suitable conditions on the functions K and f. To be more precise, suppose f is continuous on the interval $[0, T]$, $0 < T < +\infty$, and suppose that, for fixed $(t, z) \in [0, T] \times E^n$, $K(t, s, z)$ is measurable in $0 \leqq s \leqq t$, and that, for fixed $S \in [0, T]$, K is continuous for $(t, z) \in [s, T] \times E^n$. Suppose further, that there exists a function ϕ, L-integrable in $[0, T]$, such that $|K(t, s, z| \leqq \phi(s)$, $0 \leqq s \leqq t \leqq T$. These conditions are sufficient to guarantee the existence of a continuous solution on $[0, T]$ of the Volterra integral equation

$$x(t) = f(t) + \int_0^t K(t, s, x(s))\, ds, \qquad 0 \leqq t \leqq T,$$

(see [21, pp. 23–24]).

If, in addition to the above hypotheses, the function K satisfies a condition of the form

$$|K(t, s, x) - K(t^*, s, x)| \leqq \psi(s)|t - t^*|,$$

where $\psi(s)$ is L-integrable on $[0, T]$, then the set of trajectories of (2.5) form a relatively compact set. To see this, note that the integral equation (2.5) may be rewritten in the form

$$x(t) = H_u(t) + \int_0^t K(t, s, x(s))\, ds, \tag{5.1}$$

$$H_u(t) = f(t) + \int_0^t F(s, u(s))\, ds, \tag{5.2}$$

where u is an admissible control. In fact, the set of functions $\{H_u\}$ defined by (5.2) is a relatively compact set of functions; both equiboundedness and equicontinuity follow from the boundedness of the functions F and f as may be checked easily. As previously remarked, the conditions imposed on f and K are sufficient to insure the existence of a solution (5.1) for each given admissible control u. We denote the generated trajectory by x_u. The equiboundedness and equicontinuity of the set $\{x_u\}$ now follow easily using standard arguments and the properties of K and equicontinuity of the H_u.

Moreover, it is possible that, for some classes of equations, the compactness of the initial data guarantees the compactness of the trajectories. For example, Banks and Kent [5] show that, when the function g has the form $\int_{t_0-r}^{t_0} d_s\mu(t, s)x(s)$, then suitable conditions on μ and the assumption that F is linear in ϕ are sufficient to guarantee compactness of the set of trajectories given that the set of initial

conditions is compact. For precise statements, the interested reader may refer to [5].

6. Growth conditions and existence theorems for Lagrange problems. For the proof of the existence theorems for Lagrange problems, we will need, as mentioned previously (§ 5), conditions which will insure that a minimizing sequence of trajectories will contain a convergent subsequence. For this purpose, we will use a growth condition described in [12]. We remark that this condition will only insure the equiabsolute continuity of the functions $f_n(t) = x^{(n)}(t) - g(t, x_t^{(n)})$ generated by a minimizing sequence $\{x^{(n)}, u^{(n)}\}$.

DEFINITION 6.1. The function $F = (F_0, F_1, \cdots, F_n)$ is said to satisfy the growth condition (γ) if, for any $\varepsilon > 0$, there exists a nonnegative L-integrable function ψ_ε such that

$$(\gamma) \qquad |F(t, \phi, u)| \leqq \psi_\varepsilon(t) + \varepsilon F_0(t, \phi, u)$$

for all $(t, \phi, u) \in M$.

THEOREM 6.1. *Assume that the set A is closed and bounded in $E^1 \times C([-r, 0])$. Let Ω be the class of all admissible pairs $\{x, u\}$, x defined on $[t_1 - r, t_2]$, u defined on $[t_1, t_2]$, with $I[x, u] \leqq K$, $K > 0$ a fixed constant. Assume that $\tilde{F} = (F_0, F_1, \cdots, F_n)$ satisfies (γ). Then the set*

$$H = \{f : f(t) = x(t) - g(t, x_t), \{x, u\} \in \Omega\}$$

is an equiabsolutely continuous set of functions.

Proof. Note that, for $\varepsilon = 1$ we have $0 \leqq \psi_1(t) + F_0(t, \phi, u)$ for all $(t, \phi, u) \in M$. Now the set A is bounded so that its projection into E^1, I_A, is contained in some finite interval $[t_1^*, t_2^*]$. Thus

$$\int_{t_1}^{t_2} \psi_1(t)\, dt \leqq \int_{t_1^*}^{t_2^*} \psi_1(t)\, dt = K_1, \qquad 0 \leqq K_1 < +\infty.$$

Let $\varepsilon > 0$ be given and choose $\sigma = \frac{1}{2}\varepsilon(K + K_1)^{-1}$. Then for every measurable subset E of $[t_1^*, t_2^*]$ we have

$$\begin{aligned}
\int_E \left| \frac{d}{dt} f(t) \right| dt &= \int_E \left| \frac{d}{dt}[x(t) - g(t, x_t)] \right| dt \\
&= \int_E |F(t, x_t, u(t))|\, dt \leqq \int_E [\psi_\sigma(t) + \sigma F_0(t, x_t, u(t))]\, dt \\
&\leqq \int_E \psi_\sigma(t)\, dt + \sigma \int_E [F_0(t, x_t, u(t)) + \psi_1(t)]\, dt \\
&\leqq \int_E \psi_\sigma(t)\, dt + \sigma \int_{t_1}^{t_2} F_0(t, x_t, u(t))\, dt + \sigma \int_{t_1^*}^{t_2^*} \psi_1(t)\, dt \\
&\leqq \int_E \psi_\sigma(t)\, dt + \sigma(K + K_1) \leqq \int_E \psi_\sigma(t)\, dt + \varepsilon/2.
\end{aligned}$$

By integrability of the ψ_σ, we conclude that there is a $\delta > 0$ such that, if meas $(E) < \delta$, then $\int_E \psi_\sigma(t)\,dt < \varepsilon/2$. For this choice of E, we then have

$$\int_E \left|\frac{d}{dt} f(t)\right| dt \leqq \int_E \psi_\sigma(t)\,dt + \varepsilon/2 \leqq \varepsilon,$$

and hence the functions $d/dt\, f(t)$, for $f \in H$, are equiabsolutely integrable from which it follows that H is an equiabsolutely continuous set of functions.

We may now present the following existence theorem for Lagrange problems.

THEOREM 6.2. *Let $A \subset E^1 \times C([-r, 0])$ be closed and bounded, and let B be a closed subset of $E^1 \times C([-r, 0]) \times E^1 \times C([-r, 0])$. Let $U : A \to 2^{E^m}$ be given, and define a set $M \subset E^1 \times C([-r, 0]) \times E^m$ by $M = \{(t, \phi, u) : (t, \phi) \in A,\ u \in U(t, \phi)\}$. Let us assume that* (i) *$\tilde{F} = (F_0, F_1, \cdots, F_n)$ is continuous on M and satisfies condition (γ)*; (ii) *the set $\tilde{\tilde{Q}}(t, \phi)$ is closed, convex and $\tilde{\tilde{Q}}$ satisfies property* (Q) *with respect to ϕ in A*; (iii) *the set M is closed*; (iv) *the projection of the set B into its second coordinate space is contained in a compact subset of $C([-r, 0])$*; (v) *the map g is nonatomic at zero and $g[A]$ is bounded.*

Then the cost functional

$$I[x, u] = \int_{t_1}^{t_2} F_0(t, x_t, u(t))\,dt$$

has an absolute minimum in any nonempty closed class Ω of admissible pairs.

Proof. For $\varepsilon = 1$, we have that $0 \leqq \psi_1(t) + F_0(t, \phi, u)$ or $-\psi_1(t) \leqq F_0(t, \phi, u)$ for all $(t, \phi, u) \in M$. Now the set A is bounded so that I_A is contained in some finite interval $[t_1^*, t_2^*]$ and $[t_1, t_2] \subset [t_1^*, t_2^*]$. Thus

$$-\int_{t_1}^{t_2} \psi_1(t)\,dt \geqq -\int_{t_1^*}^{t_2^*} \psi_1(t)\,dt$$

and so, for every pair $\{x, u\} \in \Omega$,

$$I[x, u] = \int_{t_1}^{t_2} F_0(t, x_t, u(t))\,dt \geqq -\int_{t_1^*}^{t_2^*} \psi_1(t)\,dt = -K_1 > -\infty.$$

So, if $i = \inf I[x, u]$, then $i > -\infty$ and there exists a sequence $\{x^{(k)}, u^{(k)}\}$ in Ω such that $I[x^{(k)}, u^{(k)}] \to i$ as $k \to \infty$, each $x^{(k)}$ being defined on $[t_{1_k} - r, t_{2_k}]$ and such that the functions $x^{(k)}(t) - g(t, x_t^{(k)})$ are absolutely continuous on $[t_{1_k}, t_{2_k}]$. Furthermore, we may assume that

$$i \leqq I[x^{(k)}, u^{(k)}] = \int_{t_{1_k}}^{t_{2_k}} F_0(t, x_t^{(k)}, u^{(k)}(t))\,dt \leqq i + 1/k \leqq i + 1.$$

By Theorem 6.1, the set $H = \{f_k : f_k(t) = x^{(k)}(t) - g(t, x_t^{(k)})\}$ is an equicontinuous family, and, since A is bounded and $g[A]$ is bounded, it is easy to see that the set H is equibounded. Hence, we may use Ascoli's selection theorem to extract a subsequence, which we again call $\{f_n\}$ and which converges in the ρ-metric to a function $f \in C([t_1, t_2])$, which is absolutely continuous on $[t_1, t_2]$. Moreover, since the set of initial conditions is assumed to be compact, we may assume that the extraction has been performed in such a way that the initial functions $x_{t_{1_k}}$ converge uniformly to some function $\phi_0 \in C[-r, 0]$. Thus, as indicated above

(§ 5), there is some function $x_0 \in C([t_1 - r, t_2])$ such that at least a subsequence $x^{(n_k)} \to x_0$ in the ρ-metric.

As may be seen immediately from the proof of Theorem 6.1, the growth condition (γ) insures that the derivatives $d/dt[x^{(k)}(t) - g(t, x_t^{(k)})]$ are equiabsolutely integrable. Since A is bounded, the set I_A is bounded and hence, by a theorem of Dunford and Pettis, the set of derivatives is weakly compact in $L_1(I_A)$. Hence, at least a subsequence of derivatives converges weakly in $L_1(I_A)$ to a function $\psi(t) \in L_1(I_A)$. Since the functions f_k converge uniformly on I_A to f, certainly they converge pointwise almost everywhere to f and hence, by the absolute continuity of these functions, we have $f'(t) = \psi(t)$ almost everywhere.

The class Ω being closed, the lower closure theorem guarantees the existence of a measurable function u_0, defined on $[t_1, t_2]$, such that $\{x_0, u_0\} \in \Omega$, and

$$I[x_0, u_0] \leqq \underline{\lim}\, I[x^{(k)}, u^{(k)}] = i.$$

Since $\{x_0, u_0\} \in \Omega$, we must also have $I[x_0, u_0] \geqq i$ and so $I[x_0, u_0] = i$ and the proof is complete.

We remark that, as is well-known in nonhereditary problems, if we assume that the class Ω of admissible trajectories satisfies an L_p boundedness condition of the form

$$\int_{t_1}^{t_2} \left| \frac{d}{dt}[x(t) - g(t, x_t)] \right|^p dt \leqq N \tag{6.1}$$

for $1 < p < +\infty$, $0 \leqq N < +\infty$, we may replace the growth condition (γ) with a weaker condition involving the function F_0 only. More precisely, we may assume that there is an L-integrable function ψ such that $F_0(t, \phi, u) \geqq \psi(t)$ for all $(t, \phi, u) \in M$. In this case, the condition (γ) insures the equiabsolute continuity of the set of functions H (see Theorem 6.1). In fact, a simple application of Hölder's inequality shows that the derivatives are equiabsolutely integrable which implies, as before, the equiabsolute continuity of the set H.

The more familiar L^p boundedness condition, that is a condition on the trajectories themselves, rather than on the functions $[x(t) - g(t, x_t)]$, of the form

$$\int_{t_1}^{t_2} \left| \frac{dx}{dt} \right|^p dt \leqq N \tag{6.2}$$

for some constant $N > 0$ and integer $p > 1$, can be utilized only when we have additional information concerning the smoothness of trajectories of the hereditary system and on the function g. Specifically, we have the following.

THEOREM 6.3. *Let A be a closed bounded subset of $E^1 \times C[-r, 0]$ and assume* (i) *$\tilde{F} = (F_0, F_1, \cdots, F_n)$ is continuous on M and that there is an L-integrable function ψ such that $F_0(t, \phi, u) \geqq \psi(t)$ for all $(t, \phi, u) \in M$*; (ii) *for each $(t, \phi) \in A$, $\tilde{\tilde{Q}}(t, \phi)$ is closed and convex and $\tilde{\tilde{Q}}$ satisfies property* (Q) *with respect to ϕ in A*; (iii) *the set M is closed and the projection of the set B into its second coordinate space is a set of absolutely continuous functions which is compact in $C([-r, 0])$*; (iv) *the function g is Lipschitzian in both arguments.*

Let Ω be a nonempty closed class of admissible pairs for the Lagrange problem such that condition (6.2) *holds.*

Then there exists a pair $\{x_0, u_0\} \in \Omega$ *such that*

$$I[x, u] = \int_{t_1}^{t_2} F_0(t, x_t, u(t))\, dt$$

takes on its minimum in Ω *at* $\{x_0, u_0\}$.

Proof. From condition (i), it follows, as in the proof of Theorem 6.2, that $i = \inf I[x, u]$ is finite. Let $\{x^{(k)}, u^{(k)}\}$ be a minimizing sequence, each $u^{(k)}$ being defined on an interval $[t_{1_k}, t_{2_k}]$ and the corresponding $x^{(k)}$ being defined on $[t_{1_k} - r, t_{2_k}]$. We may assume that $i \leqq I[x^{(k)}, u^{(k)}] \leqq i + 1/k < i + 1$, and that

$$\int_{t_{1_k}}^{t_{2_k}} \left| \frac{dx^{(k)}}{dt} \right|^p dt \leqq N.$$

By the weak compactness of the unit ball in L_p, $1 < p < \infty$, we may conclude that there exists a subsequence of the minimizing sequence and some continuous function x, defined on $[t_1, t_2]$, such that $dx_k/dt \to dx/dt$ weakly and $x_k \to x$ uniformly, i.e., in the ρ-metric. Condition (iv) is enough to insure that the function $g(t, x_t)$ is absolutely continuous and, consequently, that the derivative $d/dt[x(t) - g(t, x_t)]$ exists almost everywhere in $[t_1, t_2]$.

This shows that the functions $x_k(t) - g(t, x_{kt})$ converge pointwise almost everywhere to the absolutely continuous function $x(t) - g(t, x_t)$. The remainder of the proof proceeds exactly as that of Theorem 6.2.

The restriction on the set of initial conditions, namely that all the initial functions be absolutely continuous, is necessary in the sense that, even if the map g satisfies the required Lipschitz conditions, the functions $x(t) - g(t, x_t)$ need not be absolutely continuous. In particular, note that if $g(t, x_t) = x(t - 1)$, $(x_t)(\theta) = x(t + \theta)$, $\theta \in [-1, 0]$, and we have fixed initial data $\phi(t)$, $-1 \leqq t \leqq 0$, which is continuous but nowhere differentiable, then the absolute continuity of $x(t)$ on $[t_1, t_2]$ does not insure the differentiability of $x(t) - x(t - 1)$ on the same interval. Indeed, it is examples of this sort which served Hale and co-workers as motivation for introducing the present hereditary model and thus avoiding the necessity of discussing the smoothness of the trajectories x.

7. An existence theorem for hereditary Mayer problems. We now present an existence theorem for Mayer type problems which is an extension of theorems of McShane [18] and Cesari [11] involving comparison functions. We point out that the restriction on the components of the function g, necessitated by the use of Theorem 4.1, has the effect of requiring all components not governed by a growth condition to satisfy a retarded equation. This condition is, as we have seen in § 6, automatic in the situation of the Mayer problem which arises from reformulation of a problem of Lagrange. Moreover, it does not seem possible to avoid the use of Theorem 4.1 and the assumption that the sets $Q(t, x)$ satisfy property (Q) due to the presence of the singular components. (See also the remarks after Theorem 4.1.)

THEOREM 7.1. *Let* α, n, $0 \leqq \alpha \leqq n$, *be given integers and for* $x = (x^1, \cdots, x^n)$, *let* $y = (x^1, \cdots, x^\alpha)$, $z = (x^{\alpha+1}, \cdots, x^n)$ *so that* $x = (y, z)$. *Let* A_0 *be a closed bounded subset of* $E^1 \times C([-r, 0], E^\alpha)$ *and let* S *be a closed bounded sphere in*

$C([-r, 0], E^{n-\alpha})$ *so that* $A = A_0 \times S$ *is closed and bounded in* $C([-r, 0], E^n)$. *Let* $(t, \psi) \in A_0$, *let* $U(t, \psi)$ *be a closed subset of* E^m, *let* $M_0 = \{(t, \psi, u) | (t, \psi) \in A_0, u \in U(t, \psi)\}$ *and let* $M = M_0 \times S = \{(t, \phi, u) | (t, \phi) \in A, \quad u \in U(t, \psi), \phi(\theta) = (\psi(\theta), \chi(\theta))\}$. *Let* $F = (F_1, \cdots, F_n)$ *and* H *be functions defined and continuous on* M_0 *with* $F_{\alpha+1}, \cdots, F_n$ *and* H *nonnegative.*

Assume for every $i = 1, 2, \cdots, \alpha$, *that the following growth conditions holds*:

(γ_i) *given* $\varepsilon > 0$ *there is a locally integrable function* $\xi_{i\varepsilon}(t) \geqq 0$ *such that* $|F_i(t, \psi, u)| \leqq \xi_{i\varepsilon}(t) + \varepsilon H(t, \psi, u)$ *for all* $(t, \psi, u) \in M_0$.

For every $(t, \psi) \in A_0$, *let* $Q_H(t, \psi) \subset E^{n+1}$ *be defined by*

$$Q_H(t, \phi) = \{\tilde{z} = (z^0, z^1, \cdots, z^n) | z^0 \geqq H(t, \psi, u), z^i = F_i(t, \psi, u), \quad i = 1, \cdots, \alpha, z^i \geqq F_i(t, \psi, u), \quad i = \alpha + 1, \cdots, n, u \in U(t, \psi)\}$$

and assume that $Q_H(t, \psi)$ *is convex for each* $(t, \psi) \in A_0$ *and satisfies property* (Q) *in* A_0. *Since the functions* F *and* H *do not depend on the components* $\phi^{\alpha+1}, \cdots, \phi^n$, *we will continue to write simply* $F(t, \psi, u)$ *instead of* $F(t, \phi, u)$ (*and similarly for the function* H), *where* $\phi = (\psi, \chi)$.

Moreover, assume that (i) *the set* B *is a closed subset of* $E^1 \times C([-r, 0]) \times E^1 \times C([-r, 0])$ *which is independent of the functions* $x^i_{t_2}$, $i = \alpha + 1, \cdots, n$, *and whose projection into its second coordinate space is compact*; *and* (ii) *the map* g *is bounded and has the form* $g^i = 0$, $i = \alpha + 1, \cdots, n$. *Let* $h(t_1, \phi_1, t_1, \phi_2)$ *be a bounded real-valued continuous function defined on* B, *which is monotone nondecreasing with respect to each variable* $\phi_2^{\alpha+1}, \cdots, \phi_2^n$. *Let* Ω *be the class of all admissible pairs* (*in the sense of Definition* 3.1) *for which* $H(t, y_t, u(t))$ *is L-integrable in* $[t_1, t_2]$ *and*

$$\int_{t_1}^{t_2} H(t, y_t, u(t))\, dt \leqq K_1$$

for some constant $K_1 \geqq 0$, *and assume* $\Omega \neq \varnothing$. *Then the functional* $I[x, u] = h[\eta(x)]$ *has an absolute minimum in* Ω.

Proof. By hypothesis, the cost functional h is bounded and hence $i = \inf_\Omega I[x, u] = h(\eta(x))$ is finite. Let $\{x^{(k)}, u^{(k)}\}$, each defined for $t_{1_k} \leqq t \leqq t_{2_k}$, be a minimizing sequence for I in Ω. Then $(t, x_t^{(k)}) \in A$ for $t \in [t_{1_k}, t_{2_k}]$ and, writing $x^{(k)} = (y^{(k)}, z^{(k)})$, we have $(t, y_t^{(k)}) \in A_0$, $z_t^{(k)} \in S$ for $t \in [t_{1_k}, t_{2_k}]$. Also, $u^{(k)}(t) \in U(t, y_t^{(k)})$ almost everywhere in $[t_{1_k}, t_{2_k}]$,

$$\int_{t_{1_k}}^{t_{2_k}} H(t, x_t^{(k)}, u^{(k)}(t))\, dt \leqq K_1,$$

$k = 1, 2, \cdots$, and $h(\eta(x^{(k)})) \to i$ as $k \to \infty$. Since A is bounded, the projection of A onto the t-axis is contained in some interval of the form $[t_1^*, t_2^*]$ and, since S is bounded, we have $\|x_{t_2}^{(k)i}\| \leqq a_i$, $k = 1, \cdots, n$, $i = \alpha + 1, \cdots, n$.

Let x^0 denote a new variable satisfying

$$dx^0/dt = H(t, y_t, u), \qquad x_{t_1}^0 \equiv 0.$$

Let

$$x^{(k)0}(t) = \int_{t_{1_k}}^{t} H(S, x_s^{(k)}, u^{(k)}(s))\, ds, \qquad t_{1_k} \leqq t \leqq t_{2_k}, \quad k = 1, 2, \cdots,$$

and note that $d/dt(x^{(k)0}(t)) \geqq 0$, $\|x_t^{(k)0}\| \leqq K_1$. Let S^0 be the sphere of radius K_1 in $C([-r, 0], E^1)$ and let $A' = S^0 \times A$.

Since the functions $x^{(k)i}$, $i = 1, 2, \cdots, \alpha$, $k = 1, 2, \cdots$, satisfy the growth condition (γ_i), we have as in Theorem 6.1 that the functions $g_k^i = x^{(k)i}(t) - g^i(t, x_t^{(k)})$ are equiabsolutely continuous and since, by hypothesis, A and $g[A]$ are bounded, we may extract a convergent subsequence which we again call $\{h_n\}$. As before, we may extract a further subsequence with the property that the corresponding functions $x^{(k)i}$ converge in the ρ-metric to a continuous function x^i. Writing $y = (x^1, x^2, \cdots, x^\alpha)$, we have $y^{(k)} \to y$ in the ρ-metric and so, in view of the compactness assumption on the set of initial conditions, in particular,

$$y_{t_{1_k}}^{(k)} \to y_{t_1} \quad \text{and} \quad y_{t_{2_k}}^{(k)} \to y_{t_2} \quad \text{as } k \to \infty.$$

We now consider the sequences $\{x^{(k)i}\}$, $i = 0$ and $i = \alpha + 1, \cdots, n$, of scalar functions defined, for $t_{1_k} \leqq t \leqq t_{2_k}$, by

$$x^{(k)0}(t) = \int_{t_{1_k}}^{t} H(s, y_s^{(k)}, u^{(k)}(s))\, ds,$$

$$x^{(k)i}(t) = \int_{t_{1_k}}^{t} F_i(s, y_s^{(k)}, u^{(k)}(s))\, ds, \qquad i = \alpha + 1, \cdots, n,$$

where $H \geqq 0$ and $F_i \geqq 0$. It follows that the functions $x^{(k)i}$, $i = 0$ and $i = \alpha + 1, \cdots, n$, are monotone nondecreasing and we may use Helly's theorem to extract a further subsequence, say again $\{x^{(k)i}\}$ such that $x^{(k)i}(t) \to x^i(t)$ as $k \to \infty$ for all t and such that the limits $x_{t_1}^i = \lim x_{t_{1_k}}^{(k)i}$ and $x_{t_2}^i = \lim x^{(k)i}(t_{2_k})$ exist as $k \to \infty$, $i = 0$, and $i = \alpha + 1, \cdots, n$. Note, in particular, that $x_{t_1}^0 \equiv 0$.

Since A_0 is closed, we have that $(t, y_t^{(k)}) \in A_0$ implies that $(t, y_t) \in A_0$ for all t, $t_1 \leqq t \leqq t_2$, and thus $(t, y_t, z_t) \in A = A_0 \times S$. Furthermore, $(t, x_t^0, y_t, z_t) \in A' = S^0 \times A$ for all t, $t_1 \leqq t \leqq t_2$.

Making the usual decomposition, we write, for $i = 0$ and $i = \alpha + 1, \cdots, n$, $x^i(t) = X^i(t) + S^i(t)$, $t \in [t_1 - r, t_2]$, where $X^i(t)$ is absolutely continuous on $[t_1, t_2]$ and $S^{i\prime}(t) = 0$ almost everywhere on $[t_1, t_2]$, and with $X_{t_1}^i = x_{t_1}^i$, $S_{t_1}^i \equiv 0$, X^i, S^i nondecreasing with $S^i(t) \geqq 0$ and $X_{t_2}^i(\theta) \leqq \psi_{t_2}^i(\theta)$, $\theta \in [-r, 0]$, $i = 0$ and $i = \alpha + 1, \cdots, n$.

Let $\tilde{u} = (v^0, u^1, \cdots, u^m, v^{\alpha+1}, \cdots, v^n) = (v^0, u, v)$ be an auxiliary control vector, $\tilde{u} \in E^{m+n-\alpha+1}$, and let $\tilde{U}(t, \psi) = \{\tilde{u} : u = (u^1, \cdots, u^m) \in U(t, \psi), v^0 \geqq H(t, \psi, u), v^i \geqq F_i(t, \psi, u), i = \alpha + 1, \cdots, n\}$. Let $\tilde{\tilde{F}}(t, \psi, u)$ and $\tilde{F}(t, \psi, u)$ be the vector functions

$$\tilde{\tilde{F}}(t, \psi, u) = (\tilde{F}_0, F_1, \cdots, F_\alpha, \tilde{F}_{\alpha+1}, \cdots, \tilde{F}_n),$$

$$\tilde{F}(t, \psi, u) = (F_1, \cdots, F_\alpha, \tilde{F}_{\alpha+1}, \cdots, \tilde{F}_n)$$

with $\tilde{F}_0 = v^0$, $\tilde{F}_i = v^i$, $i = \alpha + 1, \cdots, n$. Finally, consider the auxiliary differential system

$$\begin{aligned} &dx^0/dt = v^0, \\ &d/dt[x^i(t) - g^i(t, x_t)] = F_i(t, y_t, u), && i = 1, \cdots, \alpha, \\ &dx^i/dt = v^i, && i = \alpha + 1, \cdots, n, \end{aligned}$$

with constraints

$$\begin{aligned} &v^0(t) \geqq H(t, y_t, u(t)), \\ &u(t) \in U(t, y_t), \\ &v^i(t) \geqq F_i(t, y_t, u(t)), && i = \alpha + 1, \cdots, n. \end{aligned}$$

Note that $\tilde{\tilde{F}}(t, \phi, \tilde{U}(t, \psi)) = Q_H(t, \psi)$, $\tilde{F}(t, \psi, \tilde{U}) = Q(t, \psi)$, and Q_H satisfies property (Q) in A_0. Since H, $F_{\alpha+1}, \cdots, F_n$ are continuous, the set $\tilde{M}_0 = \{(t, \psi, v^0, u, v)$: $(t, \psi) \in A_0$, $u \in U(t, \psi)$, $v^0 \geqq H$, $v^i \geqq F_i$, $i = \alpha + 1, \cdots, n\}$ is closed. Thus, by Theorem 4.1, there is a measurable function $u(t) = (v^0, u, v)$, $t_1 \leqq t \leqq t_2$, with $u(t) \in U(t, y_t)$, such that

$$\begin{aligned} &dX^0/dt = v^0 \geqq H(t, y_t, u(t)) \geqq 0, \\ &d/dt[x^i(t) - g^i(t, y_t)] = F_i(t, y_t, u(t)), && i = 1, \cdots, \alpha, \\ &dX^i/dt = v^i \geqq F_i(t, y_t u(t)) \geqq 0, && i = \alpha + 1, \cdots, n, \end{aligned}$$

almost everywhere in $[t_1, t_2]$. Taking

$$z^0(t) = \int_{t_1}^{t} H(s, y_s, u(s))\, ds, \qquad z^0_{t_1} \equiv 0,$$

$$z^i(t) = x^i(t_1) + \int_{t_1}^{t} F(s, y_s, u(s))\, ds, \qquad z^i_{t_1} = x^i_{t_1},$$

$$i = \alpha + 1, \cdots, n,$$

it may be easily checked that the pair $\{X, u\}$ where $X(t) = (z^0(t), y(t), Z(t))$, $t_1 \leqq t \leqq t_2$, is an admissible pair for the auxiliary problem. Hence $\{X, u\} \in \Omega$, and $h[\eta(X)] \geqq i$. On the other hand, h is monotone nondecreasing in $\phi_2^{\alpha+1}, \cdots, \phi_2^n$ and hence

$$\begin{aligned} h[\eta(X)] &= h(t_1, x^1_{t_1}, \cdots, x^n_{t_1}, t_2, x^1_{t_2}, \cdots, x_{t_2}, X^{\alpha+1}_{t_2}, \cdots, X^n_{t_2}) \\ &\leqq h(t_1, x_{t_1}, t_2, x_{t_2}) = \lim_{k \to \infty} h[\eta(x^{(k)})] = i. \end{aligned}$$

Thus $h[\eta(X)] = i$ and this completes the proof.

Remark (added in proof). Recent improvements on the results of the final section could not be incorporated here for technical reasons. The results will be presented elsewhere.

Acknowledgments. The author would like to express his thanks to Prof. L. Cesari for his many helpful suggestions during the preparations of this paper, and to the referees, whose suggestions have helped the author to significantly simplify the presentation.

REFERENCES

[1] T. S. ANGELL, *Existence theorems for a class of optimal control problems with delay*, Doctoral thesis, Department of Mathematics, University of Michigan, Ann Arbor, Mich., 1969.

[2] ———, *Existence theorems for optimal control problems involving functional differential equations*, J. Optimization Theory Appl., 7 (1971), pp. 149–169.

[3] ———, *On the optimal control of systems governed by non-linear Volterra equations*, Ibid., to appear.

[4] ———, *Existence of optimal control without convexity and a bang-bang theorem for Volterra equations*, Ibid., to appear.

[5] H. T. BANKS AND G. A. KENT, *Control of functional differential equations of retarded and neutral type to target sets in function space*, this Journal, 10 (1972), pp. 567–593.

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

[7] ———, *Existence and lower closure theorems for abstract control problems*, this Journal, 12 (1974), pp. 27–42.

[8] M. F. BIDAUT, *Quelques résultats d'existence pour des problèmes de contrôle optimal*, C. R. Acad. Sci. Paris, Ser. A, 274 (1972), pp. 62–65.

[9] P. BRUNOVSKY, *On the necessity of a certain convexity condition for lower closure of control problems*, this Journal, 6 (1968), pp. 174–185.

[10] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints I, II*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–430.

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

[12] ———, *Existence theorems in problems of optimization*, University of Michigan mimeographed notes, Ann Arbor, Mich., 1969.

[13] ———, *Closure theorems for orientor fields and weak convergence*, Arch. Rational Mech. Anal., 55 (1974), pp. 332–356.

[14] R. DRIVER, *Existence and continuous dependence of solutions of a neutral functional differential equation*, Ibid., 19 (1965), pp. 147–166.

[15] J. K. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1971.

[16] J. K. HALE AND M. A. CRUZ, *Existence, uniqueness and continuous dependence for hereditary systems*, Ann. Mat. Pura Appl., 85 (1970), pp. 63–81.

[17] J. K. HALE AND K. R. MEYER, *A class of functional equations of neutral type*, Mem. Amer. Math. Soc., 76 (1967).

[18] E. J. MCSHANE, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.

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

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

[21] W. WALTER, *Differential and Integral Inequalities*, Springer-Verlag, New York, 1970.

[22] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

CONTROLLABILITY OF THE NONLINEAR WAVE EQUATION IN SEVERAL SPACE VARIABLES*

WILLIAM C. CHEWNING†

Abstract. On a rectangular parallelopiped in R^N, $N \geqq 2$, we consider the equation $u_{tt} = \Delta u + f(u, u_t)$, where f is a nonlinear perturbation meeting certain conditions. We prove that the above system is locally controllable at $u = 0, u_t = 0$; i.e., the set of states in a certain function space which can be reached from (0, 0) in a finite time $T < \infty$ using boundary controls is an open neighborhood of (0, 0) in that function space. These results generalize to the nonlinear case conclusions obtained by Russell for the linear wave equation, in which global controllability was established.

1. Introduction. Cirina has studied controllability for a nonlinear wave equation in one space variable in [1]. Recently, Russell has announced in [7] results which include the following.

PROPOSITION. *Let Ω be a bounded domain in R^N, $N \geqq 2$, with boundary Γ, a piecewise $(N-1)$-manifold of class C^∞. Consider the problem*

$$u_{tt} = \Delta u, \quad u(0) = u_t(0) = 0, \quad u|_\Gamma = g. \tag{1}$$

Then there is a time $T < \infty$, such that for any (u_0, v_0) in $H^2(\Omega) \times H^1(\Omega)$, a control $g \in H^{3/2}(\Gamma \times [0, T])$ exists for which (1) has a unique solution with $u(T) = u_0$, $u_t(T) = v_0$. Moreover,

$$|g|_{H^{3/2}(\Gamma \times [0,T])} \leqq K(|u_0|_{H^2(\Omega)} + |v_0|_{H^1(\Omega)}).$$

Thus Russell has established the global controllability of the wave equation in space, by controls which depend continuously on the states one intends to reach.

In Theorem 1 of [6, pp. 366–367], the authors use the inverse function theorem in R^N to prove that a nonlinear control system (governed by ordinary differential equations) is locally controllable if its linear approximation is controllable. In a similar way we shall prove that a class of nonlinear wave equations is locally controllable using the inverse function theorem in the Banach space $H^2(\Omega) \times H^1(\Omega)$, together with Russell's result on controllability for the linear wave equation. In order to keep the ideas clear and the results easily stated, we do not consider the most general possible choices for the domain Ω, the elliptic operator (Δ), the method of exercising boundary control, or the function space in which the problem is solved. Finally, one could consider more general nonlinear perturbations as well.

Details of the problem. Let X be a rectangular parallelopiped in R^N with boundary Y. For any space $V \subset R^K$, we denote by $H^r(V)$ the Sobolev space of order r on V. The space $H^2(X) \times H^1(X)$ we denote as H, and elements in H will be written as (u, v), $\binom{u}{v}$, or w depending on the situation.

Writing the nonlinear wave equation as a system, we have

$$\dot{v} = Av + F(v), \quad v(0) = 0, \quad v|_Y = h, \tag{2}$$

* Received by the editors April 19, 1974, and in revised form November 24, 1974. We regret to report the death of Professor Chewning on March 23, 1975.

† Department of Mathematics and Computer Science, University of South Carolina, Columbia, South Carolina 29208.

where $v = \binom{u}{u_t} \in H$, A is the linear operator

$$A \equiv \begin{pmatrix} 0 & I \\ \Delta & 0 \end{pmatrix}, \qquad D_A = H^3(X) \times H^2(X),$$

$$h \in H^{3/2}(Y \times [0, T]) \times H^{1/2}(Y \times [0, T]),$$

and

$$F: H \to H^1(X) \quad \text{by } F\binom{u}{u_t} = \binom{0}{f(u, u_t)}.$$

The assumptions on f are:

(a) f is continuous and is continuously Fréchet differentiable in a neighborhood of $(0, 0)$.

(b) $f(0, 0) = 0$.

(c) $$|f(u, v) - f(\hat{u}, \hat{v})|_1 \leqq (|u - \hat{u}|_2 + |v - \hat{v}|_1) \cdot C(|u|_2 + |v|_1, |\hat{u}|_2 + |\hat{v}|_1),$$

where $C: R^+ \times R^+ \to R^+$ is continuous and $|w|_k \equiv |w|_{H^k(X)}$.

(d) The Fréchet derivative of f is locally Lipschitz in a neighborhood of $0 \in H$.

Comment. It is not necessary that f be given by a function, i.e.,

$$f(u, u_t)(\mathbf{x}) \equiv g(u(\mathbf{x}, t), u_t(\mathbf{x}, t))$$

for some $g: R^2 \to R^1$. If this is the case, however, then f will satisfy the above assumptions if g has Lipschitz continuous mixed partial derivatives up through order three.

2. Existence and uniqueness of a solution to (2). Our plan is to obtain a boundary control h_u for any given $u \in H$, such that the problem

$$\dot{w} = Aw, \quad w(0) = 0, \quad w|_Y = h_u \tag{3}$$

has a unique solution $w \in C([0, T], H)$ with $w(T) = u$. If w is small, we can then solve

$$\dot{z} = Az + F(z + w), \quad z(0) = 0, \quad z|_Y = 0. \tag{4}$$

Then defining $v = w + z$, we obtain a unique solution to (2) with $v(T) = u$.

3. The linear problem (3). Our first task is to sharpen statements made in Theorem 2.1 of [7] about solutions to (3).

LEMMA 1. *For $\delta > 0$ and r a positive integer, there is a bounded linear operator $E_r: H^r(X) \to H^r(R^N)$ such that $E_r(f)$ is an extension of f to R^N and* supp $(E_r(f)) \subset X_\delta$, *the δ-neighborhood of X.*

Proof. No confusion will result from speaking of equivalence classes in $H^r(X)$ as though they were functions. Since X is a box, we can construct $\hat{X}$, a box composed of 3^N copies of X having the original X as the center box. Let the boxes in $\hat{X}$ be numbered so that $\hat{X} = \bigcup_{i=1}^{3^N} X_i$. Let the faces of X be numbered in any order, and then number the faces of each X_i in exactly the same way. $\hat{X}$ is constructed, starting with the center box, such that if X_i and X_j have a common face, in each box the face has the same number. If we start in one dimension, we can clearly

reflect a closed interval through its left- and right-hand endpoints in turn, to form three intervals that are matched as described above. One can then fit a rectangle as the middle block in a 3×3 group of nine identical rectangles with appropriate sides in contact. Inductively, one extends this to a $3 \times 3 \times 3$ packing of rectangular solids with appropriate faces in contact, etc. If $f: X \to R$, we define $f_i: X_i \to R$ as follows. Let $\psi_i: X_i \to X$ be the linear homeomorphism which maps $X_i \subset \hat{X}$ onto X with corresponding faces identified. Then $f_i(x) \equiv f \circ \psi_i(x)$. Thus if $f \in H^r(X)$, then $\hat{f} \equiv \bigcup_{i=1}^{3^N} f_i$ is in $H^r(\hat{X})$. Now let $\rho: R^N \to R$ be a C^∞ function such that $\rho|_X = 1$ and $\operatorname{supp}(\rho) \subset X_\delta \subset \hat{X}$.

We now define $E_r: H^r(X) \to H^r(R^N)$ by

$$E_r(f) = \begin{cases} \rho(x) \cdot \hat{f}(x), & x \in \hat{X} \\ 0, & x \notin \hat{X} \end{cases}.$$

It is clear that E_r is a bounded linear operator.

LEMMA 2. *Let B be a rectangular parallelopiped in R^N, and for $(f, g) \in H^2(B) \times H^1(B)$, $(f, g)|_{\partial B} = (0, 0)$ consider the problem*

$$\dot{u} = Au, \quad u|_{\partial B} = 0, \quad u(0) = \begin{pmatrix} f \\ g \end{pmatrix}. \tag{5}$$

For any fixed $T < \infty$, the solution $u(f, g)$ is in $C([0, T], H^2(B) \times H^1(B))$ and the correspondence $(f, g) \to u(f, g)$ is a bounded linear operator from $H^2(B) \times H^1(B)$ to $C([0, T], H^2(B) \times H^1(B))$.

Proof. We recall that $A \equiv \begin{pmatrix} 0 & I \\ \Delta & 0 \end{pmatrix}$. By direct separation of variables techniques, one can find a complete orthonormal basis for $L^2(B)$ consisting of eigenvectors of Δ. We term these as $\{\phi_n\}$ with corresponding eigenvalues $\{-\lambda_n^2\}$. Using direct elementary calculations and Green's identity, one can verify that $\{\phi_n\}$ is also a complete orthogonal basis for $H^1(B)$ and $H^2(B)$. By normalization, we obtain $\{\rho_n\}$ and $\{\psi_n\}$, complete orthonormal bases for $H^1(B)$ and $H^2(B)$, respectively. Moreover, $\Delta\rho_i = -\lambda_i^2 \rho_i$ and $\Delta\psi_i = -\lambda_i^2 \psi_i$; $i = 1, 2, \cdots$.

Using these eigenfunctions we construct the strongly continuous semigroup $\{R(t)\}_{t \geq 0}$ of operators on $H^2(B) \times H^1(B)$ whose infinitesimal generator is A. If such a semigroup exists, then standard results in operator semigroups imply that the mapping $M\begin{pmatrix} f \\ g \end{pmatrix} = R(t)\begin{pmatrix} f \\ g \end{pmatrix}$ is a bounded linear mapping from $H^2(B) \times H^1(B)$ to $C([0, T]; H^2(B) \times H^1(B))$. (See [5]). Let $\langle \cdot, \cdot \rangle_2, \langle \cdot, \cdot \rangle_1$ denote inner products in $H^2(B), H^1(B)$ respectively. Then the definition of $R(t)$, for $t \geqq 0$, is

$$R(t)\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{\infty} [\langle f, \psi_n \rangle_2 \cos \lambda_n t + (1/\lambda_n)\langle g, \rho_n \rangle_1 \sin \lambda_n t]\psi_n \\ \sum_{n=1}^{\infty} [-\lambda_n \langle f, \psi_n \rangle_2 \sin \lambda_n t + \langle g, \rho_n \rangle_1 \cos \lambda_n t]\rho_n \end{pmatrix}$$

It is not hard to verify that $R(0) = I$, $\{R(t)\}$ is strongly continuous, $R(t + s) = R(t)R(s)$ and $(d/dt)R(t)x = AR(t)x$.

THEOREM 1. *For $u \in H$, there is a $T < \infty$ and an $h_u \in H^{3/2}(Y \times [0, T]) \times H^{1/2}(Y \times [0, T])$ such that (3) has a unique solution $w_u \in C([0, T], H)$ with $w_u(T) = u$.*

Proof. We supply a proof for $N \geqq 3, N$ odd. The case $N \geqq 2, N$ even can be argued in a similar way after the proof of Theorem 2.1 of [7] is understood. We do not repeat the (fairly lengthy) details here.

Given $u = \begin{pmatrix} f \\ g \end{pmatrix} \in H^2(X) \times H^1(X)$, extend this to $u_\delta = \begin{pmatrix} f_\delta \\ g_\delta \end{pmatrix} \in H^2(R^N) \times H^1(R^N)$ by the operator $E_2 \times E_1$ as defined in Lemma 1. Then solve the Cauchy problem:

$$\dot{v} = Av, \quad v(0) = u_\delta, \quad v \in H^2(R^N) \times H^1(R^N). \tag{6}$$

There is a time $T < \infty$ such that $v(T)|_X = 0$. We take $h_u \equiv v(t)|_Y$ for $0 \leqq t \leqq T$. Standard theorems imply that $h_u \in H^{3/2}(Y \times [0, T]) \times H^{1/2}(Y \times [0, T])$, and that, by uniqueness, the problem

$$\dot{w} = Aw, \quad w(0) = 0, \quad w(t)|_Y = h_u(T - t)$$

has a unique solution $w_u \in H$ with $w_u(T) = u$. Actually, $w_u(t) \equiv v(T - t)|_X$. This is a sketch of Russell's proof; details appear in [7].

We observe that since $T < \infty$ and supp f_δ, supp $g_\delta \subset X_\delta$, there is a large rectangular parallelopiped $B \supset X$ such that the solution to (6) is identically zero on ∂B and outside of B for $0 \leqq t \leqq T$. (The details are supplied by considering cones of influence for the wave equation in R^N.) Thus (6), for $0 \leqq t \leqq T$, is equivalent to

$$\dot{z} = Az, \quad z(0) = \begin{pmatrix} f_\delta \\ g_\delta \end{pmatrix}, \quad z|_{\partial B} = 0. \tag{7}$$

From Lemma 2, (7) has the solution $z(t) = R(t)u_\delta$, and clearly $z(t)|_X = v(t)|_X$ for $0 \leqq t \leqq T$.

The following notation will be needed in the proof of Lemma 4. If $R^N \supset V \supset W$ and $(f, g) \in H^2(V) \times H^1(V)$, define

$$\tau(V, W): H^2(V) \times H^1(V) \to H^2(W) \times H^1(W) \quad \text{by } \tau(V, W)\begin{pmatrix} f \\ g \end{pmatrix} = (f|_W, g|_W).$$

We define the extension operator

$$E: H \to H^2(B) \times H^1(B) \quad \text{by } E(f, g) = \tau(R^N, B) \circ E_2 \times E_1(f, g).$$

Notice that $v(t)|_X = \tau(B, X) \circ R(t) \circ E(f, g)$ on $[0, T]$. It follows that on $[0, T]$,

$$w_u(t) = \tau(B, X) \circ R(T - t) \circ E(u). \tag{8}$$

We have proved that $w_u \in C([0, T], H)$ because $\tau(B, X)$ and E are bounded linear operators and $R(T - t)$ is a uniformly (in t) bounded linear operator. We also have the useful formula (8).

4. The nonlinear problem (4). From Lemma 2, we know that the operator A with trivial boundary conditions on X, is the infinitesimal generator of a strongly continuous semigroup $\{S(t)\}_{t\geqq 0}$ and standard results imply that $\|S(t)\| \leqq Me^{at}$. In the proof of Theorem 2, we need these notations:

(a) The norm of $v \in C([0, T], H)$ is $\|v\|$.

(b) The function C (defined in connection with the Lipschitz condition on the nonlinear function f) has property r if there is an $r > 0$ and a $\rho < 1$ such that $C(x, y) \leqq \rho/TMe^{aT}$ for all $(x, y) \in R^+ \times R^+ : x \leqq r, y \leqq r$.

THEOREM 2. *Suppose for some $\rho < 1$ and some $r > 0$, that the function C has property r. Then if $\|w\| < (1 - \rho)r$, (4) has a unique continuously differentiable solution $z \in C([0, T], H)$. Moreover, $\|z\| \leqq Q\|w\|$.*

Proof. We represent a solution to (4) as

$$z(t) = \int_0^t S(t - v)F[z(v) + w(v)]\, dv. \tag{9}$$

Solving for z by Picard iteration, one has

$$z_0(t) \equiv \int_0^t S(t - v)F[w(v)]\, dv$$

and

$$z_{n+1}(t) \equiv \int_0^t S(t - v)F[w(v) + z_n(v)]\, dv.$$

Justifying the assumption later, we assume that $\|w\| < r$ and $\|w + z_n\| \leqq r$ for $n = 0, 1, 2, \cdots$. Then $\|z_0\| \leqq \|w\|\rho$ and $\|z_{n+1} - z_n\| \leqq \rho\|z_n - z_{n-1}\|$. It follows that

$$\|z_n\| \leqq \|z_0\| + \sum_{k=1}^{n} \|z_k - z_{k-1}\| \leqq \rho\|w\| \sum_{k=0}^{n} \rho^k \leqq \frac{\|w\|\rho}{1 - \rho}.$$

Therefore

$$\|z_n + w\| \leqq \|z_n\| + \|w\| \leqq \|w\|(1 + \rho/(1 - \rho)) = \|w\|(1/(1 - \rho)).$$

For the above estimates to be valid, then, we must require $\|w\| \leqq (1 - \rho)r$.

If $\|w\| < (1 - \rho)r$, the iteration procedure clearly produces a Cauchy sequence in $C([0, T], H)$ and thus (4) has a solution $z(t) \equiv \lim_{n\to\infty} z_n(t)$. The differentiability of z follows from p. 6 of [5]. Easy calculations show that the right-hand derivatives of z satisfy (4) and, as z is differentiable, it must satisfy (4). The uniqueness of z is a standard consequence of the local Lipschitz condition on f.

5. Local controllability of the nonlinear equation. With ρ fixed < 1, we assume that the function C has property r.

LEMMA 3. *If $y \in H$ is small, then the problem*

$$\dot{u} = Au + F(u), \quad u(0) = 0, \quad u|_\Gamma = h_y \tag{10}$$

has a unique continuous solution $u(\cdot, y) \in C([0, T], H)$. ($h_y$ is the boundary control computed in Theorem 1 which steers (3) from 0 to y in time T.)

Proof. Given $y \in H$, we first obtain w_y, the solution to (3) with final state y. As is argued in Theorem 1, the correspondence $y \to w_y$ is a bounded linear operator from H to $C([0, T], H)$. Therefore, if y is sufficiently small, w_y will be small enough for Theorem 2 to apply and we extract a solution z_y to (4) with $w = w_y$. The function $u_y \equiv w_y + z_y$ is clearly a solution to (10); its uniqueness follows from the local Lipschitz assumption on F. The proof is completed.

We therefore have the nonlinear map G defined on a ball B about $0 \in H$:

$$G(y) = u(T, y) \equiv w_y(T) + z_y(T).$$

LEMMA 4. *G is continuously Fréchet differentiable on a neighborhood of $0 \in H$.*

Proof. $G(y) = y + \int_0^T S(T - v)F[z_y(v) + w_y(v)]\, dv$.

Proceeding formally, we compute

$$G'(y) = I + \int_0^T S(T - v)F'[z_y(v) + w_y(v)]\left[\frac{\partial z_y(v)}{\partial y} + \frac{\partial w_y(v)}{\partial y}\right] dv. \tag{11}$$

From (8) of Theorem 1, we have

$$w_y(v) = \tau(B, X) \circ R(T - v) \circ E(y),$$

so

$$\frac{\partial w_y(v)}{\partial y} = \tau(B, X) \circ R(T - v) \circ E$$

since the operator is linear. From (9), we have

$$\begin{aligned} \frac{\partial z_y(t)}{\partial y} &= \int_0^t S(t - v)F'[z_y(v) + w_y(v)]\left[\frac{\partial z_y(v)}{\partial y} + \frac{\partial w_y(v)}{\partial y}\right] dv \\ &= \int_0^t S(t - v)F'[z_y(v) + w_y(v)]\left[\frac{\partial z_y(v)}{\partial y} + \tau(B, X) \circ R(t - v) \circ E\right] dv. \end{aligned} \tag{12}$$

We note that (12) is a linear integral equation for $\partial z/\partial y$ with bounded kernel for y small, since z_y, w_y are known and $\|z_y\|, \|w_y\| \leqq K|y|$. Therefore (12) can be solved by an iterative procedure to yield a unique, bounded, continuous solution. Because F' is locally Lipschitz, the solution $\partial z_y/\partial y$ will vary continuously with y.

Returning to (11), we see that if y is small enough for w_y, z_y, and $\partial z_y/\partial y$ to exist on $[0, T]$, then the integrand in (11) is bounded on $[0, T]$ and continuous in y. It follows that $G'(y)$ is continuous in a neighborhood of $0 \in H$.

LEMMA 5. *Suppose that*

$$\int_0^T S(T - v)F'[0]\left[\frac{\partial z_y(v)}{\partial y} + \tau(B, X) \circ R(T - v) \circ E\right] dv$$

has norm in $L(H, H)$ less than one for $y = 0$. Then $G'(0)$ is a linear homeomorphism in $L(H, H)$.

Proof. Consider the expression (11) for $G'(y)$ when $y = 0$. Because $w_y = 0$ and $z_y = 0$ when $y = 0$, it follows that

$$G'(0) - I = \int_0^T S(T - v)F'[0]\left[\frac{\partial z_0(v)}{\partial y} + \tau(B, X) \circ R(T - v) \circ E\right] dv.$$

If the integral has norm less than one in $L(H, H)$, then $G'(0)$ clearly has a bounded inverse.

THEOREM 3. *Under the assumption of Lemma* 5, *there is a ball* $\hat{B}$ *about* $0 \in H$ *such that if* $x \in \hat{B}$, *there is a unique control* g_x *for which*

$$\dot{v} = Av + F(v), \quad v(0) = 0, \quad v|_Y = g_x \tag{13}$$

has a unique solution $v \in C([0, T], H)$ *with* $v(T) = x$. *Moreover* $g_x \in H^{3/2}(Y \times [0, T]) \times H^{1/2}(Y \times [0, T])$ *and* g_x *depends continuously on* x.

Proof. The mapping $G: B \to H$ is continuously differentiable on some ball B about $0 \in H$, and $G(0) = 0$. We also have demonstrated that $G'(0)$ is a linear homeomorphism. By the inverse function theorem [2, p. 268] there is a ball $\tilde{B}$ containing 0 on which G is a homeomorphism. Let $\hat{B} \subset G(\tilde{B})$.

If $x \in \hat{B}$, let $y = G^{-1}(x)$. The element y is small enough for $u(t, y)$ to exist. But $u(T, y) \equiv G(y) = x$; one can steer 0 to x in time T under the nonlinear system (13). To identify g_x with $y = G^{-1}(x)$, take h_y as it is defined in Theorem 1, i.e., $h_y = \tau(B, X) \circ R(T - t) \circ E(y)|_Y$. Let $g_x = h_y$; g_x steers 0 to $w_y(T) + z_y(T) = x$, g_x depends continuously on x and belongs to the function space named.

We note that for small x, $G^{-1}(x)$ can be approximated by $(G'(0))^{-1}x$ in the calculation of g_x for practical problems.

6. Summary. We have shown that if x is small in H, there is a boundary control g_x such that (13) has a continuous solution with $v(T) = x$. (If there were two solutions to (13), then the problem $\dot{v} = Av + F(v), v(0) = 0, v|_Y = 0$ would have a nonzero solution. This is not possible in view of the fact that F is locally Lipschitz.) Therefore we have a meaningful solution to the problem of local controllability, using boundary controls, of a certain class of nonlinear wave equations defined on a box in R^N.

REFERENCES

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

[2] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[3] A. FRIEDMAN, *Partial Differential Equations*, Holt, Reinhart and Winston, New York, 1966.

[4] P. GARABEDIAN, *Partial Differential Equations*, John Wiley, New York, 1964.

[5] G. LADAS AND V. LAKSHMIKANTHAM, *Differential Equations in Abstract Spaces*, Academic Press, New York, 1972.

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

[7] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., 52 (1973), pp. 189–211.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

GENERALIZED HILBERT NETWORKS*

VACLAV DOLEZAL†

Abstract. In the paper a general model of a nonlinear network is constructed. The model considered is a generalization of the Hilbert network introduced in [1]. It is assumed that the generalized Hilbert network consists of at most countably many lumped elements described by nonlinear multivalued operators from a subset of a Hilbert space $\mathcal{H}$ into $\mathcal{H}$. Several theorems are proved on the existence and uniqueness of the solution of the network. Also, conditions are established under which the admittance operator of a generalized Hilbert network is causal.

Introduction. The model of a Hilbert network which we construct in this paper has two ingredients: 1. an oriented locally finite graph G having at most countably many branches, which describes the interconnections of lumped elements of the network; 2. a multivalued operator $\hat{Z}$ defined on a subset of the underlying Hilbert space $\mathcal{H}$, which describes the behavior of network elements. We assume that the regime in the network is governed by Kirchhoff's laws.

As shown in [1], the oriented graph G of a network can be completely described by a linear bounded operator $\hat{a}$ on $\mathcal{H}$. Also, it was assumed there that $\hat{Z}$ is a single-valued operator defined on the entire space $\mathcal{H}$, and possibly satisfying the Lipschitz condition. However, these assumptions severely restrict the applicability of the model. For example, if a network is considered in the time-domain $[0, \tau]$, the model in [1] excludes the presence of differentiators. Similarly, if we consider the time-domain $[0, \infty)$, presence of integrators leads to difficulties.

On the other hand, in the present paper no such assumptions are made; in addition to that we allow that, in general, values of $\hat{Z}$ are subsets of $\mathcal{H}$, i.e., $\hat{Z}$ may be multivalued.

Due to this fact, our model encompasses networks containing differentiators as well as integrators, independently of whether the time-domain is a finite interval or not. Of course, since the analysis takes place in a Hilbert space, all energies associated with the network are finite, which, we believe, is a quite natural assumption.

Naturally, for this degree of generality we have to pay a price: our theorems giving necessary and sufficient conditions for existence of a solution are conceptual rather than practical in nature. On the other hand, it turns out that a quite elementary yet powerful concept of a monotonicity of $\hat{Z}$ guarantees the uniqueness of the network solution.

Basically, the ideas developed in this paper are similar to those given in the pioneering paper [3] by Minty; the approach to the problem, however, is different.

In the first part of the paper we consider the abstract network defined as a pair $\mathcal{N} = (Z, a)$, where Z, a are certain operators, and derive necessary and sufficient conditions for the existence and uniqueness of a solution. In the second part, the results on abstract networks are applied to a Hilbert network $\hat{\mathcal{N}} = (\hat{Z}, G)$ and theorems on uniqueness of the regime in $\hat{\mathcal{N}}$ are obtained. Moreover, using the

* Received by the editors February 12, 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 Grant 33568-X00.

concept of causality introduced by Saeks [2], we prove two theorems giving conditions under which the admittance operator of $\hat{\mathcal{N}}$ (expressing currents in terms of voltages) is causal.

As an example, we discuss a specific R, L, C-network in the time-domain, provided that the inductors and capacitors are linear and time-varying, and the resistors are nonlinear and multivalued.

Finally, using a theorem by Rockafellar [5], we give sufficient conditions for a network to have a solution for any vector of voltages.

1. Abstract networks. Let X, Y be nonempty sets, and let $\mathfrak{S}(Y)$ be the collection of all nonempty subsets of Y; a mapping $A : X \to \mathfrak{S}(Y)$ will be called a set mapping.

If $\mathcal{D} \subset X$, $\mathcal{D} \neq \emptyset$, we denote

$$(A\mathcal{D})^0 = \bigcup_{x \in \mathcal{D}} Ax. \tag{1.1}$$

If, in particular, A is a set mapping such that Ax is a singleton for each $x \in X$, then A will be called an operator. Since in this case A is in fact a mapping from X into Y, we have $(A\mathcal{D})^0 = A\mathcal{D}$.

Let A be a set mapping, and let $\mathcal{D} \subset X$, $\mathcal{D} \neq \emptyset$; A will be called simple on $\mathcal{D}$ if

$$x_1, x_2 \in \mathcal{D},\ x_1 \neq x_2 \Rightarrow (Ax_1) \cap (Ax_2) = \emptyset. \tag{1.2}$$

Clearly, if A is simple, then A is 1–1.

Let A be simple on $\mathcal{D}$; then we define the operator $A^- : (A\mathcal{D})^0 \to \mathcal{D}$, called the quasi-inverse of A, by the relations: if $y \in (A\mathcal{D})^0$, then $A^- y = x$, where $x \in \mathcal{D}$ is such that $y \in Ax$.

Since $\{Ax : x \in \mathcal{D}\}$ is a partitioning of $(A\mathcal{D})^0$ due to (1.2), our definition of A^- is meaningful and A^- maps $(A\mathcal{D})^0$ onto $\mathcal{D}$.

It is easy to see that A is simple if and only if for each $y \in (A\mathcal{D})^0$ there is a unique $x \in \mathcal{D}$ such that $y \in Ax$. In this case, $x = A^- y$.

Also, it is clear that if A is an operator and is simple on $\mathcal{D}$, then A^- coincides with the ordinary inverse $A^{-1} : A\mathcal{D} \to \mathcal{D}$.

If $A : X \leftarrow \mathfrak{S}(Y)$ is a set mapping and $B : Y \to Z$ is an operator, we define the set mapping $BA : X \to \mathfrak{S}(Z)$ by $(BA)x = B(Ax) \subset Z$ for each $x \in X$. Then, for any $\mathcal{D} \subset X$, $\mathcal{D} \neq \emptyset$, $((BA)\mathcal{D})^0 = B(A\mathcal{D})^0$, since $\bigcup_{x \in \mathcal{D}} (BA)x = \bigcup_{x \in \mathcal{D}} B(Ax) = B(\bigcup_{x \in \mathcal{D}} Ax)$.

If $C : U \to X$ is an operator, the set mapping AC is defined in a similar way.

Let $\mathcal{H}$ and $\mathcal{H}'$ be fixed Hilbert spaces. Let $\mathcal{D} \subset \mathcal{H}$, $\mathcal{D} \neq \emptyset$, and let $Z : \mathcal{D} \to \mathfrak{S}(\mathcal{H})$ be a set mapping; furthermore, let $a \in [\mathcal{H}, \mathcal{H}']$ (a linear bounded operator), $a \neq 0$. Then the ordered pair $\mathcal{N} = (Z, a)$ will be called an abstract network over $\mathcal{H}$.

DEFINITION. Let $\mathcal{N} = (Z, a)$, and let $e \in \mathcal{H}$; an element $i \in \mathcal{D}$ will be called *a solution of $\mathcal{N}$ corresponding to e* if

(i) there exists $v \in Zi$ such that

$$\langle c, v - e \rangle = 0 \tag{1.3}$$

for all $c \in \mathcal{H}$ with $ac = 0$,

(ii) $ai = 0$.

Denote $N_a = \{x : x \in \mathcal{H}, ax = 0\}$. Then the solution i can clearly be defined in the following equivalent way:

K_1^*: there exists a $v \in Zi$ such that $v - e \in N_a^\perp$,

K_2^*: $i \in N_a \cap \mathcal{D}$.

In the sequel we will assume that $N_a \cap \mathcal{D} \neq \emptyset$. Futhermore, let P be the orthogonal projection from $\mathcal{H}$ onto N_a. Then we have the following.

THEOREM 1.1. *Let $\mathcal{N} = (Z, a)$ be an abstract network over $\mathcal{H}$, and let $e \in \mathcal{H}$. Then $\mathcal{N}$ possesses a solution i corresponding to e if and only if*

$$e \in N_a^\perp + (Z(N_a \cap \mathcal{D}))^0. \tag{1.4}$$

Proof. (a) Let (1.4) hold. Then there exists $x \in N_a^\perp$ and $y \in (Z(N_a \cap \mathcal{D}))^0$ such that $e = x + y$. Consequently, by (1.1), there exists $i \in N_a \cap \mathcal{D}$ such that $y \in Zi$; thus $y - e = -x \in N_a^\perp$, i.e., i is a solution of $\mathcal{N}$ corresponding to e.

(b) Conversely, let i be a solution of $\mathcal{N}$ corresponding to e. Then there exists $v \in Zi$ such that $v - e \in N_a^\perp$, i.e., $e - v \in N_a^\perp$. Thus $e \in v + N_a^\perp \subset N_a^\perp + (Z(N_a \cap \mathcal{D}))^0$. Hence the proof.

Note that if $M \subset \mathcal{H}$ is nonempty, then $N_a^\perp + M = P^{-1}\{PM\}$. Hence we have the relation

$$N_a^\perp + (Z(N_a \cap \mathcal{D}))^0 = P^{-1}\{P(Z(N_a \cap \mathcal{D}))^0\}. \tag{1.5}$$

Let us now consider the uniqueness problem, i.e., to find subdomains of $\mathcal{D}$, on which the solution of a network is determined uniquely.

If $\tilde{\mathcal{D}} \subset \mathcal{D}$ and $N_a \cap \tilde{\mathcal{D}} \neq \emptyset$, we denote

$$Q(\tilde{\mathcal{D}}) = P^{-1}\{P(Z(N_a \cap \tilde{\mathcal{D}}))^0\}. \tag{1.6}$$

THEOREM 1.2. *Let $\mathcal{N} = (Z, a)$ and let $\tilde{\mathcal{D}} \subset \mathcal{D}$, $N_a \cap \tilde{\mathcal{D}} \neq \emptyset$. Then for each $e \in Q(\tilde{\mathcal{D}})$ there exists a unique solution i in $\tilde{\mathcal{D}}$ of $\mathcal{N}$ corresponding to e if and only if the set mapping PZ is simple on $N_a \cap \tilde{\mathcal{D}}$. In this case, $i = Ae$, where the operator $A : Q(\tilde{\mathcal{D}}) \to N_a \cap \tilde{\mathcal{D}}$ is defined by*

$$A = (PZ)^- P. \tag{1.7}$$

Proof. (a) Assume that PZ is simple on $N_a \cap \tilde{\mathcal{D}}$ and let $e \in Q(\tilde{\mathcal{D}})$. By (1.5) and Theorem 1.1, there exists at least one solution of $\mathcal{N}$ corresponding to e. Suppose that $i_1, i_2 \in N_a \cap \tilde{\mathcal{D}}$ are solutions corresponding to e. Then, by K_1^*, there exist $v_1 \in Zi_1$ and $v_2 \in Zi_2$ such that $v_1 - e$, $v_2 - e \in N_a^\perp$, so that $v_1 - v_2 \in N_a^\perp$. Consequently, $P(v_1 - v_2) = 0$, i.e., $Pv_1 = Pv_2$. This, however, means that $(PZi_1) \cap (PZi_2) \neq \emptyset$; hence by definition of a simple mapping, $i_1 = i_2$, i.e., the solution of $\mathcal{N}$ is determined uniquely.

(b) Assume now that for each $e \in Q(\tilde{\mathcal{D}})$ the network $\mathcal{N}$ possesses a unique solution. Suppose that for some $i_1, i_2 \in N_a \cap \tilde{\mathcal{D}}$ we have $(PZi_1) \cap (PZi_2) \neq \emptyset$. This means that there exist $v_1 \in Zi_1$ and $v_2 \in Zi_2$ such that $Pv_1 = Pv_2 = y$. It is clear that $Py = y$, and also $y \in P(Z(N_a \cap \tilde{\mathcal{D}}))^0 \subset P^{-1}\{P(Z(N_a \cap \tilde{\mathcal{D}}))^0 = Q(\tilde{\mathcal{D}})$. Thus, we have $P(v_1 - y) = 0$ and $P(v_2 - y) = 0$, i.e., $v_1 - y$, $v_2 - y \in N_a^\perp$. Hence, by K_1^*, K_2^*, both i_1

and i_2 are solutions of $\mathcal{N}$ corresponding to $y \in Q(\tilde{\mathcal{D}})$; consequently, by our hypothesis, $i_1 = i_2$, i.e., PZ is simple on $N_a \cap \tilde{\mathcal{D}}$.

To conclude the proof, assume that PZ is simple on $N_a \cap \tilde{\mathcal{D}}$. Observe that $((PZ)(N_a \cap \tilde{\mathcal{D}}))^0 = P(Z(N_a \cap \tilde{\mathcal{D}}))^0$. Construct the quasi-inverse $(PZ)^- : P(Z(N_a \cap \tilde{\mathcal{D}}))^0 \to N_a \cap \tilde{\mathcal{D}}$ as it is defined above. Choosing $e \in Q(\tilde{\mathcal{D}})$, we have $Pe \in P(Z(N_a \cap \tilde{\mathcal{D}}))^0$; consequently, $i = Ae = (PZ)^- Pe$ is well-defined and is in $N_a \cap \tilde{\mathcal{D}}$. Thus, i satisfies K_2^*. By definition of $(PZ)^-$ we have $Pe \in PZi$; thus, there exists $v \in Zi$ such that $Pe = Pv$, i.e., $v - e \in N_a^\perp$. Hence, i is the solution of $\mathcal{N}$ corresponding to e which concludes the proof.

Let $\tilde{\mathcal{D}} \subset \mathcal{D}$; motivated by Theorem 1.1 and 1.2, we will say that $\mathcal{N}$ is regular on $\tilde{\mathcal{D}}$ if for each $e \in Q(\tilde{\mathcal{D}})$ the network $\mathcal{N}$ possesses in $\tilde{\mathcal{D}}$ a unique solution i corresponding to e. Then we have the following.

THEOREM 1.3. *Let $e \in Q(\mathcal{D})$ and let $i \in \mathcal{D}$ be a solution of $\mathcal{N}$ corresponding to e. Then there exists a $\tilde{\mathcal{D}}_i \subset \mathcal{D}$ such that*

(i) $i \in \tilde{\mathcal{D}}_i$,

(ii) *$\mathcal{N}$ is regular on $\tilde{\mathcal{D}}_i$,*

(iii) *$\tilde{\mathcal{D}}_i$ is maximal, i.e., $\mathcal{N}$ is not regular on any other $\tilde{\mathcal{D}}$ which properly contains $\tilde{\mathcal{D}}_i$.*

Proof. Let $\mathcal{S}_i = \{\tilde{\mathcal{D}}^\alpha : \tilde{\mathcal{D}}^\alpha \subset \mathcal{D}, i \in \tilde{\mathcal{D}}^\alpha, PZ \text{ is simple on } N_a \cap \tilde{\mathcal{D}}^\alpha\}$. The collection $\mathcal{S}_i$ is nonempty, since $\tilde{\mathcal{D}}^0 = \{i\} \in \mathcal{S}_i$. Moreover, $\mathcal{S}_i$ is partially ordered by set inclusion; also, if $\mathcal{T} = \{\tilde{\mathcal{D}}^\alpha : \alpha \in I\} \subset \mathcal{S}_i$ is a chain, then clearly $\bigcup_{\alpha \in I} \tilde{\mathcal{D}}^\alpha \in \mathcal{S}_i$ and is an upper bound for $\mathcal{T}$. Hence, by Zorn's lemma, there exists a maximal element $\tilde{\mathcal{D}}_i$ in $\mathcal{S}_i$. Then Theorem 1.2 concludes the proof.

Note that, even in the case that $i \in \mathcal{D}$ is unique for some $e \in Q(\mathcal{D})$, the maximal subdomain $\tilde{\mathcal{D}}_i$ need not be determined uniquely. This is demonstrated by the example $N_a \cap \mathcal{D} = R^1$ and $(PZ)x = x^2$.

Let us now establish some sufficient conditions for regularity. Without loss of generality we may assume that $\tilde{\mathcal{D}} = \mathcal{D}$.

THEOREM 1.4. *Let $\mathcal{N} = (Z, a)$ be an abstract network over $\mathcal{H}$.*

(i) *If for all $x_1, x_2 \in N_a \cap \mathcal{D}$, $x_1 \neq x_2$ and all $y_1 \in Zx_1$, $y_2 \in Zx_2$ we have*

$$\langle y_1 - y_2, x_1 - x_2 \rangle \neq 0, \tag{1.8}$$

then $\mathcal{N}$ is regular on $\mathcal{D}$. If, in addition, Z is an operator, then

$$\langle Ae_1 - Ae_2, e_1 - e_2 \rangle \neq 0 \tag{1.9}$$

for all $e_1, e_2 \in Q(\mathcal{D})$ such that $Pe_1 \neq Pe_2$, where A is the admittance operator of $\mathcal{N}$ defined in Theorem 1.2 by (1.7).

(ii) *If there exist constants $c > 0$ and $p > 1$ such that for any $x_1, x_2 \in N_a \cap \mathcal{D}$ and any $y_1 \in Zx_1$, $y_2 \in Zx_2$ we have*

$$|\langle y_1 - y_2, x_1 - x_2 \rangle| \geqq c\|x_1 - x_2\|^p, \tag{1.10}$$

then $\mathcal{N}$ is regular on $\mathcal{D}$; moreover,

$$\|Ae_1 - Ae_2\| \leqq c^{-1/(p-1)} \|P(e_1 - e_2)\|^{1/(p-1)} \tag{1.11}$$

and

$$|\langle Ae_1 - Ae_2, e_1 - e_2 \rangle| \leqq c^{-1/(p-1)} \|P(e_1 - e_2)\|^{p/(p-1)}$$

for all $e_1, e_2 \in Q(\mathcal{D})$.

(iii) *If there exist constants* $c>0$ *and* $p>1$ *such that for any* $x_1, x_2 \in N_a \cap \mathscr{D}$ *and any* $y_1 \in Zx_1$, $y_2 \in Zx_2$ *we have*

$$\operatorname{Re}\langle y_1-y_2, x_1-x_2\rangle \geqq c\|x_1-x_2\|^p, \tag{1.12}$$

then $\mathscr{N}$ *is regular on* $\mathscr{D}$, (1.11) *hold and*

$$\operatorname{Re}\langle Ae_1-Ae_2, e_1-e_2\rangle \geqq 0 \tag{1.13}$$

for all $e_1, e_2 \in Q(\mathscr{D})$.

Proof. (i) Since $Px_j = x_j$ for $x_j \in N_a \cap \mathscr{D}$, $j=1, 2$, (1.8) yields

$$\langle Py_1-Py_2, x_1-x_2\rangle \neq 0 \tag{1.14}$$

whenever $x_1 \neq x_2$ and $y_j \in Zx_j$. Let $x_1, x_2 \in N_a \cap \mathscr{D}$, $x_1 \neq x_2$, and suppose that $(PZx_1)\cap(PZx_2) \neq \varnothing$. Then there exist elements $y_1 \in Zx_1$ and $y_2 \in Zx_2$ such that $Py_1 = Py_2$. This, however, contradicts (1.14); hence, $(PZx_1)\cap(PZx_2) = \varnothing$, i.e., PZ is simple on $N_a \cap \mathscr{D}$, and consequently, $\mathscr{N}$ is regular on $\mathscr{D}$ by Theorem 1.2.

Next, if Z is an operator, then (1.8) means that

$$\langle Zx_1-Zx_2, x_1-x_2\rangle \neq 0 \tag{1.15}$$

whenever $x_1, x_2 \in N_a \cap \mathscr{D}$, $x_1 \neq x_2$. Thus, we have as before for such x_1, x_2,

$$\langle PZx_1-PZx_2, x_1-x_2\rangle \neq 0. \tag{1.16}$$

Also, $PZ : N_a \cap \mathscr{D} \to (PZ)(N_a \cap \mathscr{D})$ is a 1–1 onto operator. Choose $y_1, y_2 \in (PZ)(N_a \cap \mathscr{D})$, $y_1 \neq y_2$ and put $x_1 = (PZ)^{-1}y_1$, $x_2 = (PZ)^{-1}y_2$. Since $x_1 \neq x_2$, we have by (1.16),

$$\langle y_1-y_2, (PZ)^{-1}y_1-(PZ)^{-1}y_2\rangle \neq 0. \tag{1.17}$$

Next, let $e_j \in Q(\mathscr{D})$, $j=1, 2$, be such that $Pe_1 \neq Pe_2$; since $Pe_j \in (PZ)(N_a \cap \mathscr{D})$, we can set $y_j = Pe_j$ into (1.17) and get

$$\begin{aligned} 0 &\neq \langle Pe_1-Pe_2, (PZ)^{-1}Py_1-(PZ)^{-1}Py_2\rangle \\ &= \langle Pe_1-Pe_2, Ae_1-Ae_2\rangle. \end{aligned} \tag{1.18}$$

Noting the fact that $PAe = Ae$ for any $e \in Q(\mathscr{D})$, we conclude from (1.18) that (1.9) holds.

(ii) Since (1.10) implies (1.8), $\mathscr{N}$ is regular on $\mathscr{D}$ by (i). If $x_j \in N_a \cap \mathscr{D}$ and $y_j \in Zx_j$, $j=1, 2$, then $Px_j = x_j$ and we obtain from (1.10) by the Schwarz inequality,

$$\begin{aligned} c\|x_1-x_2\|^p &\leqq |\langle y_1-y_2, P(x_1-x_2)\rangle| = |\langle Py_1-Py_2, x_1-x_2\rangle| \\ &\leqq \|Py_1-Py_1\|\cdot\|x_1-x_2\|. \end{aligned}$$

Hence

$$\|x_1-x_2\| \leqq c^{-1/(p-1)}\|Py_1-Py_2\|^{1(p-1)} \tag{1.19}$$

for $x_j \in N_a \cap \mathscr{D}$, $y_j \in Zx_j$, $j=1, 2$.

Now, choose $e_j \in Q(\mathscr{D})$, $j=1, 2$, and put $i_j = Ae_j = (PZ)^- Pe_j \in N_a \cap \mathscr{D}$. By definition of $(PZ)^-$, $Pe_j \in (PZ)i_j$; hence, there exists $y_j \in Zi_j$ such that $Py_j = Pe_j$ for

$j = 1, 2$. Putting $x_j = i_j$ into (1.19), we obtain

$$\|Ae_1 - Ae_2\| \leqq c^{-1/(p-1)} \|Pe_1 - Pe_2\|^{1/(p-1)};$$

the second inequality (1.11) follows from the first one by the Schwarz inequality.

(iii) Since (1.2) implies (1.10), $\mathcal{N}$ is regular on $\mathcal{D}$ and (1.11) hold by proposition (ii). As before, (1.12) implies that

$$\operatorname{Re} \langle Py_1 - Py_2, x_1 - x_2 \rangle \geqq 0 \tag{1.20}$$

for $x_j \in N_a \cap \mathcal{D}$, $y_j \in Zx_j$, $j = 1, 2$. Choose $e_j \in Q(\mathcal{D})$, $j = 1, 2$, and put $i_j = Ae_j = (PZ)^{-} Pe_j \in N_a \cap \mathcal{D}$. Then $Pe_j \in (PZ) i_j$, i.e., there exists $y_j \in Zi_j$ such that $Py_j = Pe_j$. Putting $x_j = i_j$ into (1.20), we get

$$0 \leqq \operatorname{Re} \langle Pe_1 - Pe_2, Ae_1 - Ae_2 \rangle = \operatorname{Re} \langle e_1 - e_2, Ae_1 - Ae_2 \rangle,$$

which concludes the proof.

Remark 1. It is easy to see that inequality (1.9) need not hold if Z is not an operator.

2. Hilbert networks. Let H be a fixed Hilbert space. If G is an oriented graph having the set of branches $\mathcal{B}$ with cardinal $c_2 \leqq \aleph_0$, let $\hat{a} \in (H^{c_2}, H^{c_1})$ be defined as in [1]. Furthermore, if $\mathcal{D} \subset H^{c_2}$, $\mathcal{D} \neq \emptyset$, let $\hat{Z} : \mathcal{D} \to \mathfrak{S}(H^{c_2})$ be a set mapping. Then the ordered pair $\hat{\mathcal{N}} = (\hat{Z}, G)$ is called a Hilbert network.

DEFINITION. Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network, and let $e \in H^{c_2}$; an element $i \in H^{c_2}$ is called a *solution of* $\hat{\mathcal{N}}$ *corresponding to* e if i is a solution of the associated abstract network $\mathcal{N} = (\hat{Z}, \hat{a})$ over H^{c_2} corresponding to e, i.e., if

K_1': there exists a $v \in \hat{Z}i$ such that $v - e \in N_{\hat{a}}^{\perp}$,

K_2': $i \in N_{\hat{a}} \cap \mathcal{D}$.

The network $\hat{\mathcal{N}}$ will be called *regular on* $\mathcal{D}$ if $\mathcal{N}$ is regular on $\mathcal{D}$.

As in [1], we can easily show that i is a solution of $\hat{\mathcal{N}}$ corresponding to $e \in H^{c_2}$ if and only if

K_1^+: there exists a $v \in \hat{Z}i$ such that $\bar{\gamma}^T \cdot (v - e) = 0$

for every $\gamma \in l^{c_2}$ satisfying the equation $a \cdot \gamma = 0$,

K_2^+: $i \in \mathcal{D}$ and $a \cdot i = 0$.

Since $N_{\hat{a}} = \hat{X} H^{c_0}$ and $\hat{X}$ is 1–1 (see Lemma 2.2 in [1]), let $\mathcal{F} \subset H^{c_0}$ be a (uniquely determined) set such that

$$\hat{X}\mathcal{F} = N_{\hat{a}} \cap \mathcal{D}. \tag{2.1}$$

Then we have by (1.5), (1.6),

$$Q(\mathcal{D}) = (\hat{X} H^{c_0})^{\perp} + (\hat{Z}\hat{X}\mathcal{F})^{0}. \tag{2.2}$$

A theorem corresponding to Theorem 1.1 which deals with the existence of a solution of a Hilbert network is merely a paraphrase of the latter and is omitted.

We will need the following.

LEMMA 2.1. *Let X, Y, Z, U be nonempty, and let $A : X \to \mathfrak{S}(Y)$ be a set mapping.*

(i) *If $B : Y \to Z$ is 1–1, then*
BA is simple $\Leftrightarrow A$ is simple.

(ii) *If $C : U \to X$ is 1–1 and onto, then*
AC is simple $\Leftrightarrow A$ is simple.

The proof is an obvious consequence of the definition of a simple operator.

THEOREM 2.1. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network. Then $\hat{\mathcal{N}}$ is regular on $\mathcal{D}$ if and only if the set mapping $\hat{X}^*\hat{Z}\hat{X}$ is simple on $\mathcal{F}$. In this case, the admittance operator $A : Q(\mathcal{D}) \to \hat{X}\mathcal{F}$ of $\hat{\mathcal{N}}$ is given by*

$$A = \hat{X}(\hat{X}^*\hat{Z}\hat{X})^-\hat{X}^*, \tag{2.3}$$

where $(\hat{X}^*\hat{Z}\hat{X})^-$ signifies the quasi-inverse of $\hat{X}^*\hat{Z}\hat{X} : \mathcal{F} \to (\hat{X}^*\hat{Z}\hat{X})\mathcal{F}$.

Proof. Denote $Y : N_{\hat{a}} \cap \mathcal{D} \to \mathcal{D}$, the inverse of $\hat{X} : \mathcal{F} \to N_{\hat{a}} \cap \mathcal{D}$; also, note the fact that $P = \hat{X}\hat{X}^*$ (see [1]). Then we have by Lemma 2.1, $\hat{X}^*\hat{Z}\hat{X}$ is simple on $\mathcal{F} \Leftrightarrow \hat{X}^*\hat{Z} = (\hat{X}^*\hat{Z}\hat{X})Y$ is simple on $N_{\hat{a}} \cap \mathcal{D} \Leftrightarrow P\hat{Z} = \hat{X}\hat{X}^*\hat{Z}$ is simple on $N_{\hat{a}} \cap \mathcal{D}$. This and Theorem 1.2 conclude the proof.

To prove formula (2.3), choose $e \in Q(\mathcal{D})$ and show first that the element $i = \hat{X}(\hat{X}^*\hat{Z}\hat{X})^-\hat{X}^*e = Ae$ is well-defined. If $e \in Q(\mathcal{D})$, then by (2.2) there exist $m \in (\hat{X}H^{c_0})^\perp_{\hat{a}} = N^\perp_{\hat{a}} = N_{\hat{X}^*}$ (nullspace of $\hat{X}^*$, see [1]) and $n \in (\hat{Z}\hat{X}\mathcal{F})^0$ such that $e = m + n$. Since $\hat{X}^*m = 0$, we have $\hat{X}^*e = \hat{X}^*n$, and consequently,

$$\hat{X}^*e \in \hat{X}^*(\hat{Z}\hat{X}\mathcal{F})^0 = (W\mathcal{F})^0, \tag{2.4}$$

where $W = \hat{X}^*\hat{Z}\hat{X}$. Hence

$$q = W^-(\hat{X}^*e) \in \mathcal{F}, \tag{2.5}$$

so that

$$i = \hat{X}q = Ae \in N_{\hat{a}} \cap \mathcal{D} \tag{2.6}$$

by (2.1); thus, i satisfies K_2'.

Next, from (2.6) it follows that $\hat{X}^*\hat{Z}i = \hat{X}^*\hat{Z}\hat{X}q = Wq$; also, (2.5) implies that $\hat{X}^*e \in Wq$, and consequently, $\hat{X}^*e \in \hat{X}^*\hat{Z}i$. This, however, means that there exists an element $v \in \hat{Z}i$ such that $\hat{X}^*e = \hat{X}^*v$. Hence, $\hat{X}^*(v - e) = 0$, i.e., $v - e \in N_{\hat{X}^*} = N^\perp_{\hat{a}}$. Thus, i satisfies K_1' too; consequently, i is the solution of $\hat{\mathcal{N}}$ corresponding to e and our theorem is proved.

THEOREM 2.2. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network and let $W = \hat{X}^*\hat{Z}\hat{X} : \mathcal{F} \to W\mathcal{F}$.*

(i) *If for all $y_1, y_2 \in \mathcal{F}$, $y_1 \neq y_2$ and all $w_1 \in Wy_1$, $w_2 \in Wy_2$, we have*

$$\langle w_1 - w_2, y_1 - y_2 \rangle_{c_0} \neq 0, \tag{2.7}$$

then $\hat{\mathcal{N}}$ is regular on $\mathcal{D}$. If, in addition, $\hat{Z}$ is an operator, then

$$\langle Ae_1 - Ae_2, e_1 - e_2 \rangle_{c_2} \neq 0 \tag{2.8}$$

for all $e_1, e_2 \in Q(\mathcal{D})$ such that $\hat{X}^*e_1 \neq \hat{X}^*e_2$, where A is the admittance operator of $\hat{\mathcal{N}}$.

(ii) *If there exist constants* $c>0$ *and* $p>1$ *such that for any* $y_1, y_2 \in \mathscr{F}$ *and any* $w_1 \in Wy_1$, $w_2 \in Wy_2$ *we have*

$$|\langle w_1 - w_2, y_1 - y_2\rangle_{c_0}| \geqq c\|y_1 - y_2\|_{c_0}^p, \tag{2.9}$$

then $\hat{\mathscr{N}}$ *is regular on* $\mathscr{D}$; *moreover,*

$$\|Ae_1 - Ae_2\|_{c_2} \leqq c^{-1/(p-1)}\|\hat{X}^*(e_1 - e_2)\|_{c_0}^{1/(p-1)} \tag{2.10}$$

and

$$|\langle Ae_1 - Ae_2, e_1 - e_2\rangle_{c_2}| \leqq c^{-1/(p-1)}\|\hat{X}^*(e_1 - e_2)\|_{c_0}^{p/(p-1)}$$

for all $e_1, e_2 \in Q(\mathscr{D})$.

(iii) *If there exist constants* $c>0$ *and* $p>1$ *such that for any* $y_1, y_2 \in \mathscr{F}$ *and any* $w_1 \in Wy_1$, $w_2 \in Wy_2$ *we have*

$$\operatorname{Re}\langle w_1 - w_2, y_1 - y_2\rangle_{c_0} \geqq c\|y_1 - y_2\|_{c_0}^p, \tag{2.11}$$

then $\hat{\mathscr{N}}$ *is regular on* $\mathscr{D}$, (2.10) *hold and*

$$\operatorname{Re}\langle Ae_1 - Ae_2, e_1 - e_2\rangle_{c_2} \geqq 0 \tag{2.12}$$

for all $e_1, e_2 \in Q(\mathscr{D})$.

Proof. Choose $x_1, x_2 \in N_{\hat{A}} \cap \mathscr{D}$ and $z_1 \in \hat{Z}x_1$, $z_2 \in \hat{Z}x_2$. Since $\hat{X}$ is a 1–1 correspondence between $\mathscr{F}$ and $N_{\hat{A}} \cap \mathscr{D}$, there exist uniquely determined elements $y_1, y_2 \in \mathscr{F}$ such that $x_1 = \hat{X}y_1$, $x_2 = \hat{X}y_2$. Thus, we have

$$\begin{aligned}\langle z_1 - z_2, x_1 - x_2\rangle_{c_2} &= \langle z_1 - z_2, \hat{X}(y_1 - y_2)\rangle_{c_2} \\ &= \langle \hat{X}^* z_1 - \hat{X}^* z_2, y_1 - y_2\rangle_{c_0},\end{aligned} \tag{2.13}$$

and

$$\hat{X}^* z_j \in \hat{X}^*\hat{Z}x_j = \hat{X}^*\hat{Z}\hat{X}y_j = Wy_j, \qquad j = 1, 2.$$

Hence, if $x_j \neq x_2$, then $y_1 \neq y_2$, and consequently by (2.13), $\langle z_1 - z_2, x_1 - x_2\rangle_{c_2} \neq 0$. Thus, by Theorem 1.4, $\mathscr{N}$ and $\hat{\mathscr{N}}$, too, is regular. As for (2.8), it suffices to note that $Pe_1 = \hat{X}\hat{X}^* e_1 \neq \hat{X}\hat{X}^* e_2 = Pe_2 \Leftrightarrow \hat{X}^* e_1 \neq \hat{X}^* e_2$; this completes the proof of (i).

The proof of (ii) and (iii) follows immediately from (ii), (iii) in Theorem 1.4 by using the equality (2.13) and the fact (see Lemma 2.2 in [1]) that $\hat{X}$ is an isometry between $\mathscr{F}$ and $N_{\hat{A}} \cap \mathscr{D}$, i.e., $\|x_1 - x_2\|_{c_2} = \|y_1 - y_2\|_{c_0}$ whenever $x_j = \hat{X}y_j$, $j = 1, 2$.

Let us now consider causality in Hilbert networks. For every $T \in R^1$, let $\mathscr{S}_T$ be an orthogonal projection of H^{c_2} into itself, and let the collection $\{\mathscr{S}_T : T \in R^1\}$ be a resolution of identity on H^{c_2} (see [2]), i.e.,

(i) $\mathscr{S}_{T_1} \leqq \mathscr{S}_{T_2}$ for each $T_1 \leqq T_2$,

(ii) for every $T_0 \in R^1$ and $x \in H^{c_2}$, $\mathscr{S}_T x \to \mathscr{S}_{T_0} x$ as $T \to T_0$, $T > T_0$,

(iii) for every $x \in H^{c_2}$, $\mathscr{S}_T x \to 0$ as $T \to -\infty$ and $\mathscr{S}_T x \to x$ as $T \to \infty$.

DEFINITION. Let $\mathscr{D} \subset H^{c_2}$, $\mathscr{D} \neq \varnothing$, let $A : \mathscr{D} \to \mathfrak{S}(H^{c_2})$ be a set mapping, and let $\mathfrak{M} \subset \mathscr{D}$, $\mathfrak{M} \neq \varnothing$; A will be called *causal on* $\mathfrak{M}$ if

$$x_1, x_2 \in \mathfrak{M}, \quad \mathscr{S}_T x_1 = \mathscr{S}_T x_2 \Rightarrow \mathscr{S}_T A x_1 = \mathscr{S}_T A x_2. \tag{2.14}$$

LEMMA 2.2. *Let* $\mathfrak{M} \subset \mathscr{D}$ *be a nonempty set such that* $\mathscr{S}_T \mathfrak{M} \subset \mathfrak{M}$ *for any* $T \in R^1$. *Then* A *is causal on* $\mathfrak{M} \Leftrightarrow \mathscr{S}_T A = \mathscr{S}_T A \mathscr{S}_T$ *on* $\mathfrak{M}$ *for every* $T \in R^1$.

The proof is the same as in the case of an operator and is omitted.

THEOREM 2.3. *Let* $\hat{\mathcal{N}} = (\hat{Z}, G)$ *be a Hilbert network and let* $W = \hat{X}^*\hat{Z}\hat{X} : \mathcal{F} \to W\mathcal{F}$. *For all* $y_1, y_2 \in \mathcal{F}$, $y_1 \neq y_2$ *and all* $w_1 \in Wy_1$, $w_2 \in Wy_2$, *let*

$$\langle w_1 - w_2, y_1 - y_2 \rangle_{c_0} \neq 0. \tag{2.15}$$

Moreover, assume that

(i) *for each* $T \in R^1$, *the projection* $\mathcal{S}_T$ *commutes with* $P = \hat{X}\hat{X}^*$,

(ii) *for each* $T \in R^1$, $\mathcal{S}_T(N_{\hat{a}} \cap \mathcal{D}) \subset N_{\hat{a}} \cap \mathcal{D}$,

(iii) *the set mapping* $P\hat{Z} = \hat{X}\hat{X}^*\hat{Z}$ *is causal on* $N_{\hat{a}} \cap \mathcal{D} = \hat{X}\mathcal{F}$. *Then the admittance operator* A *of* $\hat{\mathcal{N}}$ *is causal on* $Q(\mathcal{D})$.

Proof. Theorem 2.2 shows that, due to (2.15), $\hat{\mathcal{N}}$ is regular on $\mathcal{D}$. Also, from the proof of the same theorem it follows that (2.15) is equivalent to the following condition: for any $x_j \in N_{\hat{a}} \cap \mathcal{D}$ and $z_j \in \hat{Z}x_j$, $j = 1, 2$,

$$\langle z_1 - z_2, x_1 - x_2 \rangle_{c_2} \neq 0 \tag{2.16}$$

whenever $x_1 \neq x_2$.

First, we are going to show that the quasi-inverse $(P\hat{Z})^- : [P\hat{Z}(N_{\hat{a}} \cap \mathcal{D})]^0 \to N_{\hat{a}} \cap \mathcal{D}$ is causal on $[P\hat{Z}(N_{\hat{a}} \cap \mathcal{D})]^0$. By Lemma 2.2 and (ii), $\mathcal{S}_T(P\hat{Z}) = \mathcal{S}_T(P\hat{Z})\mathcal{S}_T$ on $N_{\hat{a}} \cap \mathcal{D}$ for every $T \in R^1$.

Next, choose $T \in R^1$ and $u_1, u_2 \in N_{\hat{a}} \cap \mathcal{D}$ such that $\mathcal{S}_T u_1 \neq \mathcal{S}_T u_2$. Then $\mathcal{S}_T u_j \in N_{\hat{a}} \cap \mathcal{D}$, $j = 1, 2$, due to (ii), and by (2.16),

$$\langle z_1 - z_2, \mathcal{S}_T u_1 - \mathcal{S}_T u_2 \rangle_{c_2} \neq 0 \tag{2.17}$$

for any $z_j \in \hat{Z}\mathcal{S}_T u_j$. Since $\mathcal{S}_T u_j = P\mathcal{S}_T(\mathcal{S}_T u_j)$, (2.17) yields

$$\langle \mathcal{S}_T P z_1 - \mathcal{S}_T P z_2, \mathcal{S}_T u_1 - \mathcal{S}_T u_2 \rangle_{c_2} \neq 0.$$

Hence, for any $u_j \in N_{\hat{a}} \cap \mathcal{D}$, $\mathcal{S}_T u_1 \neq \mathcal{S}_T u_2$ and any $p_j \in \mathcal{S}_T P\hat{Z}\mathcal{S}_T u_j = \mathcal{S}_T P\hat{Z}u_j$,

$$\langle p_1 - p_2, \mathcal{S}_T u_1 - \mathcal{S}_T u_2 \rangle_{c_2} \neq 0. \tag{2.18}$$

Now, choose $v_j \in [PZ(N_{\hat{a}} \cap \mathcal{D})]^0$ such that $\mathcal{S}_T(P\hat{Z})^- v_1 \neq \mathcal{S}_T(P\hat{Z})^- v_2$ and let $u_j = (P\hat{Z})^- v_j$; note that $u_j \in N_{\hat{a}} \cap \mathcal{D}$. Then $v_j \in P\hat{Z}u_j$, so that $\mathcal{S}_T v_j \in \mathcal{S}_T P\hat{Z}u_j$. Consequently, we can put $p_j = \mathcal{S}_T v_j$ into (2.18) and get

$$\langle \mathcal{S}_T v_1 - \mathcal{S}_T v_2, \mathcal{S}_T(P\hat{Z})^- v_1 - \mathcal{S}_T(P\hat{Z})^- v_2 \rangle_{c_2} \neq 0. \tag{2.19}$$

However, (2.19) shows that $\mathcal{S}_T v_1 - \mathcal{S}_T v_2$ cannot be zero; hence we have the implication $v_j \in [P\hat{Z}(N_{\hat{a}} \cap \mathcal{D})]^0$, $j = 1, 2$, $\mathcal{S}_T(P\hat{Z})^- v_1 \neq \mathcal{S}_T(P\hat{Z})^- v_2 \Rightarrow \mathcal{S}_T v_1 \neq \mathcal{S}_T v_2$, i.e., the operator $(P\hat{Z})^-$ is causal on $[P\hat{Z}(N_{\hat{a}} \cap \mathcal{D})]^0$.

To conclude the proof, let $e_1, e_2 \in Q(\mathcal{D})$ be such that $\mathcal{S}_T e_1 = \mathcal{S}_T e_2$. Then $P\mathcal{S}_T e_1 = P\mathcal{S}_T e_2$, so that by (i), $\mathcal{S}_T P e_1 = \mathcal{S}_T P e_2$; thus, by causality of $(P\hat{Z})^-$, $\mathcal{S}_T A e_1 = \mathcal{S}_T(P\hat{Z})^- P e_1 = \mathcal{S}_T(P\hat{Z})^- P e_2 = \mathcal{S}_T A e_2$, i.e., A is causal on $Q(\mathcal{D})$, which is what we wanted to show.

THEOREM 2.4. *In Theorem* 2.3,

(a) *assumption* (ii) *can be replaced by the stronger condition*

$$\mathcal{S}_T\mathcal{D} \subset \mathcal{D} \quad \textit{for each } T \in R^1, \tag{2.20}$$

(b) *assumption* (iii) *can be replaced by the stronger condition "*$\hat{Z}$ *is causal on* $\mathcal{D}$*"*.

Proof. (a) We have

$$\mathcal{S}_T(N_{\hat{a}} \cap \mathcal{D}) \subset (\mathcal{S}_T N_{\hat{a}}) \cap (\mathcal{S}_T \mathcal{D}) \subset (\mathcal{S}_T P H^{c_2}) \cap \mathcal{D}$$

$$= (P\mathcal{S}_T H^{c_2}) \cap \mathcal{D} \subset (PH^{c_2}) \cap \mathcal{D} = N_{\hat{a}} \cap \mathcal{D},$$

i.e., (ii) is satisfied.

(b) If $\hat{Z}$ is causal on $\mathcal{D}$, it is causal on $N_{\hat{a}} \cap \mathcal{D}$, too; thus, by Lemma 2.2, $\mathcal{S}_T \hat{Z} = \mathcal{S}_T \hat{Z} \mathcal{S}_T$ on $N_{\hat{a}} \cap \mathcal{D}$ for each $T \in R^1$. Hence, by (i), $\mathcal{S}_T P \hat{Z} = P \mathcal{S}_T \hat{Z} = P \mathcal{S}_T \hat{Z} \mathcal{S}_T = \mathcal{S}_T P \hat{Z} \mathcal{S}_T$, i.e., (iii) holds; hence the proof.

A quite natural resolution of identity on H^{c_2}, and consequently, a natural concept of causality, can be obtained as follows.

Let $\{E_T : T \in R^1\}$ be a resolution of identity on H, and let $c \leqq \aleph_0$ be fixed. For any $T \in R^1$, define $\mathcal{S}_T : H^c \to H^c$ by

$$\mathcal{S}_T[x_k] = [E_T x_k], \tag{2.21}$$

$x = [x_k] \in H^c$. Then we have the following.

PROPOSITION 1. *The collection* $\{\mathcal{S}_T : T \in R^1\}$ *is a resolution of identity on* H^c.

Proof. Clearly, $\mathcal{S}_T$ is a bounded linear operator from H^c into itself, since for any integer $N > 0$ and $x \in H^c$ we have $\sum_{k=1}^{N} \|E_T x_k\|^2 \leqq \sum_{k=1}^{N} \|E_T\|^2 \cdot \|x_k\|^2 \leqq \|x\|_c^2$.

Moreover, $\mathcal{S}_T$ is a projection. Indeed, for any $T \in R^1$ and $x \in H^c$, $\mathcal{S}_T^2 x = [E_T^2 x_k] = [E_T x_k] = \mathcal{S}_T x$ so that $\mathcal{S}_T^2 = \mathcal{S}_T$.

Also, if $x, y \in H^c$, $\langle \mathcal{S}_T x, y \rangle_c = \sum_k \langle E_T x_k, y_k \rangle = \sum_k \langle x_k, E_T y_k \rangle = \langle x, \mathcal{S}_T y \rangle_c$; thus, $\mathcal{S}_T = \mathcal{S}_T^*$.

Next, let $T_1 \leqq T_2$. Then, for any $x \in H^c$,

$$\langle (\mathcal{S}_{T_2} - \mathcal{S}_{T_1}) x, x \rangle_c = \sum_k \langle E_{T_2} x_k - E_{T_1} x_k, x_k \rangle \geqq 0.$$

Hence, $\mathcal{S}_{T_1} \geqq \mathcal{S}_{T_1}$ and condition (i) in the definition of a resolution of identity is satisfied.

If $c < \aleph_0$, (ii) is trivially satisfied. Thus, let $c = \aleph_0$ and choose $T_0 \in R^1$ and $x \in H^c$. If $\varepsilon > 0$, find $N > 0$ so large that $\sum_{k=N+1}^{\infty} \|x_k\|^2 < \varepsilon^2/8$; by our hypothesis on E_T, there exist intervals $I_i = [T_0, T_i)$, $i = 1, 2, \cdots, N$, such that $\|E_T x_i - E_{T_0} x_i\|^2 < \varepsilon^2/(2N)$ for each $T \in I_i$ and $i = 1, 2, \cdots, N$. Putting $I = \bigcap_{i=1}^{N} I_i$, we have for $T \in I$,

$$\|\mathcal{S}_T x - \mathcal{S}_{T_0} x\|_c^2 = \sum_{i=1}^{\infty} \|E_T x_i - E_{T_0} x_i\|^2 = \sum_{i=1}^{N} \|\cdots\|^2 + \sum_{i=N+1}^{\infty} \|\cdots\|^2$$

$$< N \cdot \frac{\varepsilon^2}{2N} + \sum_{i=N+1}^{\infty} (\|E_T\| + \|E_{T_0}\|)^2 \|x_i\|^2 < \varepsilon^2.$$

Hence, $\mathcal{S}_T x \to \mathcal{S}_{T_0} x$ as $T \to T_0$, $T > T_0$.

The verification of (iii) is similar.

Moreover, we have the following.

PROPOSITION 2. *The collection* $\{\mathcal{S}_T : T \in R^1\}$ *satisfies condition* (i) *in Theorem* 2.3.

Proof. Let projections $\mathcal{S}_T : H^{c_2} \to H^{c_2}$ and $\mathcal{S}'_T : H^{c_0} \to H^{c_0}$ be defined by (2.21). Then

$$\hat{X}\mathcal{S}'_T = \mathcal{S}_T\hat{X}, \qquad \mathcal{S}'_T\hat{X}^* = \hat{X}^*\mathcal{S}_T \tag{2.22}$$

for every $T \in R^1$. Indeed, let $X = [\xi_{ik}]$ be the $c_2 \times c_0$ matrix generating the operator $\hat{X}$ (see [1]), and let $u = [u_k] \in H^{c_0}$; then we have by linearity and continuity of E_T,

$$\hat{X}\mathcal{S}'_T u = [\xi_{ik}] \cdot [E_T u_k] = \left[\sum_k \xi_{ik}(E_T u_k)\right] = \left[E_T\left(\sum_k \xi_{ik} u_k\right)\right] = \mathcal{S}_T\hat{X}u.$$

Hence, the first equation (2.22) holds; the second one follows by taking the adjoints.

Now, we have $P\mathcal{S}_T = \hat{X}\hat{X}^*\mathcal{S}_T = \hat{X}\mathcal{S}'_T\hat{X}^* = \mathcal{S}_T\hat{X}X^* = \mathcal{S}_T P$, what we intended to show.

Summarizing our results, we see that, when dealing with causality defined via projections (2.21), conditions (i)–(iii) in Theorem 2.3 can be replaced by the simple assumption that $\hat{Z}$ is causal on $\mathcal{D}$ and $\mathcal{S}_T\mathcal{D} \subset \mathcal{D}$ for every $T \in R^1$.

On the other hand, condition (2.20) may sometimes be inconvenient when dealing with specific cases of networks. We will show that this condition can be traded for a different assumption; indeed, we have the following theorem.

THEOREM 2.5. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network. Assume that*

(i) *there exist constants $c > 0$ and $p > 1$ such that, for every $T \in R^1$, all $x_j \in N_{\hat{a}} \cap \mathcal{D} = \hat{X}\mathcal{F}$ and $z_j \in \hat{Z}x_j$, $j = 1, 2$, we have*

$$|\langle z_1 - z_2, \mathcal{S}_T(x_1 - x_2)\rangle_{c_2}| \geqq c\|\mathcal{S}_T(x_1 - x_2)\|^p_{c_2}, \tag{2.23}$$

(ii) *for each $T \in R^1$, the projection $\mathcal{S}_T$ commutes with $P = \hat{X}\hat{X}^*$. Then $\hat{\mathcal{N}}$ is regular on $\mathcal{D}$ and the admittance operator A of $\hat{\mathcal{N}}$ is causal on $Q(\mathcal{D})$.*

Proof. First, choose $x_j \in N_{\hat{a}} \cap \mathcal{D}$ and $z_j \in \hat{Z}x_j$, $j = 1, 2$. Since $\mathcal{S}_T(x_1 - x_2) \to x_1 - x_2$ as $T \to \infty$, (2.23) yields by continuity,

$$|\langle z_1 - z_2, x_1 - x_2\rangle_{c_2}| \geqq c\|x_1 - x_2\|^p_{c_2}.$$

Thus, by Theorem 1.4, $\hat{\mathcal{N}}$ is regular on $\mathcal{D}$ and, in particular, the set mapping $P\hat{Z}$ is simple on $N_a \cap \mathcal{D}$.

Next, choose a $T \in R^1$ and $x_j \in N_{\hat{a}} \cap \mathcal{D}$, $j = 1, 2$; *since by (ii)*, $\mathcal{S}_T(x_1 - x_2) = \mathcal{S}_T P(x_1 - x_2) = P\mathcal{S}_T\{\mathcal{S}_T(x_1 - x_2)\}$, (2.23) implies that

$$|\langle w_1 - w_2, \mathcal{S}_T x_1 - \mathcal{S}_T x_2\rangle_{c_2}| \geqq c\|\mathcal{S}_T x_1 - \mathcal{S}_T x_2\|^p_{c_2} \tag{2.24}$$

whenever $w_j \in \mathcal{S}_T P\hat{Z}x_j$, $j = 1, 2$.

Now, let $y_1, y_2 \in [P\hat{Z}(N_{\hat{a}} \cap \mathcal{D})]^0$, and put $x_j = (P\hat{Z})^- y_j \in N_{\hat{a}} \cap \mathcal{D}$, $j = 1, 2$. Then $y_j \in P\hat{Z}x_j$, and consequently, $\mathcal{S}_T y_j \in \mathcal{S}_T P\hat{Z}x_j$. Thus, we can put $w_j = \mathcal{S}_T y_j$, $j = 1, 2$, into (2.24) and get

$$\begin{aligned} &|\langle \mathcal{S}_T y_1 - \mathcal{S}_T y_2, \mathcal{S}_T(P\hat{Z})^- y_1 - \mathcal{S}_T(P\hat{Z})^- y_2\rangle_{c_2}| \\ &\qquad \geqq c\|\mathcal{S}_T(P\hat{Z})^- y_1 - \mathcal{S}_T(P\hat{Z})^- y_2\|^p. \end{aligned} \tag{2.25}$$

Hence, if $\mathcal{S}_T(P\hat{Z})^- y_1 \neq \mathcal{S}_T(P\hat{Z})^- y_2$, (2.25) shows that $\mathcal{S}_T y_1 \neq \mathcal{S}_T y_2$, i.e., the operator $(P\hat{Z})^-$ is causal on $[P\hat{Z}(N_{\hat{a}} \cap \mathcal{D})]^0$.

Finally, let $e_1, e_2 \in Q(\mathscr{D})$ and let $\mathscr{S}_T e_1 = \mathscr{S}_T e_2$. Then, by (ii), $\mathscr{S}_T P e_1 = P\mathscr{S}_T e_1 = P\mathscr{S}_T e_2 = \mathscr{S}_T P e_2$, and consequently, by causality of $(P\hat{Z})^-$,

$$\mathscr{S}_T A e_1 = \mathscr{S}_T (P\hat{Z})^- P e_1 = \mathscr{S}_T (P\hat{Z})^- P e_2 = \mathscr{S}_T A e_2,$$

which concludes the proof.

Remark 2. If inequality (2.23) is satisfied even for all $x_j \in \mathscr{D}$, $z_j \in \hat{Z}x_j$ and all $T \in R^1$, then assumption (i) in Theorem 2.5 is trivially satisfied, too.

From Theorem 2.5 we easily get the following result.

COROLLARY. *Let the resolution of identity* $\{\mathscr{S}_T : T \in R^1\}$ *be defined by* (2.21). *Furthermore, let* $\hat{\mathscr{N}} = (\hat{Z}, G)$ *be a Hilbert network and let* $W = \hat{X}^* \hat{Z} \hat{X} : \mathscr{F} \to W\mathscr{F}$. *Assume that there exist constants* $c > 0$ *and* $p > 1$ *such that*

$$|\langle w_1 - w_2, \mathscr{S}'_T(y_1 - y_2)\rangle_{c_0}| \geqq c\|\mathscr{S}'_T(y_1 - y_2)\|^p_{c_0} \tag{2.26}$$

for all $T \in R^1$, $y_j \in \mathscr{F}$ *and* $w_j \in Wy_j$, $j = 1, 2$, *where* $\mathscr{S}'_T : H^{c_0} H^{c_0}$ *is defined by* (2.21). *Then* $\hat{\mathscr{N}}$ *is regular on* $\mathscr{D}$ *and the admittance operator* A *of* $\hat{\mathscr{N}}$ *is causal on* $Q(\mathscr{D})$.

Proof. By Proposition 2, $\mathscr{S}_T P = P\mathscr{S}_T$ for every $T \in R^1$, i.e., condition (ii) in Theorem 2.5 is satisfied. Also, by (2.22),

$$\hat{X}\mathscr{S}'_T = \mathscr{S}_T \hat{X}. \tag{2.27}$$

Choose $T \in R^1$, $x_j \in N_{\hat{a}} \cap \mathscr{D} = \hat{X}\mathscr{F}$ and $z_j \in \hat{Z}x_j$, $j = 1, 2$. Then there exists uniquely determined $y_j \in \mathscr{F}$ such that $x_j = \hat{X}y_j$ and we have by (2.27),

$$\begin{aligned} \nu = |\langle z_1 - z_2, \mathscr{S}_T(x_1 - x_2)\rangle_{c_2}| &= |\langle z_1 - z_2, \hat{X}\mathscr{S}'_T(y_1 - y_2)\rangle_{c_2}| \\ &= |\langle \hat{X}^* z_1 - \hat{X}^* z_2, \mathscr{S}'_T(y_1 - y_2)\rangle_{c_0}|. \end{aligned}$$

However, since $\hat{X}^* z_j \in \hat{X}^* \hat{Z} x_j = \hat{X}^* \hat{Z} \hat{X} y_j = W y_j$, we have by (2.26), $\nu \geqq c\|\mathscr{S}'_T(y_1 - y_2)\|^p_{c_0}$.

On the other hand, since $\hat{X}$ is an isometry between $N_{\hat{a}}$ and H^{c_0}, it follows by (2.27) that

$$\|\mathscr{S}_T(x_1 - x_2)\|_{c_2} = \|\mathscr{S}_T \hat{X}(y_1 - y_2)\|_{c_2} = \|\hat{X}\mathscr{S}_T(y_1 - y_2)\|_{c_2} = \|\mathscr{S}'_T(y_1 - y_2)\|_{c_0}.$$

Hence, inequality (2.23) is satisfied and Theorem 2.5 concludes the proof.

Let us now consider an example of a specific network.

Example. Let G be an oriented graph having the set of branches $\mathscr{B}$ with cardinal $c_2 \leqq \aleph_0$ and assume that H is the real space $L_2[0, \tau]$, $\tau > 0$. Furthermore, let us make the following assumptions:

(a) For every index j, let r_j be a set mapping from R^1 into $\mathfrak{S}(R^1)$; assume that there exists a constant $\alpha > 0$ such that for any $\sigma_1, \sigma_2 \in R^1$ and any $\omega_1 \in r_j(\sigma_1)$, $\omega_2 \in r_j(\sigma_2)$ we have

$$\alpha(\sigma_1 - \sigma_2)^2 \leqq (\omega_1 - \omega_2)(\sigma_1 - \sigma_2), \qquad j = 1, 2, \cdots. \tag{2.28}$$

Moreover, let there exist an integer $N > 0$ and a constant $\beta > 0$ such that, for all $j > N$, $\sigma \in R^1$ and $\omega \in r_j(\sigma)$ we have

$$|\omega| \leqq \beta|\sigma|. \tag{2.29}$$

(b) For every pair of indices i, k, let $L_{ik}(t)$ be a real function having a derivative everywhere on $[0, \tau]$. Assume that there exists $\gamma > 0$ such that

$$|L_{ik}(t)| \leqq \gamma, \qquad |L'_{ik}(t)| \leqq \gamma \tag{2.30}$$

for all $t \in [0, \tau]$. Moreover, let $[L_{ik}(t)]$ and $[L'_{ik}(t)]$ be ribbon, symmetric and positive semidefinite $c_2 \times c_2$ matrices on $[0, \tau]$ (see [1]).

(c) For every pair of indices i, k, let $S_{ik}(t)$ be a real function having a bounded derivative on $[0, \tau]$ and let a $\delta > 0$ exist such that

$$|S_{ik}(t)| \leqq \delta \tag{2.31}$$

for all $t \in [0, \tau]$. Assume that $[S_{ik}(t)]$ and $-[S'_{ik}(t)]$ are ribbon, symmetric and positive semidefinite matrices on $[0, \tau]$.

Finally, let $i_0 \in l^{c_2}$ be a vector satisfying the equation $d^T \cdot i_0 = 0$, where d is the structural matrix of G.

Let $\mathscr{D} \subset L_2^{c_2}[0, \tau]$ be the set of all c_2-vectors $x = [x_k(t)]$ such that each $x_k(t)$ is absolutely continuous on $[0, \tau]$, $x' = [x'_k(t)] \in L_2^{c_2}[0, \tau]$ and $x(0) = i_0$.

Let the set mapping $\hat{Z}$ be defined on $\mathscr{D}$ by

$$\hat{Z}y = [Z_{ik}] \cdot [y_k], \tag{2.32}$$

where each set mapping Z_{ik} is defined by

$$(Z_{ik}\xi)(t) = (L_{ik}(t)\xi)' + \eta_{ik} + S_{ik}(t) \int_0^t \xi(\sigma)\, d\sigma. \tag{2.33}$$

Here $\eta_{ik} \in r_{ik}(\xi(t))$, $r_{ik} = 0$ for $i \neq k$ and $r_{ii} = r_i$, $i = 1, 2, \cdots$.

Let us consider the network $\hat{\mathscr{N}} = (\hat{Z}, G)$. We are going to show that, under the assumptions made, $\hat{\mathscr{N}}$ is regular on $\mathscr{D}$. Thus, we consider an L, R, C-network with time-varying inductors and capacitors, and with time-invariant nonlinear multivalued resistors, whose initial current regime is described by the vector i_0. Note that, due to condition (2.28), the characteristic r_j of a resistor may look as indicated in Fig. 1.

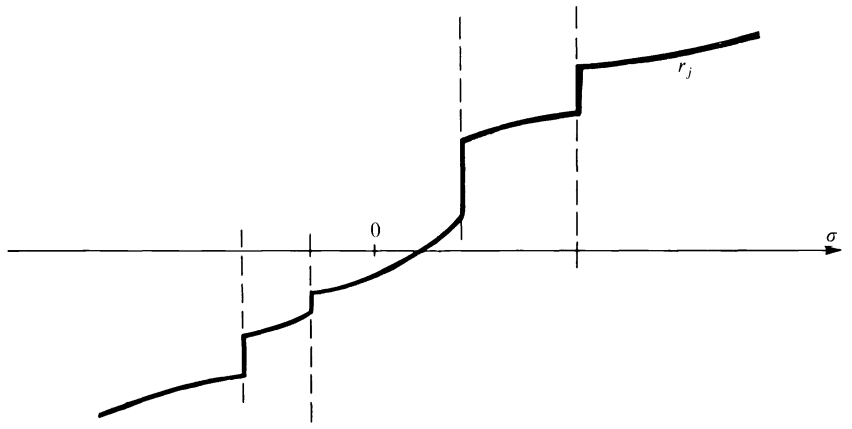


FIG. 1.

To prove our claim, define operators V_{ik}, U_{ik} and the set mappings R_{ik} by

$$(V_{ik}\xi)(t) = (L_{ik}(t)\xi)', \qquad (U_{ik}\xi)(t) = S_{ik}(t) \int_0^t \xi(\sigma)\, d\sigma,$$
$$(R_{ik}\xi)(t) = r_{ik}(\xi(t)), \qquad t \in [0, \tau]; \tag{2.34}$$

also, let the operators $\hat{V}$, $\hat{U}$ and the set mapping $\hat{R}$ be defined on $\mathscr{D}$ by

$$\hat{V}y = [V_{ik}] \cdot [y_k], \quad \hat{U}y = [U_{ik}] \cdot [y_k], \quad \hat{R}y = [R_{ik}] \cdot [y_k]. \tag{2.35}$$

Then, clearly $\hat{Z} = \hat{V} + \hat{R} + \hat{U}$.

Since

$$\hat{V}y = \{[L_{ik}(t)] \cdot [y_k]\}' = [L'_{ik}(t)] \cdot [y_k] + [L_{ik}(t)] \cdot [y'_k],$$

it follows from (2.30) and the fact that $[L_{ik}(t)]$ is a ribbon matrix that $\hat{V}y \in L_2^{c_2}[0, \tau]$ for each $y \in \mathscr{D}$.

Furthermore, in [1] we have shown that, for any x, $y \in L_2^{c_2}[0, \tau]$, $x = [x_k]$, $y = [y_k]$, we have $\langle x, y \rangle_{c_2} = \int_0^\tau x^T(\sigma) \cdot y(\sigma)\, d\sigma$. Thus, if $x_1, x_2 \in \mathscr{D}$, we can write

$$J = \langle \hat{V}x_1 - \hat{V}x_2, x_1 - x_2 \rangle_{c_2} = \int_0^\tau (x_1 - x_2)^T \cdot \{L \cdot (x_1 - x_2)\}'\, d\sigma$$

$$= [(x_1 - x_2)^T \cdot L \cdot (x_1 - x_2)]_0^\tau - \int_0^\tau (x_1' - x_2')^T \cdot L \cdot (x_1 - x_2)\, d\sigma.$$

Also,

$$J = \int_0^\tau (x_1 - x_2)^T \cdot L' \cdot (x_1 - x_2)\, d\sigma + \int_0^\tau (x_1 - x_2)^T \cdot L \cdot (x_1' - x_2')\, d\sigma.$$

Since $x_1(0) = x_2(0) = i_0$, we get due to the symmetry of L,

$$J = \tfrac{1}{2}(x_1 - x_2)^T(\tau) \cdot L(\tau) \cdot (x_1 - x_2)(\tau) + \tfrac{1}{2} \int_0^\tau (x_1 - x_2)(\sigma) \cdot L'(\sigma) \cdot (x_1 - x_2)(\sigma)\, d\sigma. \tag{2.36}$$

Hence, due to positive semidefiniteness of L and L', $J \geqq 0$.

Similarly, (2.29) shows that $\hat{R}x \subset L_2^{c_2}[0, \tau]$ whenever $x \in \mathscr{D}$. Also, routine calculations confirm that (2.28) implies that

$$\langle w_1 - w_2, x_1 - x_2 \rangle_{c_2} \geqq \alpha \|x_1 - x_2\|_{c_2}^2 \tag{2.37}$$

whenever $x_1, x_2 \in \mathscr{D}$ and $w_1 \in \hat{R}x_1$, $w_2 \in \hat{R}x_2$.

Finally, in [1] we have shown that $\hat{U}x \in L_2^{c_2}[0, \tau]$ whenever $x \in \mathscr{D}$ and also that

$$\langle \hat{U}x, x \rangle_{c_2} = \tfrac{1}{2} y^T(\tau) \cdot S(\tau) \cdot y(\tau) - \tfrac{1}{2} \int_0^\tau y^T(\omega) \cdot S'(\omega) \cdot y(\omega)\, d\omega \geqq 0, \tag{2.38}$$

where $y(t) = \int_0^t x(\sigma)\, d\sigma$.

Thus, by (2.36), (2.37) and (2.38), the set mapping $\hat{Z}$ satisfies the condition

$$\langle w_1 - w_2, x_1 - x_2 \rangle_{c_2} \geqq \alpha \|x_1 - x_2\|_{c_2}^2 \tag{2.39}$$

whenever $x_1, x_2 \in \mathscr{D}$ and $w_1 \in \hat{Z}x_1$, $w_2 \in \hat{Z}x_2$. Since $N_a \cap \mathscr{D} \subset \mathscr{D}$, (2.39) holds for all $x_1, x_2 \in N_a \cap \mathscr{D}$, too. Hence by Theorem 1.4 (iii), $\hat{\mathscr{N}}$ is regular on $\mathscr{D}$, what we wished to show.

Also, by (1.11), the admittance operator A of $\hat{\mathscr{N}}$ satisfies the Lipschitz condition with constant α^{-1}.

It is easy to show that the operator A is causal on $Q(\mathscr{D})$ provided the causality is defined via truncations on $L_2[0, \tau]$.

Indeed, let η be a continuous increasing function on R^1 which maps R^1 onto $(0, \tau)$. For every $T \in R^1$ and $u \in L_2[0, \tau]$, let

$$(E_T u)(t) = \begin{cases} u(t) & \text{for } t \in [0, \eta(T)], \\ 0 & \text{for } t \in (\eta(T), \tau]. \end{cases}$$

Then $\{E_T : T \in R^1\}$ is a resolution of identity on $L_2[0, \tau]$ and $\{\mathscr{S}_T : T \in R^1\}$, defined by (2.21), is a resolution of identity on $L_2^{c_2}[0, \tau]$ by Proposition 1.

Note that, in this context, causality of an operator M clearly means that $x_1(t) = x_2(t)$ on $[0, T']$, $T' < \tau \Rightarrow (Mx_1)(t) = (Mx_2)(t)$ on $[0, T']$.

Next, by Proposition 2, each projection $\mathscr{S}_T$ commutes with P, i.e., condition (ii) in Theorem 2.5 is satisfied.

Furthermore, it is clear that, for every $T_0 \in R^1$ and $x, z \in H^{c_2}$,

$$\langle z, \mathscr{S}_{T_0} x \rangle_{c_2} = \int_0^{\eta(T_0)} z^T(\sigma) \cdot x(\sigma)\, d\sigma. \tag{2.40}$$

However, this together with (2.36), (2.37) and (2.38) shows immediately that our set mapping $\hat{Z}$ satisfies the condition (2.23) for all $x_1, x_2 \in \mathscr{D}$; hence, by Remark 2 and Theorem 2.5, A is causal on $Q(\mathscr{D})$, what we wanted to prove.

Concluding the paper, let us make a few comments on our results. As it is apparent from Theorem 2.2, quite simple conditions guarantee the uniqueness of a network solution. On the other hand, condition (1.4) giving the existence, involves the set $Q(\mathscr{D})$ which is hard to describe. In particular, it would be useful to find conditions under which $Q(\mathscr{D}) = H^{c_2}$, i.e., when a Hilbert network possesses a solution for any vector of voltages in H^{c_2}.

From (1.6) it follows that $Q(\mathscr{D}) = H^{c_2}$ if and only if

$$[P\hat{Z}(N_{\hat{a}} \cap \mathscr{D})]^0 = N_{\hat{a}}. \tag{2.41}$$

It is clear that (2.41) does not hold in general, unless we make additional assumptions on $\hat{Z}$ and $\mathscr{D}$. Fortunately, it turns out that (2.41) is satisfied, if $\mathscr{D} \supset N_{\hat{a}}$ and $P\hat{Z}$ is a maximal monotone and coercive mapping.

Indeed, let $\tilde{\mathscr{H}}$ be a real Hilbert space and let $M : \tilde{\mathscr{H}} \to \mathfrak{S}(\tilde{\mathscr{H}})$ be a set-mapping; as known, M is called monotone on $\tilde{\mathscr{H}}$ if for all $x_1, x_2 \in \mathscr{H}$ and $z_1 \in Mx_1$, $z_2 \in Mx_2$,

$$\langle z_1 - z_2, x_1 - x_2 \rangle \geqq 0. \tag{2.42}$$

Moreover, M is called maximal monotone on $\tilde{\mathscr{H}}$ if $M' : \tilde{\mathscr{H}} \to \mathfrak{S}(\tilde{\mathscr{H}})$ monotone, $M'x \supset Mx$ for all $x \in \tilde{\mathscr{H}}$ implies that $M' = M$.

We say that M is coercive if

$$\lim_{a \to \infty} a^{-1} \inf \{\langle z, x \rangle : z \in Mx, x \in \tilde{\mathscr{H}}, \|x\| \geqq a\} = \infty. \tag{2.43}$$

Then, as proved by Rockafellar (see [5, Thm. 9] and [6, Thm. 3]), we have the following assertion: if M is maximal monotone and coercive, then $(M\tilde{\mathscr{H}})^0 = \tilde{\mathscr{H}}$.

Now, since $N_{\hat{a}}$ is closed in H^{c_2}, and thus a Hilbert space of its own right, and $\hat{X}$ is a norm preserving isomorphism between H^{c_0} and $N_{\hat{a}}$, we immediately get the following result on Hilbert networks.

THEOREM 2.6. *Let H be a real Hilbert space and let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network. Assume that*

(i) $\mathcal{D} \supset N_{\hat{a}}$,

(ii) *the set mapping $W = \hat{X}^* \hat{Z} \hat{X} : H^{c_0} \to WH^{c_0}$ is maximal, monotone and coercive on H^{c_0}. Then for any $e \in H^{c_2}$, the network $\hat{\mathcal{N}}$ possesses a solution.*

COROLLARY. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be such that*

(i)′ $\mathcal{D} \supset N_{\hat{a}}$,

(ii)′ *W satisfies the inequality* (2.11) *for all $y_1, y_2 \in H^{c_0}$ and is maximal on H^{c_0}. Then for any $e \in H^{c_2}$, $\hat{\mathcal{N}}$ possesses a unique solution i.*

(The proof is obvious).

Another theorem on existence can be derived from a theorem by Minty (see Theorem 2 in [4]), but we omit the details.

REFERENCES

[1] V. DOLEZAL, *Hilbert networks I*, this Journal, 12 (1974), pp. 755–778.

[2] R. SAEKS, *Causality in Hilbert space*, SIAM Rev., 12 (1970), pp. 357–383.

[3] G. J. MINTY, *Monotone networks*, Proc. Royal Soc. London Ser. A, 257 (1960), pp. 194–212.

[4] ———, *On the solvability of nonlinear functional equations of "monotonic" type*, Pacific J. Math., 14 (1964), pp. 249–255.

[5] R. T. ROCKAFELLAR, *Convex functions, monotone operators and variational inequalities*, Theory and Applications of Monotone Operators, A. Ghizzetti, ed., Proc. NATO Adv. Study Inst., Venice, Italy, Edizioni Orderisi, Gubbio, 1969, pp. 35–65.

[6] ———, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75–88.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

EFFICIENTLY CONVERGING MINIMIZATION METHODS BASED ON THE REDUCED GRADIENT*

DANIEL GABAY AND DAVID G. LUENBERGER†

Abstract. This paper presents three computational methods which extend to nonlinearly constrained minimization problems the efficient convergence properties of, respectively, the method of steepest descent, the variable metric method, and Newton's method for unconstrained minimization. Development of the algorithms is based on use of the implicit function theorem to essentially convert the original constrained problem to an unconstrained one. This approach leads to practical and efficient algorithms in the framework of Abadie's generalized reduced gradient method. To achieve efficiency, it is shown that it is necessary to construct a sequence of approximations to the Lagrange multipliers of the problem simultaneously with the approximations to the solution itself. In particular, the step size of each iteration must be determined by a linesearch for a minimum of an approximate Lagrangian function.

1. Introduction. Many computational methods have been proposed to find the minimum of a real-valued function $f(x)$ over the n-dimensional real space R^n. When both the values of the function f and its derivatives $\partial f/\partial x_i$ for $i=1,\cdots,n$ are available at every point x, gradient-related techniques are generally favored. These schemes iteratively construct, from an initial point x^0, a monotonically improving sequence of approximate solutions x^k according to a recurrence formula of the form

$$x^{k+1}=x^k-\alpha_k p^k, \qquad k=0,1,\cdots, \tag{*}$$

where p^k is a direction determined on the basis of the gradient $\nabla f(x^k)$, and α_k is a positive scalar chosen to achieve a descent in the value of the objective:

$$f(x^{k+1})\leqq f(x^k).$$

The parameter α_k controls the size of the step k and influences the convergence properties of the algorithm (*).

The speed of convergence is strongly dependent on the step size choice. This dependence is well understood in the case of the method of steepest descent, which takes for p^k the direction of the gradient $\nabla f(x^k)$ itself, as originally proposed by Cauchy [6]. The best performance is obtained for the optimal steepest descent method in which α_k is chosen to achieve a local minimum of f along $\nabla f(x^k)$. As first exhibited by Kantorovitch [17], the sequence $\{x^k\}$ converges to x^* linearly, i.e., at least as fast as a geometric progression. The sharpest possible estimate for the ratio of this progression is asymptotically given by $(M-m/M+m)^2$, where M and m are, respectively, the largest and smallest eigenvalues of F^*, the matrix of second order partial derivatives of f at x^*. We refer to this as the *natural rate of convergence* of the problem. It represents the fastest possible speed for a steepest descent algorithm. It provides, therefore, a standard by which the performance of other schemes can be evaluated through comparison with this efficient natural

* Received by the editors April 23, 1973, and in revised form December 17, 1974.

† Department of Engineering-Economic Systems, Stanford University, Stanford, California 94305. The first author is now with Centre National de la Recherche Scientifique, Paris, France. This research was supported by the National Science Foundation under Grants GK32870 and GK 29237.

rate. And, indeed, the rate of convergence of other gradient-related algorithms can be expressed relative to this ratio [21].

In practice, many problems either arise or can be formulated as constrained optimization problems. In this case, the minimum of the function f is sought among the values it takes while the variable point x is restricted to a given subset $\mathcal{S}$. This subset is called the feasible region and is assumed to be described by a finite number of constraint equalities and inequalities. Without lack of generality, we can formally write

$$\mathcal{S} = \{x \in R^n | h_i(x) = 0,\ i = 1, 2, \cdots, m;\ a_j \leqq x_j \leqq b_j,\ j = 1, 2, \cdots, n\} \quad \text{with } m \leqq n,$$

where a_j and b_j are real numbers and can take the values $-\infty$ and $+\infty$.

One of the most successful gradient-related methods to solve this nonlinearly constrained problem is Abadie's generalized reduced gradient algorithm (GRG) [1]. As an extension to the nonlinear case of the upper bounding simplex method for linear programming [8], this method introduces a partition of the variables into *m basic variables*, denoted by the vector $x_B = (x_{B_1}, \cdots, x_{B_m})$, and $n - m$ remaining *independent variables* forming the vector $x_R = (x_{R_1}, \cdots, x_{R_{n-m}})$, such that

$$a_{B_i} < x_{B_i} < b_{B_i}, \qquad i = 1, \cdots, m,$$

$$a_{R_j} \leqq x_{R_j} \leqq b_{R_j}, \qquad j = 1, \cdots, n - m.$$

The independent variables are changed on the basis of the reduced gradient [28], [11], obtained by "pricing-out" the nonbasic components of the gradient $\nabla f(x)$, as the reduced costs are obtained in the simplex method. When the constraints h_i are nonlinear, Abadie's proposal consists in decomposing each iteration in two phases. Starting from a feasible point, a move is performed along a direction tangent to $\mathcal{S}$ based on the reduced gradient. It is followed by a restoration move, achieved by adjusting the m basic variables in order to satisfy the constraint equations. The resulting algorithm [2] is ranked first in efficiency among all available techniques in the comparison studies conducted by Colville on a series of test problems [7], [3]. The selection of the size of the tangent move is, in large part, responsible for the current complexity of the code, since, if this parameter is too large, the restoration may be impossible or may lead to a feasible point which does not constitute an improvement of the objective function. Therefore, following this approach, one is often forced to try several step lengths for the first phase in order to obtain a satisfactory point at the end of the second phase. Such repeated trials significantly increase the computation time and, even though a procedure is developed so as to insure convergence, the rate of convergence may be far from optimal.

The object of this work is to propose generalized gradient related methods for nonlinearly constrained problems which properly extend the efficient convergence of the optimal methods of the unconstrained case. We restrict our analysis to problems without bound constraints on the variables and refer to [15] for the alterations necessary to treat the general case. The implicit function theorem [10] provides a natural and convenient framework to study the appropriate restrictions of the original methods of unconstrained minimization to the constraint set $\mathcal{S}$ itself (rather than to the subspace tangent to $\mathcal{S}$). The theorem

conceptually allows one to express the basic variables as functions of the independent variables, thus converting the original problem to an unconstrained one:

$$\text{Minimize } \phi(x_R) = f(x_B(x_R), x_R).$$

Solving this reduced problem by the gradient-related methods of unconstrained minimization leads, in the original space, to schemes in which the independent variables are moved on the basis of the gradient of the reduced function, which turns out to be Wolfe's reduced gradient. The basic variables are altered correspondingly to maintain feasibility. We thus extend the efficient convergence properties of the method of steepest descent, the variable metric method, and Newton's method to nonlinearly constrained minimization.

These ideal extended gradient-related methods cannot be implemented exactly, since it is not possible in practice to generate arcs along $\mathcal{S}$. We are led to consider more practical schemes which accurately approximate the arcs of the ideal methods.

In § 3 we define an implementable generalized reduced steepest descent algorithm, combining at each iteration a tangent phase and a restoration phase. To achieve efficiency, it is shown that it is necessary to construct a sequence of approximations to the Lagrange multipliers of the problem simultaneously with the approximations to the solution itself. Each combined step then accurately approximates the arc of the ideal scheme, provided that the step size is determined by a linesearch for a minimum of the approximate Lagrangian, a procedure which has been tentatively proposed on other occasions [19], [23].

In §§ 4 and 5, we show how the framework of the reduced unconstrained problem can establish guidelines to define practical and efficient algorithms extending, respectively, the variable metric method [14] and Newton's method [18] to nonlinearly constrained minimization and inheriting their superlinear and second order rates of convergence.

2. Notation. We denote n-dimensional vectors by notation such as $x = (x_1, \cdots, x_n)$. Unless otherwise specified, they are regarded as column vectors. For any matrix A, ${}^tA$ denotes its transpose.

Given a function $f: R^n \to R$, its gradient at x is the n-row vector $\nabla f(x) = ((\partial f/\partial x_1)(x), \cdots, (\partial f/\partial x_n)(x))$. For any subset $K \subset \{1, \cdots, n\}$, we denote by $\nabla_K f(x)$ the vector of components $(\partial f/\partial x_i)(x)$ with $i \in K$. We denote the matrix of the second order partial derivatives, the Hessian, by $F(x)$.

For a mapping $h: R^n \to R^m$ with components h_i, $\nabla h(x)$ represents the $m \times n$ Jacobian matrix with element (i, j) given by $(\partial h_i/\partial x_j)(x)$. The second derivative of h is best regarded as the m-tuple $H = (H_1, H_2, \cdots, H_m)$, where H_i is the Hessian of h_i. We denote the m-tuple of the associated quadratic forms by ${}^tx \cdot H \cdot x = ({}^txH_1x, {}^txH_2x, \cdots, {}^txH_mx)$ for any $x \in R^n$. We define the operator $\times$, associating an element λ of R^m and an m-tuple H of $R^{m \times n \times n}$ into an element of $R^{n \times n}$, by

$$\lambda \times H = \lambda_1 H_1 + \cdots + \lambda_m H_m;$$

for any $x \in R^n$, we have

$${}^tx(\lambda \times H)x = \lambda({}^tx \cdot H \cdot x).$$

We denote by (P) the problem

$$\text{(P)} \qquad \begin{aligned} &\text{Minimize } f(x) \\ &\text{Subject to } h(x) = 0. \end{aligned}$$

3. The generalized reduced steepest descent method for nonlinearly constrained minimization.

3.1. The idealized reduced gradient method. The implicit function theorem has historically played a fundamental role in the theory of constrained minimization problems, since it provides the tool required to establish the existence of Lagrange multipliers. Basically it reduces problem (P) to an unconstrained minimization (at least locally) by solving the implicit constraint equations. We assume in all the following that f and h_i are twice continuously differentiable and possess bounded third order derivatives.

Assuming that the constraints are regular, i.e., that the gradient vectors $\nabla h_1(x), \cdots, \nabla h_m(x)$ are linearly independent for all x, then the implicit function theorem guarantees the local existence of a mapping $\psi : R^{n-m} \to R^m$ such that $x_B = \psi(x_R)$. It is well known that

$$\nabla\psi(x) = -\nabla_B h(x)^{-1} \nabla_R h(x), \tag{1}$$

where the argument x stands indifferently for the independent variables $x_R \in R^{n-m}$ and for the n-tuple $(\psi(x_R), x_R)$; and it can be shown that the second derivative of ψ is given by

$$\Psi(x) = -\nabla_B h(x)^{-1} \times ['T(x) \cdot H(x) \cdot T(x)], \tag{2}$$

where $T(x)$ is the $n \times (n-m)$ matrix

$$T(x) = \begin{bmatrix} -\nabla_B h(x)^{-1} \nabla_R h(x) \\ I_{n-m} \end{bmatrix}. \tag{3}$$

This matrix represents the mapping of R^n onto $\mathcal{T}(x)$, the tangent subspace to $\mathcal{S}$ at x. We can view (P) in terms of the reduced problem (R) in R^{n-m}:

$$\text{(R)} \qquad \text{Minimize } \phi(x_R) = f[\psi(x_R), x_R],$$

defined at least in a neighborhood of a solution point x^* of (P). The gradient of ϕ is called the *reduced gradient* and its transpose is an $(n-m)$-dimensional column vector denoted by $r(x)$. The chain rule for derivatives leads to

$$\nabla\phi(x) = \nabla_B f(x) \nabla\psi(x) + \nabla_R f(x).$$

Using (1), we get

$$'r(x) = \nabla_R f(x) - \nabla_B f(x)[\nabla_B h(x)]^{-1} \nabla_R h(x). \tag{4}$$

Among the methods of steepest descent for solving (R), the optimal steepest descent algorithm provides the best performing algorithm. It consists of a series of moves in R^{n-m} along the reduced gradients at the successive iterates x_R^k. The size of each step is determined by a linesearch along $r^k = r(x_R^k)$ for a local minimum point of ϕ.

In practice, it is usually not possible to achieve explicitly the elimination of the dependent variables x_B and it is therefore necessary to solve problem (P) in the original space R^n. However, the above study of the reduced problem (R) shows that the most natural algorithm consists, for its kth iteration, starting from the feasible point $x^k = (x_B^k, x_R^k)$, in moving the independent variables x_R along the reduced gradient r^k, while maintaining feasibility by an alteration of the basic variables x_B. This defines an arc $x^k(\beta)$ on $\mathcal{S}$ emanating from x^k. The projection of this arc on the subspace R of the independent variables, parallel to the basic subspace B, is the straight line in the (negative) direction of r^k; hence

$$x_R^k(\beta) = x_R^k - \beta r^k. \tag{5}$$

To satisfy the constraint equations, the basic variables must satisfy

$$x_B^k(\beta) = \psi(x_R^k - \beta r^k).$$

Assuming that the constraints are uniformly regular (i.e., that there exists a scalar $\gamma > 0$ such that $\|\nabla_B h(x)\| \geqq \gamma$ for all x), we can write

$$x_B^k(\beta) = \psi(x_R^k) - \beta \nabla\psi^k r^k + (\beta^2/2)' r^k \cdot \Psi^k \cdot r^k + O(\|r^k\|^3).$$

Using (1) and (2) we obtain

$$x_B^k(\beta) = x_B^k - \beta(-(\nabla_B h^k)^{-1}\nabla_R h^k) r^k - \beta^2 (\nabla_B h^k)^{-1} q^k + O(\|r^k\|^3), \tag{6}$$

where q^k is the m-dimensional column vector of components

$$q_i^k = \tfrac{1}{2} r^{kt} T^k H_i(x^k) T^k r^k. \tag{7}$$

This reduced gradient method, illustrated in Fig. 1, belongs to a class of techniques for nonlinearly constrained problems proposed by Altman [4] and was

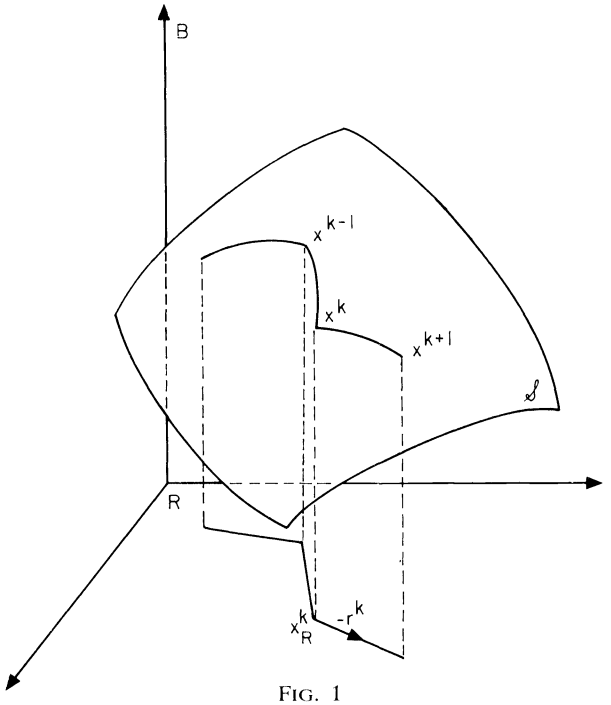


FIG. 1

presented first in [21] in this specific set-up. To efficiently extend the optimal steepest descent method, the step size parameter β_k must be chosen to achieve a local minimum of f along the curve $x^k(\beta)$:

$$\beta_k = \text{Argmin}\,\{f[x^k(\beta)] \mid \beta \geqq 0\}. \tag{8}$$

The speed of convergence of the sequence $\{x^k\}$ to x^* is then asymptotically given by the Kantorovitch-ratio $(M-m/M+m)^2$, where M and m are the extreme eigenvalues of Φ^*, the Hessian of ϕ at x^*. This defines the natural rate of convergence for reduced gradient methods, since this algorithm ideally extends the efficient performance of steepest descent methods for the unconstrained case. Our motivation is to find an efficient way to at least approximately find the parameter β_k of the ideal method, without actually searching the ideal curve. This, in turn, will lead to an algorithm achieving the natural rate.

A familiar formulation of this result is obtained through the introduction of the *Lagrangian function* for (P), $l : R^n \times R^m \to R$ defined by

$$l(x, \lambda) = f(x) + \lambda h(x).$$

At every regular point x and for the partition of $R^n = B \oplus R$, we define the *reduced Lagrange multiplier* as the m-dimensional row vector

$$\lambda(x) = -\nabla_B f(x) \nabla_B h(x)^{-1}. \tag{9}$$

We can thus interpret the reduced gradient in terms of the gradient of the Lagrangian, since

$$\nabla_x l[x, \lambda(x)] = [\nabla_B l, \nabla_R l] = [0, {}'r(x)].$$

We can also evaluate the Hessian $\Phi(x)$ of the function $\phi(x)$:

$$\begin{aligned}
\Phi(x) &= {}'\nabla\psi(x) F_{BB}(x) \nabla\psi(x) + F_{BR}(x) \nabla\psi(x) + {}'\nabla\psi(x) F_{RB}(x) \\
&\quad + F_{RR}(x) + \nabla_B f(x) \Psi(x) \\
&= {}'T(x) F(x) T(x) + \lambda(x) \times (T(x) \cdot H(x) \cdot T(x)) \\
&\qquad\qquad \text{(using (1), (2), (3), (9))} \\
&= {}'T(x) L[x, \lambda(x)] T(x),
\end{aligned}$$

where $L(x, \lambda)$ is the Hessian, with respect to x, of the Lagrangian l. ${}'TLT$ represents a restriction to $\mathcal{T}(x)$ of the Hessian of the Lagrangian.

3.2. The step size selection. The reduced gradient method presented above is idealized from a computational viewpoint, since it is in practice impossible to generate the arcs $x^k(\beta)$. We can devise an implementable version of the reduced gradient algorithm, which is really an approximation, using first order information, of the idealized scheme. Calculation of the step along the curve from x^k to $x^k(\beta_k)$ is replaced by a combination of two phases, as depicted in Fig. 2. In the first phase, a move is made along the tangent to the ideal curve $x^k(\beta)$. (It has already been established that this tangent direction is given by $-T^k r^k$.) This step, charac-

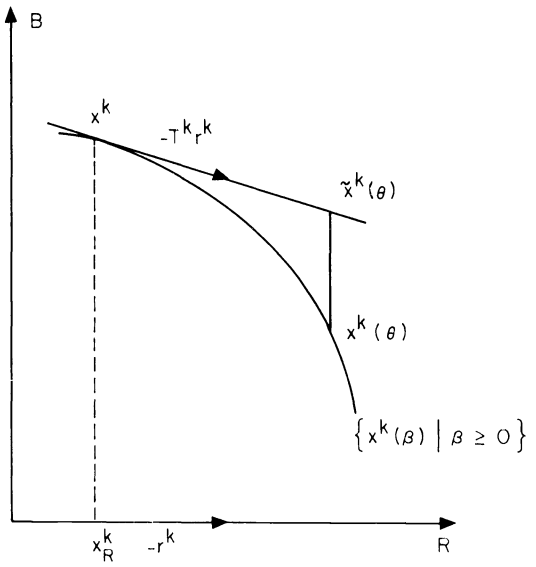


FIG. 2

terized by the step length parameter θ, leads to the point

$$\tilde{x}^k(\theta) = x^k - \theta T^k r^k,$$

which generally does not satisfy the constraint equations. A restoration phase back to a feasible point $x^k(\theta)$ of $\mathcal{S}$ is needed and is performed by adjusting the basic variables.

The efficiency of this algorithm depends, to a large extent, upon the selection of the step length θ_k. The study of the idealized reduced gradient method shows that, in order to achieve convergence at the natural rate, our practical algorithm must use a step size parameter which asymptotically satisfies

$$\theta_k = \beta_k + O(\|r^k\|).$$

We can compute an estimate of the value of the objective along the arc $\{x^k(\beta)\}$:

$$f[x^k(\beta)] = f(x^k) - \beta \nabla f^k T^k r^k + (\beta^2/2)' r^{k\prime} T^k F^k T^k r^k$$

$$- \beta^2 \nabla_B f^k (\nabla_B h^k)^{-1} q^k + O(\|r^k\|^3).$$

Introducing $\lambda^k = \lambda(x^k)$ as defined by (9) and using the definition (7) of q^k, this can be written

$$f[x^k(\beta)] = f(x^k) - \beta \|r^k\|^2 + (\beta/2)' r^{k\prime} T^k (F^k + \lambda^k \times H^k) T^k r^k + O(\|r^k\|^3). \tag{10}$$

Consider now the value of the Lagrangian function $l(\cdot, \lambda^k)$ along the direction

tangent at x^k to the arc $\{x^k(\beta)\} : \tilde{x}^k(\beta) = x^k - \beta T^k r^k$:

$$l[\tilde{x}^k(\beta), \lambda^k] = l(x^k, \lambda^k) - \beta \nabla l(x^k, \lambda^k) T^k r^k + \frac{\beta^2}{2} {}^t r^{kt} T^k L^k T^k r^k + O(\|r^k\|^3)$$

$$= f(x^k) - \beta \|r^k\|^2 + \frac{\beta^2}{2} {}^t r^{kt} T^k L^k T^k r^k + O(\|r^k\|^3).$$

Therefore the function $l(\cdot, \lambda^k)$ takes along the tangent direction $-T^k r^k$ up to second order the value of the objective f at the feasible point with the same independent coordinate.

Hence a simple and efficient procedure to find an approximation of the order of $\|r^k\|$ to the ideal step size β_k consists in selecting the step length parameter of the tangent phase in order to minimize $l(x, \lambda^k)$ along the direction $-T^k r^k$. This rule defines the parameter α_k:

$$\alpha_k = \text{Argmin}\,\{l(x^k - \alpha T^k r^k, \lambda^k) | \alpha \geqq 0\},$$

which we refer to as the *Lagrangian step size.* This selection rule can be easily incorporated in our algorithm. We describe below the detailed procedures to handle the possible computational difficulties associated with the restoration phase and to insure the convergence of the sequence of iterates.

3.3. The generalized reduced steepest descent algorithm. We must first notice that the choice of the Lagrangian step size does not represent any additional computational work since the evaluation of λ^k is a necessary step in the calculation of the reduced gradient r^k.

Under the assumption of uniform regularity of the constraints, the restoration phase from a point $\tilde{x}^k = [\tilde{x}^k_B, \tilde{x}^k_R]$ can always be performed, at least conceptually. A computationally efficient procedure consists of solving the system of equations

$$h_i(x_{B_1}, \cdots, x_{B_m}, \tilde{x}^k_{R_1}, \cdots, \tilde{x}^k_{R_{n-m}}) = 0, \qquad i = 1, \cdots, m,$$

for the unknown variables $x_{B_1}, \cdots, x_{B_m}$ using a modified Newton's method. Starting from $y^0 = \tilde{x}^k_B$, such a method constructs successive approximations $y^i \in R^m$, according to the recurrence

$$y^{i+1} = y^i - [\nabla_B h(x^k)]^{-1} h(y^i, \tilde{x}^k_R), \qquad i = 1, 2, \cdots.$$

(The inverse of the basic Jacobian at x^k has already been computed to evaluate λ^k and r^k.)

The convergence of this method for obtaining a feasible point has been established by Kantorovitch [18], provided that the starting point is sufficiently close to x^k and that the matrices H_i and $\nabla_B h(x)^{-1}$ are bounded in this neighborhood. It may be necessary to decrease the step length θ_k of the tangent phase, initially defined by the Lagrangian step size α_k, by scaling down θ_k by a factor $\rho_1 \in (0, 1)$, possibly several times, until the restoration phase is successful.

Assuming that the level sets of the Lagrangian function

$$\mathscr{L}[f(x^k)] = \{x \in R^n | l(x, \lambda^k) \leqq f(x^k)\}$$

are compact, the Lagrangian step sizes α_k are bounded. There exists, therefore, a neighborhood of a solution x^* of (P) and a corresponding integer N such that, for

all $k > N$, the iterates x^k are in this neighborhood and $\|r^k\|$ is small enough to guarantee the convergence of the modified Newton's method from the starting point $y^o = \tilde{x}_B^k(\alpha_k)$.

It is important to provide a rule for the step size which insures that the sequence $\{x^k\}$ is convergent. Each iteration must result in a descent in the value of the objective, and convergence can be established if this improvement is sufficient enough. Sufficient descent is achieved in our algorithm by enforcing the test for the step length θ_k first proposed by Armijo [5] in the framework of unconstrained optimization; namely, θ_k is scaled down by a factor $\rho_2 \in (0, 1)$ until

$$f[x^k(\theta_k)] < f(x^k) - \sigma\theta_k\|r^k\|^2, \tag{11}$$

where σ is a positive parameter chosen in $(0, \frac{1}{2})$. The Taylor expansion of f, considered as a function of θ, leads to

$$f[x^k(\theta_k)] - f(x^k) = -\theta_k\|r^k\|^2 + O(\theta_k^2).$$

Hence, after at most a finite number of scalings by the factor $\rho_2 \in (0, 1)$ from the initial determination $\theta_k = \alpha_k$, the test (11) will be satisfied.

We can give now a detailed description of the algorithm in a pseudo-ALGOL format. The method depends on the parameters ε, σ, ρ_1, and ρ_2 which must be specified in advance, with $\varepsilon > 0$, $\sigma \in (0, \frac{1}{2})$, $\rho_1 \in (0, 1)$, $\rho_2 \in (0, 1)$. The tolerance parameter ε expresses the accuracy required in the satisfaction of the constraints. The damping factors ρ_1, ρ_2 are selected according to the nonlinearity of the problem. (They are taken as $\frac{1}{2}$ or $\frac{1}{10}$ in the GRG method of Abadie [2]).

GRSD ALGORITHM (Generalized reduced steepest descent method).

Step 0. Select a feasible $x^0 \in R^1$; set $k = 0$.

Step 1. Procedure "check regularity assumption":
if x^k is not regular, then stop; else partition

$$x^k = (x_B^k, x_R^k) \quad \text{and} \quad \nabla h(x^k) = [B^k, D^k].$$

Step 2. Compute the reduced Lagrange multiplier:

$$\lambda^k = -\nabla_B f(x^k)(B^k)^{-1};$$

compute the reduced gradient:

$$'r^k = \nabla_R f(x^k) + \lambda^k D^k.$$

Step 3. Procedure "stopping rule":
if $r^k = 0$, then stop; *comment*: x^k is a solution candidate.

Step 4. Procedure "move in the tangent plane":
compute the direction $p^k = T^k r^k$;

Step 5. Compute the Lagrangian stepsize α_k such that

$$\alpha_k = \text{Argmin}\,\{l(x^k - \alpha p^k, \lambda^k) | \alpha \leqq 0\};$$

set $\theta = \alpha_k$ and $\tilde{x}^k(\theta) = x^k - \theta p^k$.

Step 6. Procedure "restoration of the constraints":
set $i = 0$; set $y^0 = \tilde{x}_B^k(\theta)$;
while $(\|h[y^i, \tilde{x}_R^k(\theta)]\| > \varepsilon)$ and $i <$ itermax do:
set $y^{i+1} = y^i - (B^k)^{-1}h(y^i, \tilde{x}_R^k)$ and set $i = 1 + 1$.

Step 7. If $\|h[y^f, \tilde{x}_R^k(\theta)]\| > \varepsilon$, then
set $\theta = \rho_1 \theta$, go to Step 6;
else set $x^k(\theta) = (y^f, \tilde{x}_R^k(\theta))$.
Step 8. If $f[x^k(\theta)] > f(x^k) - \theta\sigma\|r^k\|^2$,
then set $\theta = \rho_2 \theta$, go to Step 6.
Step 9. Set $\theta_k = \theta$ and $x^{k+1} = x^k(\theta_k)$.
Step 10. Set $k = k+1$, and go to Step 1.

3.4. Convergence properties. We prove first that the rule given by (11), adopted to determine the step length, guarantees the convergence of the algorithm.

THEOREM 1 (Global convergence). *Assume that f is bounded from below, and that the level set $\mathscr{L}[f(x^0)]$ is compact. Let $\{x^k\}$ be the sequence of regular feasible points constructed by the GRSD algorithm. Then every cluster point of $\{x^k\}$ is a critical point.*

Proof. It follows from the assumptions that the sequence $\{f(x^k)\}$ is monotonically decreasing and that θ_k is positively bounded from below. □

The above theorem, as well as establishing global convergence, also guarantees that, after a finite number of iterations, $\|r^k\|$ is small enough so that the restoration phase does not offer any computational difficulties. We can now prove also that the Lagrangian step size α_k satisfies the test (11) for k large enough.

PROPOSITION 1. *Let $\{x^k\}$ be a sequence, constructed by the GRSD algorithm, converging to x^*, a critical point of* (P) *which satisfies the sufficient second order optimality conditions. There exists an integer N, such that, for all $k > N$, the step lengths θ_k are determined directly by the values of the Lagrangian step sizes α_k.*

Proof. The definition of the Lagrangian step size leads to

$$\alpha_k = \frac{\|r^k\|^2}{{}^t r^k {}^t T^k L^k T^k r^k} + O(\|r^k\|),$$

which yields

$$f[x^k(\alpha_k)] - f(x^k) = -\frac{1}{2} \frac{\|r^k\|^4}{{}^t r^k {}^t T^k L^k T^k r^k} + O(\|r^k\|^3). \tag{12}$$

Assuming that x^* fulfills the sufficient second order optimality conditions [13], the matrix ${}^t T^* L^* T^*$ is positive definite (at x^* and therefore in a domain around x^*); hence, for k large enough the first term of the right-hand side of (12) is negative and dominates the second term which is only of the order $O(\|r^k\|^3)$. Thus, for k large enough, the point $x^k(\alpha_k)$ satisfies the test of Step 8, and we can make the choice $\theta = \alpha_k$; then $x^{k+1} = x^k(\alpha_k)$. □

The introduction of the Lagrangian step size is a very powerful device. We have just established that, once the algorithm has approached close enough to a solution, this parameter defines a tangent move from which the restoration phase can be performed successfully without using complex scaling down procedures. This choice leads therefore to a computationally simple and convergent algorithm. We can further prove that the convergence itself is efficient by showing that the algorithm converges at the natural rate.

THEOREM 2 (Local rate of convergence). *Assume that the sequence* $\{x^k\}$ *constructed by the GRSD algorithm converges to* x^*, *an isolated local minimizer of* f, *subject to the constraints* $h(x)=0$. *Let* M *and* m *be, respectively, the largest and smallest eigenvalues of the matrix* ${}^tT^*L^*T^*$, *the restriction of the Hessian of the Lagrangian to the tangent subspace to* $\mathcal{S}$ *at* x^*. *Then the sequence* $\{x^k\}$ *converges linearly to* x^* *with asymptotic ratio* $(M-m/M+m)^2$.

Proof. The proof is a generalization of a similar estimate for the rate of convergence of the optimal steepest descent method.

For small r^k, equation (12) gives an estimate of the decrease of the objective function during the kth iteration:

$$f(x^{k+1})-f(x^k)=-\frac{1}{2}\frac{\|r^k\|^4}{{}^tr^k{}^tT^kL^kT^kr^k}+O(\|r^k\|^3). \tag{13}$$

Introducing the error vector $y^k=x^k-x^*$, we have

$$\begin{aligned} f(x^k)-f(x^*)&=l(x^k,\lambda^k)-l(x^*,\lambda^k)\\ &=\nabla l(x^k,\lambda^k)y^k-\tfrac{1}{2}{}^ty^kL^ky^k+O(\|y^k\|^3). \end{aligned}$$

Let z^k be the vector of independent components of y^k. Using a first order Taylor expansion of the implicit function $y^k=\psi(z^k)$, we obtain

$$y^k=T^kz^k+O(\|z^k\|^2)=O(\|z^k\|).$$

Using a first order Taylor expansion of ∇l, we derive an estimate for r^k:

$${}^tr^k=[{}^ty^kL^k+O(\|y^k\|^2)]T^k={}^tz^k{}^tT^kL^kT^k+O(\|z^k\|^2).$$

Hence

$$f(x^k)-f(x^*)=\tfrac{1}{2}{}^tz^k{}^tT^kL^kT^kz^k+O(\|z^k\|^3),$$

or, in terms of r^k,

$$f(x^k)-f(x^*)=\tfrac{1}{2}{}^tr^k({}^tT^kL^kT^k)^{-1}r^k+O(\|y^k\|^3).$$

Denoting by F^k the matrix ${}^tT^kL^kT^k$, we have, for the ratio of the successive errors,

$$\frac{f(x^{k+1})-f(x^*)}{f(x^k)-f(x^*)}=\left(1-\frac{({}^tr^kr^k)^2}{{}^tr^kF^k{}^tr^k(F^k)^{-1}r^k}\right)(1+O(\|y^k\|)).$$

Let us introduce the normalized vectors $u^k=r^k/\|r^k\|$, which converge to u^*; the Kantorovitch inequality [17] gives for the positive definite matrix F^*,

$$\frac{({}^tu^*u^*)^2}{{}^tu^*F^*u^*{}^tu^*F^{*-1}u^*}\geqq\frac{4mM}{(M+m)^2}.$$

Hence we obtain the desired result:

$$\lim_{k\to\infty}\frac{f(x^{k+1})-f(x^*)}{f(x^k)-f(x^*)}\leqq\left(\frac{M-m}{M+m}\right)^2. \qquad \square$$

We have thus been able to establish that the convergence characteristics of the GRSD algorithm are simple and complete extensions of the corresponding

properties of the unconstrained steepest descent method. The algorithm has been run satisfactorily on several examples, derived from Colville's tests [7] (where only the constraints active at the optimum were considered and treated as equality constraints), and its computational performance has been consistently comparable with the results obtained with a GRG algorithm of similar sophistication. We describe in the next subsection certain situations where the new method performs even better than the original scheme.

3.5. Comparison with the generalized gradient method. Since the GRSD method and Abadie's GRG algorithm [2] use the same procedures to determine the direction of the tangent phase and to perform the restoration, the step length determination is the key difference between the two methods. In the latter, the point in the tangent direction is initially chosen to achieve a local minimum of the objective function. Although the resulting performance of the GRG algorithm is often satisfactory [3], this selection rule does not constitute the proper extension of the optimal steepest descent method and does not exhibit the efficient properties achieved by the Lagrangian step size in the GRSD algorithm.

For example, if the initial choice of the step length is systematically small compared to the ideal step size, the convergence of the GRG algorithm may be slowed. Such a situation arises when the restricted Hessian $'TFT$ of the objective is ill-conditioned compared to the corresponding Hessian $'TLT$ of the Lagrangian. This phenomenon is illustrated by the following problem:

$$\text{Minimize } 5x_1^2+3x_2^2+5x_3^2+x_4^2-9x_1+7x_2-x_3-6x_4$$

$$\text{Subject to } h_1(x)=x_1^2+x_2^2+x_3^2+x_4^2+x_1-7x_2+3x_3-5x_4+4=0,$$

$$h_2(x)=2x_1^2+x_2^2+2x_3^2+3x_2+5x_3-4x_4-9=0.$$

The objective achieves its minimum value 5 at the solution point $x^*=(1, 1, 1, 1)$; the corresponding Lagrange multiplier vector is $\lambda^*=(1, -2)$. The Hessians at x^* of the objective and of the Lagrangian are respectively

$$F^*=\begin{bmatrix}10 & & & \\ & 6 & & \\ & & 10 & \\ & & & 2\end{bmatrix} \text{ and } L^*=\begin{bmatrix}4 & & & \\ & 4 & & \\ & & 4 & \\ & & & 4\end{bmatrix}.$$

The starting point is taken as the feasible point $x^0=(3, 2, -1, 4)$. The partition chosen treats x_2 and x_3 as the basic variables and x_1 and x_4 as the independent ones. While it takes 19 iterations of the GRG algorithm to reach an approximate solution where the reduced gradient is in norm less than 10^{-3}, the GRSD algorithm reaches the same precision in only 6 iterations. Since each step of both methods consists of the same operations and necessitates about the same amount of computational work, the GRSD method is about three times faster.

4. A generalized reduced variable metric method for nonlinearly constrained minimization. For unconstrained minimization, the conjugate gradient method [14] or the variable metric algorithm [13], which exhibit superlinear rates

of convergence, are sometimes preferred to the method of steepest descent. It is, therefore, natural to seek a way of combining these efficient schemes with the reduced gradient technique in order to solve constrained problems. But the appealing properties of these methods rely, to a substantial degree, on the fact that, at each step, the objective function is accurately minimized along the direction of search. This is not an obstacle when the constraints are linear, and simple as well as efficient combination schemes have been proposed in this framework [25].

The only available extension of the Fletcher–Powell method to nonlinearly constrained minimization has been proposed by Davies [9] in the context of Rosen's gradient projection [27]. The restoration phase is, however, a source of difficulty, ignored by Davies but acknowledged by Murtagh and Sargent [24], since the new feasible iterate is not likely to exactly achieve a local minimum of f. This leads to a possible deterioration of the convergence properties of the algorithm.

4.1. The idealized reduced variable metric method. A natural and efficient generalization to the constrained case can be provided within the implicit function framework we have already adopted to extend the method of steepest descent. The key idea consists again in viewing problem (P) in terms of the reduced unconstrained problem (R). The minimization of $\phi(x_R)$ is then, at least ideally, performed by the variable metric method in the subspace R of the independent variables. The kth iteration of this scheme proceeds from x_R^k by searching for the minimum of $\phi(x_R)$ along a direction s^k, defined by

$$s^k = G^k r^k,$$

where G^k is an $(n-m)\times(n-m)$ matrix updated according to the formula

$$G^{k+1} = G^k - \frac{G^k(r^{k+1}-r^k)'(r^{k+1}-r^k)G^k}{'(r^{k+1}-r^k)G^k(r^{k+1}-r^k)} + \frac{(x_R^{k+1}-x_R^k)'(x_R^{k+1}-x_R^k)}{'(x_R^{k+1}-x_R^k)(r^{k+1}-r^k)},$$

which approximates the inverse Φ, the Hessian of ϕ.

In practice it is necessary to solve the problem in the original space R^n, since the reduction to the form (R) can generally be achieved only conceptually. The ideal scheme consists, therefore, in defining a curve $\{x^k(\beta)\}$ on $\mathcal{S}$ emanating from $x^k = (x_B^k, x_R^k)$, its projection on R, parallel to the basic subspace B, being the straight line in the negative direction of s^k. To extend the Fletcher–Powell method, the next point x^{k+1} must be chosen to achieve a local minimum of f along the arc $\{x^k(\beta)|\beta \geqq 0\}$.

By construction, this method exhibits the convergence properties of the variable metric method in the $(n-m)$-dimensional subspace R. In particular, the rate of convergence of this variable metric method is actually superlinear. Moreover, the conceptual framework adopted shows that we have only to construct a sequence of $(n-m)\times(n-m)$ matrices G^k instead of $n\times n$ matrices as proposed in [9].

4.2. A generalized variable reduced metric method. The method developed in the previous paragraph is an idealized version, since it is computationally

impossible to generate the curves $\{x^k(\beta)\}$. But we can derive from it a practically implementable algorithm which asymptotically generates the same points. Again, this is achieved by a move along the direction $p^k = T^k s^k$ of the tangent subspace $\mathcal{T}^k$, followed by a restoration.

To obtain the best possible approximation of the arc of the idealized scheme, we are led, as in § 3.2, to define the step length parameter in terms of the Lagrangian step size, achieving a local minimum of the updated Lagrangian $l(x, \lambda^k)$ along p^k. We have established in § 3.3 that this provides a first order approximation to the step size of the ideal search for the point achieving a local minimum of f along the arc $x^k(\beta)$.

An even better method would be to adapt this generalization technique in conjunction with the version of the variable metric method proposed recently for the unconstrained situation by Oren and Luenberger [26], the self-scaling variable metric algorithm. It exhibits rapid convergence even when the minimization step is performed only approximately, while the Fletcher–Powell algorithm is adversely affected by even a small error in the step size.

5. A generalized Newton's method for nonlinearly constrained minimization. In spite of the very appealing fast convergence of Newton's method for the minimization of unconstrained convex functions (when second order information is available and when the dimension n of the problem is not too large to prohibit storage and inversion of an $n \times n$ matrix), very little effort has been devoted to extend the method to constrained situations. Levitin and Polyak [20] were the first to study a Newton's scheme for such cases. They proposed an implementable algorithm which considers only a linearized version of the constraints and which uses the inverse of the $n \times n$ Hessian F of the objective function to compute each iteration. This does not seem to be the most suitable approach, since it ignores the nonlinearity of $\mathcal{S}$ and therefore does not fully capture the essence of the problem to second order. It is preferable to explicitly incorporate the second order information available.

5.1. The idealized reduced Newton's method. An ideal method can again be conceived by viewing problem (P) in terms of the reduced problem (R) and by adopting Newton's method in the subspace R to find the unconstrained minimum of $\phi(x_R)$. The derivatives of ϕ have already been computed in § 3.1:

$$\nabla\phi(x) = {}^t r(x),$$
$$\Phi(x) = {}^t T(x) L[x, \lambda(x)] T(x).$$

Hence the kth iteration consists of a move from x_R^k along the (negative of the) direction

$$p^k = ({}^t T^k L^k T^k)^{-1} r^k. \tag{14}$$

To guarantee a descent in the value of the objective, we must assume that ${}^t TLT$ is positive definite and we must sometimes use a damping parameter $\theta \in (0, 1]$ to reduce the size of the step along $-p^k$, until the point

$$x_R^{k+1} = x_R^k - \theta_k p^k$$

satisfies a descent condition; a test like Armijo's rule, for example [5]. There is no need, however, to determine the step length by an accurate minimization procedure, as in the previous gradient-related methods, to obtain efficient convergence, since asymptotically $\theta_k = 1$ will yield convergence of order 2. If the matrix ${}^tT^kL^kT^k$ is not positive definite, p^k must be modified to preserve the descent character of the algorithm. Computationally efficient schemes [12], [16] for the unconstrained case can be applied in this case as well.

In the original space R^n, a step of this ideal scheme consists in moving along an arc $\{x^k(\beta)|0 \leqq \beta \leqq 1\}$ *of* $\mathcal{S}$, the projection on R of which is the straight line in the negative direction of p^k. By introducing this ideal scheme, we conclude that it is necessary to invert only $(n-m)\times(n-m)$ matrix. We can also derive from it a practical algorithm, by approximating the search along the arc to second order.

5.2. The generalized reduced Newton's algorithm. Let us consider the move of the independent variables

$$x_R^k(\theta) = x_R^k - \theta p^k, \quad \text{with} \quad \theta \in (0, 1].$$

To satisfy the constraint equations, the basic variables must be altered to

$$x_B^k(\theta) = \psi(x_R^k - \theta p^k) = \psi(x_R^k) - \theta \nabla\psi^k p^k + \frac{\theta^2}{2} {}^tp^k \cdot \Psi^k \cdot p^k + O(\|p^k\|^3),$$

$$x_B^k(\theta) = x_B^k - \theta(-B^{-1}D)p^k - \theta^2 B^{-1} q^k + O(\|p^k\|^3),$$

where q^k is now the m-dimensional column vector with components

$$q_i^k = \tfrac{1}{2} {}^tp^k {}^tT^k H_i(x^k) T^k p^k. \tag{15}$$

We thus obtain a second order approximation of the form

$$\tilde{x}^k(\theta) = \begin{bmatrix} \tilde{x}_B(\theta) \\ \tilde{x}_R(\theta) \end{bmatrix} = x^k - \theta T^k p^k - \theta^2 V^k q^k, \tag{16}$$

where

$$T^k = \begin{bmatrix} -(B^k)^{-1} D^k \\ I \end{bmatrix}, \qquad V^k = \begin{bmatrix} (B^k)^{-1} \\ 0 \end{bmatrix}.$$

From a geometric viewpoint, we can interpret this approximation as a move along the osculating parabola c^k to the ideal curve $\{x^k(\beta)|\beta > 0\}$, i.e., the parabola of origin x^k parameterized by θ as

$$\tilde{x}^k(\theta) = \theta t^k + \theta^2 v^k$$

in the 2-dimensional variety containing x^k and spanned by the vectors t^k and v^k, where

$$t^k = -T^k p^k, \qquad v^k = -V^k q^k.$$

This curve is the natural extension, for second order approximation, to the tangent t^k. (See Fig. 3.)

In general, however, the points $\tilde{x}^k(\theta)$ are not feasible, and a move back to $\mathcal{S}$ is again performed by altering the basic variables through a modified Newton's

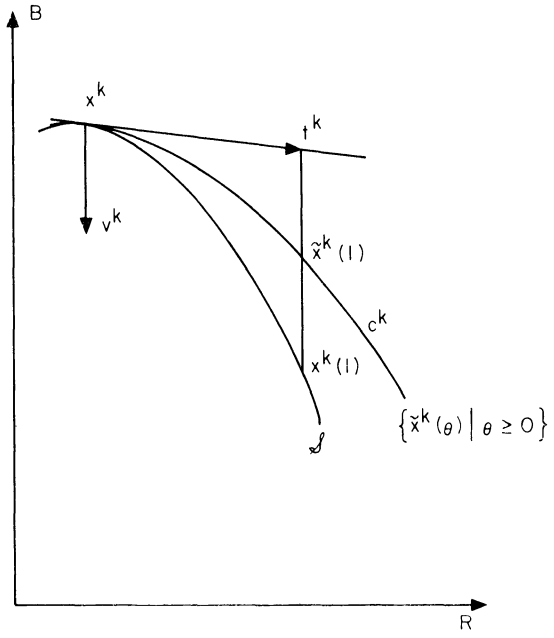


FIG. 3

method to solve the system

$$h[y, \tilde{x}_R^k(\theta)] = 0.$$

As mentioned in § 3.3, the restoration move can be a source of difficulty, since the modified Newton's method may fail to converge or may lead to a new feasible point which does not represent an improvement over x^k in the objective function. We handle these difficulties by successive halving of θ_k from the initial value $\theta_k = 1$ eventually finding a new feasible iterate x^{k+1} such that, given a scalar $\sigma \in (0, \frac{1}{2})$,

$$f(x^{k+1}) \leqq f(x^k) - \theta_k \frac{\sigma}{\|{}^t T^k L^k T^k\|} \|r^k\|^2. \tag{17}$$

We emphasize, however, that these difficulties are less frequent than with first order methods, since the approximations of the constraints used in the present scheme are valid to within second order.

We now present our algorithm. It is of the same form as the GRSD algorithm in § 3.3 except for Steps 3, 4 and 5.

GRN ALGORITHM (Generalized reduced Newton's method).

Step 0. Select a feasible $x^0 \in R^n$; set $k = 0$.

Step 1. Check regularity assumption.

Step 2. Compute λ^k, r^k.

Step 3. Stopping rule: if $r^k = 0$, then stop.

Step 4. Procedure "move along the osculating parabola":

compute $p^k = ({}^t T^k L^k T^k)^{-1} r^k$;

compute $q_i^k = \frac{1}{2} {}^t p^k {}^t T^k H_i^k T^k p^k$ for $i = 1, \cdots, m$.

Step 5. Set $\theta = 1$ and $\tilde{x}^k(\theta) = x^k - \theta T^k p^k - \theta^2 V^k q^k$;
Steps 6, 7. Restoration of the constraints.
Step 8. If $f[x^k(\theta)] > f(x^k) - \theta(\sigma/\|{}^tT^kL^kT^k\|)\|r^k\|^2$,
then set $\theta = \frac{1}{2}\theta$, go to Step 6.
Step 9. Set $\theta_k = \theta$ and $x^{k+1} = x^k(\theta_k)$.
Step 10. Set $k = k+1$ and go to Step 1.

Since $-p^k$ is a direction of descent, the test (17) is satisfied after at most a finite number of halvings of the original step size $\theta_k = 1$. This selection rule guarantees the convergence of a subsequence of $\{x^k\}$ to a critical point, as established in § 3.4.

We can also show that, after a finite number of iterations, no halving of the step length is necessary.

PROPOSITION 2. *Let $\{x^k\}$ be a sequence, constructed by the GRN algorithm, converging to x^*, an isolated local minimizer of* (P). *Assume that ${}'T^*L^*T^*$ is positive definite. Then there exists an integer N, such that, for all $k > N$, we may take $\theta_k = 1$.*

Proof. An expansion of $f[\tilde{x}^k(1)]$ to second order gives

$$f[\tilde{x}^k(1)] = f(x^k) - \nabla f(x^k)(T^k p^k + V^k q^k) + \tfrac{1}{2}{}^t(T^k p^k + V^k q^k)F^k(T^k p^k + V^k q^k) + O(\|p^k\|^3).$$

Using (15), we obtain

$$f[\tilde{x}^k(1)] = f(x^k) - {}^tr^k p^k + \tfrac{1}{2}{}^t p^k {}^tT^k L^k T^k p^k + O(\|p^k\|^3).$$

By definition (16), we have

$$\|h[\tilde{x}^k(1)]\| = O(\|r^k\|^3).$$

Therefore, if $\|r^k\|$ is not too large, which occurs for k large enough since $r^k \to 0$, the modified Newton's method converges to a feasible point $x^k(1)$. We derive, using the definition (14) of p^k,

$$f[x^k(1)] = f(x^k) - \tfrac{1}{2}{}^tr^k({}^tT^kL^kT^k)^{-1}r^k + O(\|r^k\|^3). \tag{18}$$

For any $\sigma \in (0, \frac{1}{2})$, there exists an N large enough such that

$$f[x^k(1)] < f(x^k) - \frac{\sigma}{\|{}^tT^kL^kT^k\|}\|r^k\|^2 \quad \text{for all } k > N;$$

therefore the test of Step 8 is satisfied for the step length $\theta_k = 1$ and the new iterate $x^{k+1} = x^k(1)$. □

The estimate (18) shows that the choice $\theta_k = 1$ achieves, at least asymptotically, the best possible decrease in the objective along the parabola c^k.

THEOREM 3. *Assume that f and h_i are three times continuously differentiable. Assume that $\{x^k\}$ converges to x^*, an isolated local minimizer of* (P). *Then this convergence is of order at least* 2.

Proof. Let us introduce the error vector $y^k = x^k - x^*$ and partition it as $y^k = (w^k, z^k)$. Since x^{k+1} and x^* are feasible, we have

$$h(x^{k+1}) - h(x^*) = 0 = \nabla h(x^*)y^{k+1} + O(\|y^{k+1}\|^2) = \nabla_B h(x^*)w^{k+1} + \nabla_R h(x^*)z^{k+1} + O(\|y^{k+1}\|^2).$$

Since h is uniformly regular, we get

$$\|w^{k+1}\| = O(\|z^{k+1}\|).$$

According to Proposition 2, we have, for k large enough,

$$x_R^{k+1} = x_R^k - ({}^tT^k L^k T^k)^{-1} r^k.$$

The study of this iterative process shows that its convergence is of second order. Hence

$$\|z^{k+1}\| = O(\|z^k\|^2).$$

Since

$$\|y^{k+1}\| \leqq \|z^{k+1}\| + \|w^{k+1}\| = O(\|y^k\|^2),$$

we obtain the rate of convergence of order 2:

$$\|x^{k+1} - x^*\| \leqq c\|x^k - x^*\|^2. \qquad \square$$

Recently Mangasarian [22] has proposed a Newton's method for nonlinearly constrained minimization which exhibits quadratic convergence. The role of the Lagrange multipliers is also central to his approach, although he uses more general Lagrangian functions than in this paper. Feasibility is not required at each iteration, but it is necessary to compute the inverse of an $n \times n$ matrix.

6. Numerical experience. The GRN algorithm has been tested on the quadratic problem described in § 3.5. Convergence was quite rapid. From the same starting point as used before, the problem was solved in 3 iterations, yielding a value for the solution with 7 exact digits.

A nonquadratic test problem in 5 variables and 3 constraints was also run. From an initial approximation defined as the solution rounded to one decimal place, full precision was achieved after a single iteration.

7. Conclusion. The algorithms presented in this paper are of both practical and theoretical interest for nonlinearity constrained minimization. On the practical side, our methods efficiently generalize to this class of problems the appealing convergence properties of the optimum steepest descent method, the variable metric method, and Newton's method for unconstrained minimization. Computational results indicate that they can provide significant savings in computer time as compared to the existing schemes, particularly when the constraints are highly nonlinear. There is, of course, room for further improvement. The requirement of maintaining feasibility may cause excessive time expenditure in the restoration phase of each iteration if a high degree of accuracy is demanded in the satisfaction of the constraints. It is, however, possible to adaptively improve the accuracy requirements of the restoration phase as the minimization procedure progresses. Further investigation of such restoration schemes might, therefore, lead to faster computational performance. Other areas for further research include the extension of our methods to problems with inequality constraints and the development of effective rules for updating the partition between basic and independent variables.

On the theoretical side, our study has shown how to fully exploit the viewpoint associated with the implicit function theorem in order to define computational algorithms for the solution of nonlinear programming problems. One of the key observations in this perspective is the necessity of constructing a sequence of approximate Lagrange multipliers $\{\lambda^k\}$ simultaneously with the sequence of approximate solutions $\{x^k\}$. The resulting interplay of Lagrangian methods in a primal framework should lead to useful new results.

REFERENCES

[1] J. ABADIE, J. CARPENTIER AND C. HENSGEN, *Generalization of the Wolfe reduced gradient method to the case of nonlinear constraints*, ES/TIMS Joint European Meeting, Warsaw, Poland, 1966 (available from Electricité de France, EDF Note HR 7595, April 5th 1967).

[2] J. ABADIE AND J. GUIGOU, *Gradient réduit généralisé*, EDF Note HI 069/02, April 15th, 1969.

[3] ———, *Numerical experiments with the* GRG *method*, Integer and Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1970, pp. 529–536.

[4] M. ALTMAN, *A generalized gradient method for the conditional minimum of a functional*, Bull. Acad. Polon. Sci. Sér. Sci. Math., 14 (1966), pp. 445–451.

[5] L. ARMIJO, *Minimization of functions having Lipschitz-continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[6] A. CAUCHY, *Méthode générale pour la résolution des systèmes d'équations simultanées*, C. R. Acad. Sci. Paris, 25 (1847), pp. 536–538.

[7] A. R. COLVILLE, *A comparative study on nonlinear programming codes*, IBM Tech. Rep. 320-2949, New York Scientific Center, 1968.

[8] G. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, N.J., 1963.

[9] D. DAVIES, *Some practical methods of optimization*, Integer and Nonlinear Programming, J. Abadie, ed., 1970, pp. 87–117.

[10] J. DIEUDONNE, *Foundations of Modern Analysis*, Academic press, New York, 1960.

[11] P. FAURE AND P. HUARD, *Résolution de programmes mathématiques à fonction non linéaire par la méthode du gradient réduit*, Rev. Française Recherche Opérationnelle, 9 (1965), pp. 167–206.

[12] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968.

[13] R. FLETCHER AND M. J. D. POWELL, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.

[14] R. FLETCHER AND C. M. REEVES, *Function minimization by conjugate gradients*, Ibid., 7 (1964), pp. 149–154.

[15] D. GABAY, *Efficient convergence of reduced gradient algorithms for nonlinearly constrained optimization*, Ph.D. dissertation, Dept. of Engineering-Economic Systems, Stanford University, Palo Alto, Calif., 1973.

[16] J. GREENSTADT, *On the relative efficiencies of gradient methods*, Math. Comp., 21 (1967), pp. 360–367.

[17] L. V. KANTOROVITCH, *Functional analysis and applied mathematics*, Uspehi Mat. Nauk., 3 (1948), pp. 89–185; English transl., Nat. Bur. of Standards Rep. No. 1509, 1952.

[18] L. V. KANTOROVITCH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, Fizmatgiz, Moscow, 1959; English transl., Pergamon Press–Macmillan, New York, 1964.

[19] H. J. KELLEY AND J. L. SPEYER, *Accelerated gradient projection*, Symposium on Optimization, Nice, 1969, A. V. Balakrishnan, ed., Springer-Verlag, Berlin, 1970, pp. 151–158.

[20] E. S. LEVITIN AND B. T. POLYAK, *Constrained Minimization Methods*, Ž. Vyčisl. Mat. i Mat. Fiz., 6 (1966), pp. 787–823 = USSR Comp. Math. and Math. Phys., 6 (1968), pp. 1–50.

[21] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass., 1973.

[22] O. L. MANGASARIAN, *Unconstrained Lagrangians in nonlinear programming*, Tech. Rep. 174, Computer Sciences Department, University of Wisconsin, Madison, 1973.

[23] A. MIELE, H. Y. HUANG AND J. C. HEIDEMAN, *Sequential gradient restoration algorithm for the minimization of constrained functions—Ordinary conjugate gradient versions*, J. Optimization Theory Appl., 4 (1969), pp. 213–243.

[24] B. A. MURTAGH AND R. W. H. SARGENT, *A constrained minimization method with quadratic convergence*, Optimization, R. Fletcher, ed., Academic Press, London, 1969, pp. 215–246.

[25] G. P. MCCORMICK, *The variable reduction method for nonlinear programming*, Management Sci., 17, (1970), pp. 146–160.

[26] S. S. OREN AND D. G. LUENBERGER, *The self-scaling variable metric algorithm* (SSVM), Fifth Hawaiian Internat. Conf. on System Sciences, 1972.

[27] J. B. ROSEN, *The gradient projection method for nonlinear programming. Part I: Linear constraints*; *Part II: Nonlinear constraints*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 181–217; 9 (1961), pp. 514–532.

[28] R. WOLFE, *Methods of nonlinear programming*, Recent Advances in Mathematical Programming, R. L. Graves and P. Wolfe, eds., McGraw Hill, New York, 1963, pp. 67–86.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

FULL "BANG" TO REDUCE PREDICTED MISS IS OPTIMAL*

V. E. BENEŠ†

Abstract. Consider the stochastic control problem of minimizing the final value expectation $El(k'z_1)$ by choosing a measurable control law $u(\cdot,\cdot)$, subject to the stochastic differential equation $dz_t = A(t)z_t\,dt + B(t)u(t,z_t)\,dt + C(t)\,dw_t$, $0 \leqq t \leqq 1$, for the process z, and to the boundedness condition $u: [0,1] \times R^d \to [-1,1]^r$, with $w_\cdot$ a Wiener process, $k \neq 0$ a given vector, and $l(\cdot)$ an even positive function increasing in $x > 0$. C. G. Hilborn, Jr. and others have conjectured that one optimal law takes the form of full "bang" in the direction of reducing the "predicted miss", defined as the expected value of $k'z_1$ with control identically zero. Using the maximum principle for parabolic operators, we prove this conjecture in the setting of the exponential functionals which express the derivatives of measures induced by translations in Wiener space.

1. Introduction. We consider the stochastic control problem of minimizing the final value expectation $El(k'z_1)$ by choosing a control law $u(\cdot,\cdot)$, subject to the boundedness condition $u: [0,1] \times R^d \to [-1,1]^r$ on the measurable function $u(\cdot,\cdot)$, and subject to the stochastic DE (differential equation)

$$dz_t = A(t)z_t\,dt + B(t)u(t,z_t) + C(t)\,dw_t, \qquad 0 \leqq t \leqq 1, \tag{1}$$

for the process $z_\cdot$, with $w_\cdot$ a Wiener process, k a given vector, and $l(\cdot)$ an even positive function, increasing in $x > 0$. Our interest in this problem arose from reading an unpublished work of C. G. Hilborn, Jr., who conjectured that one optimal control law took the form of full "bang" in the direction of reducing the "predicted miss", defined as the expected value of $k'z_1$ with control identically zero. This conjecture is proved here in the setting of Girsanov's theorem for the exponential functionals which express the derivatives of measures induced by translations in Wiener space.

Girsanov's theorem [1] serves to connect these functionals with stochastic control theory [2]. It states that for φ a nonanticipative Brownian functional with $\int |\varphi|^2\,dt < \infty$ a.s., and $d\tilde{P} = \exp\zeta(\varphi)\,dP$ with $\zeta(\varphi) = \int \varphi\,dw - \frac{1}{2}\int |\varphi|^2\,dt$ and $E\exp\zeta = 1$, the translated functions $w_t - \int_0^t \varphi\,ds$ are a Wiener process under $\tilde{P}$. This result is used [2] in stochastic control theory as follows: it is assumed that the controlled system satisfies a functional DE $dx_t = f(t,x,u(t,x))\,dt + dw_t$; here f represents system dynamics and u is a particular control law; a "solution" is provided by the Wiener functions w_t under $\tilde{P}$ with $\varphi = f(t,w,u(t,w))$, in the sense that there is a Wiener process W_t such that

$$w_t = \int_0^t f(s,w,u(s,w))\,ds + W_t.$$

This idea has been exploited, in stochastic control for existence proofs [2] for optimal laws, for Hamilton–Jacobi conditions [3] for optimality, and for direct proofs of optimality [4]. Actually we shall use a slightly more involved version [1] of Girsanov's theorem than that quoted above, in order to take account of the matrix $C(\cdot)$ that modifies the noise in equation (1).

* Received by the editors September 14, 1973, and in revised form November 25, 1974.
† Bell Laboratories, Murray Hill, New Jersey 07974.

The notion of "predicted miss", as it is to be used here, is introduced in § 2. Our basic assumptions and formulation of the problem, and the resulting representation for the cost of a control law, are in §§ 3 and 4, respectively. Section 5 gives an informal outline of the basic comparison arguments that prove optimality. The next five sections, 6 through 10, are mostly heuristic, and aim to exhibit the analytical and probabilistic reasons for the relevance of predicted miss. The optimality proof begins in earnest in § 11 with a study of the sgn of the gradient of the value function. After a brief but necessary digression on smoothing control laws (§ 12), the comparison of control laws is carried out in § 13. There follow three appendices that are of technical nature.

Appendix A establishes various requisite properties of the functionals used in representing the cost of using a control law. Appendix B is concerned with the validity of the hypothesis $E \exp \zeta(\varphi) = 1$ in Girsanov's theorem, and more particularly with that of Girsanov's Lemma 7 [1]. This lemma has been used by several authors to prove the above hypothesis; its meaning and validity have also been questioned. We give a reconstruction of Girsanov's argument for Lemma 7. For the case of principal physical interest, viz., linear growth of φ, we give a new short proof of $E \exp \zeta(\varphi) = 1$, not depending on Lemma 7 at all. Appendix C, finally, answers a question of Balakrishnan about the measure of the set of points at which switching occurs in the optimal regime.

2. Predicted miss. If a process $z_\cdot$ satisfying (1) starts from the point z at the time t, and no, i.e., zero, control is exerted, then the expected value of $k'z_1$ is given by $s(t)'z$, where $s(\cdot)$ satisfies the "adjoint" equation

$$\dot{s}(t) = -A(t)'s(t), \qquad s(1) = k.$$

This function $s(t)'z$ is called the *predicted miss*. It has previously appeared in the stochastic control literature [5], [6] as the basis of conjectured optimal or near-optimal laws for problems with boundedness and/or "finite fuel" constraints, but nowhere has it been proved to give an optimal policy. It has been guessed that if $s(t)'z_t$ is positive, then maximum control effort should go to reducing $s(t)'z_t$, and inversely if it is negative. For the finite fuel case, it is likely that there is a central region of space-time in which no effort should be made; determining this region is a problem orders of magnitude harder than the one we are solving, and it is not considered here. But with simple boundedness constraints on the control, it was conjectured that for the purposes of final value control in which one seeks to minimize say the distance of the final point from a subspace $k'z = 0$, the information comprised in the state could be compressed into the one-dimensional statistic $s(t)'z_t$ without loss, and that in fact one optimal law had the form

$$u(t, z) = -\operatorname{sgn} B(t)'s(t)s(t)'z, \tag{2}$$

where the sgn of a vector is the vector of sgna of the components.

We shall show that this is right, in the sense that the $u(\cdot, \cdot)$ as defined in (2) achieves the infimum of $El(k'z_1)$ over all measurable control laws restricted in value to $[-1, 1]^k$, when the "solution" z_t corresponding to a given control law is constructed by use of Girsanov's theorem, as will be done in § 3. Notice that this optimal control in no way depends on the noise modifying function $C(\cdot)$, whose

role, under the positivity condition $C(t)C(t)' > 0$, will become purely that of changing the time scale in a suitable representation of the (optimal) value function.

3. Assumptions and formulation. We assume that $k \neq 0$, and that $A(\cdot)$, $B(\cdot)$ and $C(\cdot)$ are (respectively $d \times d$, $r \times d$, and $d \times d$ matrix-valued continuous functions, with $C(\cdot)$ meeting the uniform elliptic condition that $C(t)C(t)' - cI$ be positive definite for some $c > 0$; $C(\cdot)$ is then also nonsingular. For the convergence of integrals it will be convenient to assume that $l(x) = O(\exp \kappa|x|)$ for some $\kappa > 0$, in addition to being even, and increasing in $x > 0$.

Let the class $\mathscr{A}$ of admissible control laws consist of all measurable functions $u: [0, 1] \times R^d \to [-1, 1]^r$. Since the Itô theory of stochastic differential equations is not available for (1) because u is not Lip, we shall construct solutions, or rather solution measures, by using Girsanov's theorem. We define, for each $u \in \mathscr{A}$, $z \in R^d$, and $s \in [0, 1]$ a solution of (1) that starts at z at time s and corresponds to use of control law u. Let w_t be a d-dimensional Wiener process defined on a probability space $(\Omega, \mathscr{B}, P)$; to solve (1) take the functions

$$z_t = z + \int_s^t C(u)\, dw_u$$

under the measure $\tilde{P}$ defined by $d\tilde{P} = \exp \zeta\, dP$ with

$$\zeta = \int_s^1 C(u)^{-1} g(u, z_u)\, dw_u - \frac{1}{2}\int_s^1 |C(u)^{-1} g(u, z_u)|^2\, du,$$

$$g(t, z) = A(t)z + B(t)u(t, z).$$

Under P, the functions z_t form an Itô process with respect to w corresponding to drift zero and diffusion $C(u)$; assuming the linear growth condition $|g(u, z)|^2 \leqq \kappa(1 + |z|^2)$, it can be shown that $E \exp \zeta = 1$; it then follows from Girsanov's theorem that under $\tilde{P}$, the translated functions

$$W_t = w_t - w_s - \int_s^t C(u)^{-1} g(u, z_u)\, du, \qquad s \leqq t \leqq 1,$$

form a Wiener process $W_\cdot$, and the original functions z_t form an Itô process with respect to W corresponding to drift $g(u, z_u)$ and diffusion $C(u)$. Thus

$$dz_t = g(t, z_t)\, dt + C(t)\, dW_t$$

in the sense that $z_t = z + \int_s^t C(u)\, dw_u$ and

$$z + \int_s^t C(u)\, dw_u = z + \int_s^t g(u, z + \int_s^u C(v)\, dw_v)\, du + \int_s^t C(u)\, dW_u.$$

4. Representation of the cost. If the system is run from time s to time 1 using control law u and starting in state z, the expected cost incurred is

$$J_{s,z}[u] = El(k'z_1) \exp \int_s^1 C(t)^{-1} g(t, z_t)\, dw_t - \frac{1}{2}\int_s^1 |C(t)^{-1} g(t, z_t)|^2\, dt, \tag{3}$$

where $g(t, y) = A(t)y + B(t)u(t, y)$ and $z_t = z + \int_s^t C(v)\, dw_v$. This is an explicit representation, peculiar to u. By setting $\tau = 1 - s =$ "time to go" and changing

variables a bit we can write this cost or value for the control law u as $v(\tau, z) = El(k'[z + \int_0^\tau C(1-\tau+t)\,dw_t]\exp\zeta$, with ζ given by

$$\int_0^t C(1-\tau+t)^{-1}g\left(1-\tau+t, z+\int_0^t C(1-\tau+s)\,dw_s\right)dw_t$$

$$-\frac{1}{2}\int_0^\tau |C(1-\tau+t)^{-1}g(1-\tau+t, z+\int_0^t C(1-\tau+s)\,dw_s)|^2\,dt.$$

The eventual cost is then $J[u] = v(1, z)$.

5. Outline of the argument. A standard way of approaching a Markov control problem like the one we have posed is to look for a sufficiently smooth solution $V(\tau, z)$ of the Bellman–Hamilton–Jacobi equation

$$V(0, z) = l(k'z), \tag{4}$$

$$V_\tau = \min_{u\in[-1,1]^r} \tfrac{1}{2}\operatorname{tr} C(1-\tau)'D_2VC(1-\tau) + \nabla V'[A(1-\tau)z + B(1-\tau)u],$$

where D_2 is the matrix $(\partial^2/\partial z_i\,\partial z_j)$. This equation is difficult to attack, even numerically. It does suggest again, though, that there may be an optimal bang-bang control law.

We prefer to work with specific control laws, for which the corresponding cost functions satisfy PDE's similar to (4), but without the min. When $u(\cdot,\cdot)$ is smooth, say Lip, the solution $z_\cdot$ of the stochastic DE (1) can be constructed in the usual Itô's way, and the corresponding value functions $v(\tau, z) = E\{l(k'z_1)|z_{1-\tau} = z\}$ will satisfy the backward PDE

$$v(0, z) = l(k'z), \tag{5}$$

$$v_\tau = \tfrac{1}{2}\operatorname{tr} C(1-\tau)'D_2vC(1-\tau) + \nabla v'[A(1-\tau)z + B(1-\tau)u(1-\tau, z)].$$

Our method will be as follows: (i) to show that any admissible control law can be approximated by a smooth one in such a way that nearly the same cost is incurred; (ii) to single out a special class of control laws, viz., the smooth laws depending oddly on predicted miss; (iii) to show that for these laws the cost $v(\tau, z)$ and its gradient ∇v can be calculated from a *one-dimensional* problem; (iv) to use the maximum principle for (5) to compare these laws to other, arbitrary ones. In the course of these comparisons it will turn out that given any law whatever there is a law in the class that is (i) at least as good as the given one and (ii) arbitrarily close to the law $\sigma(t, z) = -\operatorname{sgn} B(t)'s(t)s(t)'z$. From these facts it readily follows that σ is optimal in the sense that it achieves $\inf J[u]: u\in\mathscr{A}$.

6. Reduction to one dimension for laws depending on predicted miss. The next four sections are heuristic in presentation. While only a few of the results are actually needed for the optimality argument, we feel that together they shed so much light on the structure of the problem that we eagerly include them here. They dispel much of the ad hoc character of the basic proof.

The class $\mathscr{P}$ of admissible control laws depending only on predicted miss consists of $u\in\mathscr{A}$ of the form

$$u(t, z) = y(t, s(t)'z)$$

for some measurable $y\colon [0,1]\times R\to[-1,1]^r$. We shall exhibit precise analytical and probabilistic senses in which calculating the cost or value function for laws in $\mathscr{P}$ reduces to a one-dimensional problem.

7. Stochastic equation for predicted miss. For $u\in\mathscr{P}$ smooth, let us take the stochastic DE $z_s = z$, $u(t,z_t) = y(t,s(t)'z_t)$, $dz_t = A(t)z_t\,dt + B(t)u(t,z_t)\,dt + C(t)\,dw_t$, $s\leqq t\leqq 1$, seriously and define, with Hilborn, a one-dimensional process x_t as $x_t = s(t)'z_t$, $s\leqq t\leqq 1$. This is the expected value of $k'z_1$, given z_t, if control were zero henceforth. Clearly $x_1 = s(1)'z_1 = k'z_1$, and with $s = 1-\tau$, $x_{1-\tau} = s(1-\tau)'z$. Taking the Itô differential, we find

$$\begin{aligned} dx_t &= \dot{s}(t)'z_t\,dt + s(t)'\,dz_t \\ &= s(t)'B(t)y(t,x_t)\,dt + s'(t)C(t)\,dw_t, \qquad 1-\tau\leqq t\leqq 1, \end{aligned} \tag{6}$$

and the expected cost is

$$v(\tau,z) = E\{l(x_1)|x_{1-\tau} = s(1-\tau)'z\},$$

where x solves (6). Thus we can expect the cost for $u\in\mathscr{P}$ to have the form $v(\tau,z) = \xi(\tau,s(1-\tau)'z)$, where $\xi(\tau,x)$ solves a one-dimensional parabolic PDE associated with (6); this is shown in the next section.

8. Composition with predicted miss. Following the hint of the previous paragraphs we now note that if u is smooth and belongs to $\mathscr{P}$, with $u(t,z) = y(t,s(t)'z)$, and if ξ is a solution of $\xi(0,x) = l(x)$,

$$\xi_\tau = s(1-\tau)'B(1-\tau)y(1-\tau,x)\xi_x + \tfrac{1}{2}\sigma(1-\tau)\xi_{xx} \tag{7}$$

with $\sigma(t) = |s(t)'C(t)|^2$, then the composition

$$v(\tau,z) = \xi(\tau,s(1-\tau)'z)$$

satisfies the expected cost equation (5) associated with the law u, for we have

$$\begin{aligned} v_\tau &= \xi_1(\tau,s(1-\tau)'z) - \xi_2(\tau,s(1-\tau)'z)\dot{s}(1-\tau)'z \\ &= s(1-\tau)'B(1-\tau)y(1-\tau,s(1-\tau)'z)\xi_2(\tau,s(1-\tau)'z) \\ &\quad + \tfrac{1}{2}\sigma(1-\tau)\xi_{22}(\tau,s(1-\tau)'z) + \xi_2(\tau,s(1-\tau)'z)[A'(1-\tau)s(1-\tau)]'z. \end{aligned}$$

Since $\nabla v = \xi_2(\tau,s(1-\tau)'z)s(1-\tau)$ and

$$s'CC's|_{1-\tau}\xi_{22}(\tau,s(1-\tau)'z) = \operatorname{tr} C(1-\tau)'D_2vC(1-\tau),$$

the PDE (5) for expected cost using u follows. This composition result represents an analytical sense in which a "predicted miss" law $u\in\mathscr{P}$ reduces the problem to one dimension. Of course, this reduction does not show that in solving the control problem only laws $u\in\mathscr{P}$ need be considered; this must be done separately (§ 13). What the reduction does do is enable us to calculate $\operatorname{sgn}\nabla v$ for $u\in\mathscr{P}$ with y odd in x.

9. A time substitution. The next task is to "get rid" of the noise modifying function $s'CC's$ which appears in the equation for ξ; this will be done by using another composition, this time a time substitution, to represent the solution ξ

of (7). We solve for ξ as $\xi(\tau, x) = \psi(T - t(1 - \tau), x)$, where $T = t(1)$,

$$t(\tau) = \int_0^\tau \sigma(r)\, dr, \qquad \sigma(v) = |s(v)'C(v)|^2 > 0, \tag{8}$$

and ψ satisfies $\psi(0, x) = l(x)$,

$$\psi_r = \frac{1}{2}\psi_{xx} + \left.\frac{s'By}{\sigma}\right|_{t^{-1}(T-r),x} \psi_x.$$

For then

$$\begin{aligned}\xi_\tau &= \psi_1(T - t(1 - \tau), x)\sigma(1 - \tau) \\ &= \frac{1}{2}\sigma(1 - \tau)\psi_{22}(T - t(1 - \tau), x) + \left.\frac{s'By}{\sigma}\right|_{t^{-1}(t(1-\tau)),x} \psi_2(T - t(1 - \tau), x)\sigma(1 - \tau) \\ &= \frac{1}{2}\sigma(1 - \tau)\xi_{xx} + s'By\Big|_{1-\tau,x} \xi_x.\end{aligned}$$

10. Reduction to one dimension: Probability version. Let us now see the same facts probabilistically from the integral for $v(\tau, z)$. Indeed, all the parts of the preceding "analytical" reduction can be found "inside" the integral. We shall find, among the constituents for v itself, a Wiener process $w_\cdot^*$ in R such that

$$v(\tau, z) = E\{l(y_{t(1)}) | y^*_{t(1-\tau)} = s(1 - \tau)'z\},$$

where $y_\cdot^*$ solves the stochastic DE

$$\begin{aligned} y^*_{t(1-\tau)} &= s(1 - \tau)'z, \\ dy_r^* &= \left.\frac{s'By}{\sigma}\right|_{t^{-1}(r),y^*(r)} dt + dw^*, \qquad t(1 - \tau) \leqq r \leqq t(1),\end{aligned}$$

$t(\cdot)$ being the time substitution defined in (8). It is easy to see that $x_v = y^*(t(v))$, that the stochastic DE for $y_\cdot^*$ corresponds to the PDE for ψ in § 9, and that the time substitution relating x_v and y_r^* mirrors that defining ξ from ψ. Note that the PDE's are formulated for "time to go" while the processes and $t(\cdot)$ itself are defined for elapsed time; this circumstance explains the use of $T = t(1)$ in § 9, the ranges of validity of the DE for $y_\cdot^*$ and $x_\cdot$, etc.

Let $s = 1 - \tau$ be fixed and consider the functions on $s \leqq t \leqq 1$,

$$W_t = w_t - w_s - \int_s^t C(v)^{-1}[A(v)z_v + B(v)y(v, s(v)'z_v)]\, dv,$$

where $z_v = z + \int_s^v C(r)\, dw_r$ as in § 3. According to Girsanov's theorem these form a Wiener process under the measure $\exp \zeta\, dP$ where, as in § 3,

$$\zeta = \zeta_s^1(G^{-1}[A(\cdot)\cdot + B(\cdot)y(\cdot, s(\cdot)'\cdot)]_{v,z_v}).$$

We can now express $k'z_1$ as the value x_1 at 1 of a process $x_\cdot$ defined as $x_t = s(t)'z_t$, $s \leqq t \leqq 1$: this is obvious, because $s(1) = k$ by construction. But we can get still another expression for $k'z_1$ by examining the stochastic equation satisfied by $x_\cdot$.

Exploiting the relationship between $w_\cdot$ and $W_\cdot$, we find in analogy with § 7,

$$
\begin{aligned}
dx_t &= \dot{s}(t)z_t\,dt + s(t)'\,dz_t \\
&= -s(t)'A(t)z_t\,dt + s(t)'C(t)\,dw_t \\
&= -s(t)'A(t)z_t + s(t)'[C(t)\,dW_t + [A(t)z_t + B(t)y(t, x_t)]]\,dt, \qquad (9) \\
dx_t &= s(t)'B(t)y(t, x_t)\,dt + s(t)'C(t)\,dW_t, \qquad s \leqq t \leqq 1.
\end{aligned}
$$

It follows, noting that $x_{1-\tau} = s(1-\tau)'z$, that

$$
k'z_1 = x_1 = s(1-\tau)'z + \int_{1-\tau}^{1} s(t)'B(t)y(t, x_t)\,dt + \int_{1-\tau}^{1} s(t)'C(t)\,dW_t.
$$

Thus $v(\tau, z) = E\{l(x_1)|x_{1-\tau} = s(1-\tau)'z\}$ with x a solution of (9). Introduce again, finally, the time substitution $v \to t(v)$ such that

$$
t(v) = \int_0^v s'CC's\,dr,
$$

which is strictly monotone because $CC' > 0$, and k, and hence $s(\cdot)$, does not vanish. According to a known result of McKean [7, p. 46],

$$
w_v^* = \int_{1-\tau}^{t^{-1}(v)} s(u)'C(u)\,dW_u, \qquad t(1-\tau) \leqq v \leqq t(1),
$$

under the same $\exp \zeta\,dP$ that makes $W_\cdot$ Wiener, is a *one-dimensional* Wiener process w^* to which the transformed process $y_v^* = x_{t^{-1}(v)}$, $t(1-\tau) \leqq v \leqq t(1)$, is related by the equation

$$
\begin{aligned}
y_v^* &= s(1-\tau)'z + \int_{1-\tau}^{t^{-1}(v)} s(u)'B(u)y(u, x_u)\,du + \int_{1-\tau}^{t^{-1}(v)} s'C\,dW \\
&= s(1-\tau)'z + \int_{t(1-\tau)}^{v} s'By\Big|_{u, y_u^*}\,du + w_v^*.
\end{aligned}
$$

11. The sgn of the gradient Δv. We now turn to the proof of optimality. The first task is to show that for smooth $u \in \mathscr{A}$ with $u(t, z) = y(t, s(t)'z)$ and $y(\cdot, x)$ odd in x we have $\operatorname{sgn} \nabla v = \operatorname{sgn} s(1-\tau)'zs(1-\tau)$. Knowing $\operatorname{sgn} \nabla v$ will enable us to use the maximum principle to compare control laws; this use [4] of the maximum principle is similar to that in Wonham's optimality lemma [8, p. 321], and is not unrelated to Pontryagin's maximum principle in deterministic optimal control. Lemmas A.1–A.9 are in the Appendix.

As is customary we use $C^{m,n}$ to mean the class of functions $f\colon R \times R^d \to R^d$ (or R, etc.) which are m (resp. n) times continuously differentiable in the first (resp. second) variable, and $C_b^{m,n}$ to mean the subclass for which all these derivatives are bounded.

LEMMA 1. *If $u \in \mathscr{P} \cap C_b^{1,3}$, and u is of the form $u(t, z) = y(t, s(t)'z)$ with $y(t, x) = -y(t, -x)$ (i.e., y is odd in its second argument), then*

$$
\operatorname{sgn} \nabla v = \operatorname{sgn} s(1-\tau)'zs(1-\tau).
$$

Proof. There is a unique solution $v(\tau, z)$ to the problem $v(0, z) = l(k'z)$, $v_\tau = \frac{1}{2}\operatorname{tr} C(1-\tau)'D_2vC(1-\tau) + [A(1-\tau)z + B(1-\tau)u(1-\tau, z)]'\nabla v$, with $v = O(\exp \kappa|z|^2)$ expressible as $v(\tau, z) = \xi(\tau, s(1-\tau)'z)$, with $\xi(\tau, x)$ the unique solution of $\xi(0, x) = l(x)$, $\xi_\tau = s'CC's|_{1-\tau}\xi_{xx} + s'By|_{1-\tau,x}\xi_x$ that is of exponential type. In turn, ξ is expressible as $\xi(\tau, x) = \psi(T - t(1-\tau), x)$, where $t(v) = \int_0^v s'CC's\,dr$ and ψ is the unique solution of $\psi(v, x) = l(x)$, $\psi_r = \frac{1}{2}\psi_{xx} + s'By/s'CC's|_{t^{-1}(T-r),x}\psi_x$ that is of exponential type. These facts follow from the existence and uniqueness theorems for Cauchy problems [9, p. 25, Thm. 12 and p. 44, Thm. 10], and from the elementary calculus with compositions presented in §§ 8 and 9. From Lemma A.8 it is seen that the integral

$$El(k'z_1)\exp \zeta_s^1(G(v)^{-1}g(v, z_v)), \qquad z_v = z + \int_s^v C(r)\,dw_r$$

with $g(v, z) = A(v)z + B(v)u(v, z)$ satisfies the same equation as $v(\tau, z) = A(v)z + B(v)u(v, z)$ satisfies the same equation as $v(\tau, z)$ and so is equal to it, being of exponential type. *Simile* ξ and ψ are expressible as expectations, and in particular, using the even and odd characters of $l(\cdot)$ and $y(t, \cdot)$, we deduce by the first passage time argument of [4] that $\operatorname{sgn} \psi_2(t, x) = \operatorname{sgn} x$. Since

$$v(\tau, z) = \psi(T - t(1-\tau), s(1-\tau)'z)$$

we find at once that

$$\operatorname{sgn} \nabla v = \operatorname{sgn} \psi_2(T - t(1-\tau), s(1-\tau)'z)s(1-\tau) = \operatorname{sgn} s(1-\tau)'zs(1-\tau).$$

12. Smoothing of control laws. To smooth control laws $u(t, z)$ from $\mathscr{A}$, which are defined only on $[0, 1] \times R^d$, we shall extend them to R^{d+1} by equating them to 0 when $t \notin [0, 1]$. For measurable functions $f: R^{d+1} \to [-1, 1]^r$ we use the smoothings $f \to S_\delta f$ defined by

$$S_\delta(f)(y) = (2\delta)^{-d-1}\int_C f(y + v)\,dv,$$

where C = cube of side 2δ centered on the origin in R^{d+1}. The restriction to $[0, 1] \times R^d$ of a smoothed extension of a member of $\mathscr{A}$ is again in $\mathscr{A}$, by convexity. $(S_\delta)^m u$ approaches u in L_2 as $\delta \to 0$ and has bounded mth partials. In particular $(S_\delta)^3 u(t, z) \in C_b^{3,3}$.

13. Comparison of control laws. We next show that given any control law $g \in \mathscr{A}$ and any $\varepsilon > 0$, there is another law $u \in \mathscr{P}$, depending only on predicted miss, which is as good as g to within ε, and which is within ε in norm of our natural guess candidate $\sigma(t, z) = -\operatorname{sgn} B(t)'s(t)s(t)'z$.

LEMMA 2. *$g \in \mathscr{A}$, $\varepsilon > 0 \Rightarrow \exists u \in \mathscr{A} \ni \|u - \sigma\| < \varepsilon$ and*

$$J[u] \leqq J[g] + \varepsilon.$$

Proof. Choose δ by Lemma A.9 so that $\|(S_\delta)^3 g - g\|$ is so small that

$$J[(S_\delta)^3 g] \leqq J[g] + \varepsilon/2.$$

Let $\phi: R \to [-1, 1]$ be a C_b^3-function such that

$$\phi(x) = -\operatorname{sgn} x \text{ outside } |x| \leqq \delta,$$
$$\phi(x) = -\phi(-x),$$
$$\phi(x) \leqq 0 \quad \text{for } x \geqq 0$$

and define u and h both in $\mathscr{P}$ by

$$u_i(t, z) = \phi(B(t)'s(t)_i s(t)'z), \qquad i = 1, \cdots, d,$$
$$h_i(t, z) = (S_\delta)^3 g(t, z)_i u_i^2(t, z), \qquad i = 1, \cdots, d.$$

Let B_i be the (t, z)-set on which $B(t)'s(t)_i s(t)'z \geqq 0$. Then on B_i we have $u_i \leqq 0$, so that if $(S_\delta)^3 g(t, z)_i \geqq 0$, then $h_i \geqq u_i$, because $h_i \geqq u_i$; in the opposite case that $(S_\delta)^3 g(t, z)_i \leqq 0$ it is apparent that $u_i^2 (S_\delta)^3 g(t, z)_i \geqq u_i$. Dually, $h_i \leqq u_i$ on the complement of B_i. Thus componentwise, $h \geqq u$ on $s'zB's \geqq 0$ and $h \leqq u$ on $s'zB's \leqq 0$. It is clear that δ can be further reduced, if necessary, so that both $\|u - \sigma\| < \varepsilon$ and (by Lemma A.9)

$$J[h] \leqq J[(S_\delta)^3 g] + \varepsilon/2.$$

Now define the operator $L[g]$ by

$$L[g] = \tfrac{1}{2} \operatorname{tr} C(1 - \tau)' D_2 C(1 - \tau) + g(1 - \tau, z)'\nabla = \frac{\partial}{\partial \tau}.$$

L is parabolic because of the uniform ellipticity assumption $CC' - cI > 0$. Set $\xi(\tau, z) = J_{\tau,z}[u]$, $\eta(\tau, z) = J_{\tau,z}[h]$ to obtain, by Lemma A.8,

$$L[u]\xi = 0, \qquad L[h]\eta = 0.$$

By construction u is an odd function of $B(t)'s(t)s(t)'z$ so that $\operatorname{sgn} u(t, z) = -\operatorname{sgn} B(t)' \cdot s(t)s(t)'z$. It follows from Lemma 1 that $\operatorname{sgn} \nabla\xi = \operatorname{sgn} s(1 - \tau)s(1 - \tau)'z$, and the inequalities between u and h imply that

$$\begin{aligned}[B(1 - \tau)u(1 - \tau, z)]'\nabla\xi &= u(1 - \tau, z)'B(1 - \tau)'\nabla\xi \\ &\leqq h(1 - \tau, z)'B(1 - \tau)'\nabla\xi.\end{aligned}$$

Therefore $L[u = h]\xi \leqq 0$, and so $L[h](\eta - \xi) = -L[h]\xi \leqq -L[u]\xi \leqq 0$. Since $\eta - \xi$ is of exponential type, the maximum principle for parabolic operators [9, p. 43, e.g.] implies that $\eta - \xi \geqq 0$, that is, $J_{\tau,z}[h] \geqq J_{\tau,z}[u]$. It follows that $J[u] \leqq J[h] \leqq J[f] + \varepsilon$. Lemma 2 is proved.

The following basic justification of the full "bang" to reduce predicted miss policy now follows at once from this last lemma.

THEOREM 1. *For $k \neq 0$, for l, ∇l, and $D_2 l$ of exponential type, for $A(\cdot)$. $B(\cdot)$, and $C(\cdot)$ continuous with $CC' - cI > 0$ for some $c > 0$, the control law $\sigma(t, z) = -\operatorname{sgn} B(t)'s(t)s(t)'z$ achieves*

$$\inf_{u \in \mathscr{A}} J[u].$$

This result readily extends to the more general criterion $El(k'z_1) + E\int_0^1 L(t, k'z_t)\,dt$ containing an averaged time-integral of a suitable function of trajectory, and to the case of noisy observations of z_t, in which case $s(t)'z_t$ is replaced by $s(t)'\hat{z}_t$, with $\hat{z}_t$ the Kalman filter estimate of z_t, and $C(\cdot)$ is replaced by a more involved matrix depending on $C(\cdot)$ itself, on the observation equation, and on the covariance matrix arising from the filtering problem.

Appendix A. Some analytical properties of such functionals as $J[u]$ in (3) are established here as prerequisite to the optimality proof in §§ 11–13. Although these prerequisites are few in number, some of their proofs have many steps that are ancillary to the main arguments, so it was natural to put them all in an appendix. What we need are the following: (a) conditions implying that $\exp\zeta$ belongs to L_α for some $\alpha > 1$; (b) order estimates for $v(\tau, z)$; (c) continuous differentiability of v and ∇v in z; (d) order estimates for ∇v and $D_2 v$; (e) the natural PDE for $v(\tau, z)$; (f) continuity of v in the L_2 topology of control laws. For (c) and (e) we shall assume suitable smoothness of the drift coefficient.

Since the above results do not depend on having a linear system with linearly entering control, and since they are basic also to further studies, we shall establish them for a general drift coefficient $g(t, z)$ in place of the function

$$A(t)z + B(t)u(t, z)$$

that appears in (a) or (c). Growth or smoothness of $g(\cdot, \cdot)$ will be postulated as it is needed. Also, we use a general cost function $k(\cdot)$ in place of the $l(k'\cdot)$ in the original problem, with $|\nabla k| = O(\exp\kappa|z|)$, $|(D_2 k)_{ij}| = O(\exp\kappa|z|)$. In the next 8 lemmas,

$$v(\tau, z) = Ek(z_1)\exp\zeta_{1-\tau}^1(g(\cdot, z.)),$$

$$z_t = z + \int_{1-\tau}^t C(v)\,dw_v,$$

$$\zeta(z) = \int_{1-\tau}^1 g(t, z_t)\,dw_t - \frac{1}{2}\int_{1-\tau}^1 |g(t, z_t)|^2\,dt.$$

Arguments of ζ are often omitted, and convenient changes of variable are used without too much explanation.

LEMMA A.1. *If* $g:[0, 1] \times R^d \to R^d$ *with* $|g(t, y)|^2 \leqq \kappa(1 + |y|^2)$, *then given* $\varepsilon > 0$ *there exist* $\alpha > 1$ *and* K *depending only on* ε, κ, *and* $C(\cdot)$, *such that*

$$\sup_{0 \leqq \tau \leqq 1} E\alpha \exp\zeta_0^\tau(g) < K\exp\varepsilon|z|^2,$$

where

$$\zeta_0^\tau(g) = \int_0^\tau g(1 - \tau + t, z_t)\,dw_t - \frac{1}{2}\int_0^\tau |g(1 - \tau + t, z_t)|^2\,dt,$$

$$z_t = z + \int_0^t C(1 - \tau + u)\,dw_u.$$

Proof. Girsanov's Lemma 7 [1] implies that $E \exp \zeta_0^\tau(\alpha g) = 1$. By his Theorem 1, then, for each $\alpha > 1$, under $\exp \zeta_0^\tau(\alpha g)\, dP$ the functions

$$Z_t = z_t - \alpha \int_0^t C(1 - \tau + u) g(1 - \tau + u, z_u)\, du, \qquad 0 \leqq t \leqq \tau,$$

form a Gaussian process of the same kind as z_t. By the c_2 and Gronwall inequalities, for $0 \leqq t \leqq \tau \leqq 1$ and

$$\beta = \kappa \sup_{0 \leqq s \leqq 1} \|C(s)\|^2 \quad \text{(operator norm)},$$

$$|z_t|^2 \leqq 2|Z_t|^2 + 2\alpha\beta \int_0^t (1 + |z_s|^2)\, ds$$

$$\leqq 2(\alpha^2\beta + \sup_{0 \leqq s \leqq t} |Z_s|^2) \exp 2\alpha^2\beta.$$

Now write

$$E \exp \alpha\zeta_0^\tau(g) = E \exp \{\zeta_0^\tau(\alpha g) + \frac{\alpha^2 - \alpha}{2} \int_0^\tau |g(1 - \tau + u, z_u)|^2\, du\}$$

$$\leqq \exp \tfrac{1}{2}\kappa(\alpha^2 - \alpha) E \exp \{\zeta_0^\tau(\alpha g) + \kappa(\alpha^2 - \alpha)[\alpha^2\beta + \sup_{0 \leqq s \leqq \tau} |Z_s|^2]\, e^{2\alpha^2\beta}\}.$$

Since for each $\alpha > 1$, $\{Z_t, 0 \leqq t \leqq \tau\}$ under $\exp \zeta_0^\tau(\alpha g)\, dP$ has the same distributions as z_t, the last expectation simplifies to give

$$E \exp \alpha\zeta_0^\tau(g) \leqq \exp \kappa(\alpha^2 - \alpha)[\tfrac{1}{2} + \alpha^2\beta\, e^{2\alpha^2\beta} + 2|z^2|]$$

$$\cdot E \exp 2\kappa(\alpha^2 - \alpha) \sup_{0 \leqq s \leqq \tau} |z_s - z|^2.$$

Recalling that

$$\sup_{0 \leqq s \leqq \tau} |z_s - z|^2 = \sup_{0 \leqq s \leqq \tau} \left| \int_0^{1-\tau+s} G\, dw - \int_0^{1-\tau} G\, dw \right|^2 \leqq 4 \sup_{0 \leqq s \leqq 1} \left| \int_0^s G\, dw \right|^2$$

we find

$$E \exp \alpha\zeta_0^\tau(g) \leqq h(\alpha, z) E \exp 8\kappa(\alpha^2 - \alpha) \sup_{0 \leqq s \leqq 1} \left| \int_0^s G\, dw \right|^2$$

with $h(\alpha, z) = O(\exp 2\kappa(\alpha^2 - \alpha)|z|^2)$ uniformly in $\alpha \geqq 1$. Doob's submartingale inequality implies that the sup in the exponent is finite a.s., so a result of Landau and Shepp [10, p. 377, Thm. 5] or [11] assures us that the last expectation above is finite for $\alpha > 1$ small enough, and depends only on α, κ, and $G(\cdot)$.

LEMMA A.2. *If* $l(x) = O(e^{\kappa|x|})$, *then for every* $\varepsilon > 0$,

$$v(\tau, z) = O(\exp \varepsilon|z|^2)$$

uniformly in $0 \leqq \tau \leqq 1$.

Proof. $v(\tau, z) \leqq \text{const.}\, E\, e^{\kappa|k'z_1|}\, e^{\zeta_1^{1-\tau}(g)}$, where g satisfies the conditions of Lemma 1; since z_τ is a Gaussian process, $\sup_{0 \leqq \tau \leqq 1} E \exp \beta|k'z_\tau| < \infty$; now use Hölder's inequality and Lemma A.1.

LEMMA A.3. *If* $g \in C_b^{0,2}$, *then with probability* 1, $\nabla\zeta$ *exists and equals*

$$\int_0^\tau J(1-\tau+s, z_s)\,dw_s - \int_0^\tau (Jg)(1-\tau+s, z_s)\,ds,$$

where J *is the Jacobian matrix function*

$$J_{ij} = \frac{\partial g_j}{\partial x_i}, \quad \textit{i.e.,}\ J = \nabla g' = (\nabla g_1, \cdots, \nabla g_d).$$

Proof. Set for $h \in R^d$, $|h| = 1$,

$$\varphi_j(\varepsilon) = \frac{g_j(1-\tau+s, \varepsilon h + z_s) - g_j(1-\tau+s, z_s)}{\varepsilon} - h'\nabla g_j(1-\tau+s, z_s)$$

so that Taylor's formula gives, since $g \in C_b^{0,2}$,

$$\varphi_j(\varepsilon) = \varepsilon \int_0^1 (h'\nabla)^2 g_j(1-\tau+s, t\varepsilon h + z_s)(1-t)\,dt,$$

$$E\left|\int_0^\tau \varphi_j(\varepsilon)\,dw_s\right|^2 = O(\varepsilon^2) \quad \text{uniformly in } \tau, z.$$

Evidently $|\varphi_j(\varepsilon+\xi) - \varphi_j(\varepsilon)| \leqq \text{const.}|\xi|$ if h is a unit vector, so

$$E\left|\int_0^\tau [\varphi_j(\varepsilon+\xi) - \varphi_j(\varepsilon)]\,dw_s\right|^2 = O(|\xi|^2).$$

Thus the argument for Kolmogorov's sample continuity theorem shows that an ε-separable version of $\int_0^\tau \varphi_j(\varepsilon)\,dw_s$ is a.s. a continuous function of ε vanishing at $\varepsilon = 0$, i.e.,

$$\nabla \int_0^\tau g(1-\tau+s, z_s)\,dw_s = \int_0^\tau J(1-\tau+s, z_s)'\,dw_s.$$

Similarly, setting

$$\psi_j(\varepsilon) = \frac{|g_j(1-\tau+s, \varepsilon h + z_s)|^2 - |g_j(1-\tau+s, z_s)|^2}{\varepsilon} = 2(h'\nabla g_j)\Big|_{\substack{1-\tau+s \\ z_s}}$$

$$= \varepsilon \int_0^1 (h'\nabla)^2 |g_j(1-\tau+s, t\in h + z_s)|^2 (1-t)\,dt$$

we can show, from the linear growth of g and from $g \in C_b^2$, that $E|\int_0^\tau \psi_j(\varepsilon)\,ds|^2 = O(\varepsilon^2)$ and that

$$E\left|\int_0^\tau \psi_j(\varepsilon+\xi)\,ds - \int_0^\tau \psi_j(\varepsilon)\,ds\right|^2 = O(|\xi|^2).$$

By the sample continuity argument just used, $\int_0^\tau \psi_j(\in)\,ds$ is continuous and vanishes at $\varepsilon = 0$, so that a.s.

$$\nabla \int_0^\tau |g(1-\tau+s, z_s)|^2\,ds = \int_0^\tau (Jg)(1-\tau+s, z_s)\,ds.$$

LEMMA A.4. *If $g \in C_b^{0,3}$, then with probability* 1, $D_2\zeta$ *exists and equals*

$$\int_0^\tau D_2 g(1 - \tau + s, z_s)\, dw_s - \int_0^\tau (JJ' + g'D_2 g)\Big|_{\substack{1-\tau+s \\ z_s}} ds.$$

Proof of this result is exactly analogous to that of the previous lemma using the linear growth of g, and the boundedness of J and $D_2 g$.

LEMMA A.5. *If $g \in C_b^{0,3}$, then $v \in C^1$ as a function of z, with*

$$\nabla v(\tau, z) = E\, e^{\zeta(z)}[\nabla k(z_\tau) + k(z_\tau)\nabla\zeta(z)],$$

$$|\nabla v| = O(\exp \varepsilon |z|^2) \quad \textit{uniformly in } \tau \textit{ for every } \varepsilon > 0.$$

Proof. One could appeal to the absolute convergence of the differentiated integrand under E. Instead, note that

$$v(\tau, z + h) - v(\tau, z) = E\, e^{\zeta(z)}[k(h + z_\tau) - k(z_\tau)] + Ek(h + z_\tau)[e^{\zeta(z+h)} - e^{\zeta(z)}].$$

Since $k \in C^2$ and $\zeta(z) \in C^2$ a.s. we can expand by Taylor:

$$k(h + z_\tau) - k(z_\tau) = h'\nabla k(z_\tau) + \int_0^1 (h'\nabla)^2 k(th + z_\tau)(1 - t)\, dt,$$

$$e^{\zeta(z+h)} - e^{\zeta(z)} = h'\nabla\zeta(z)\, e^{\zeta(z)} + \int_0^1 (h'\nabla)^2\, e^{\zeta(z+th)}(1 - t)\, dt,$$

whence

$$v(\tau, z + h) - v(\tau, z) - E\, e^{\zeta(z)}[h'\nabla k(z_\tau) + k(z_\tau)h'\zeta(z)]$$
$$= \int_0^1 (1 - t)E[e^{\zeta(z)}(h'\nabla)^2 k(th + z_\tau) + k(h + z_\tau)(h'\nabla)^2\, e^{\zeta(z+th)}]\, dt.$$

The first term is easily seen to be $O(|h|^2)$, by Lemma 1 and the order assumption on $D_2 k$. Then

$$(h'\nabla)^2\, e^{\zeta(z+th)} = e^{\zeta(z+th)}\{|h'\nabla\zeta(z + th)|^2 + h'D_2\zeta(z + th)h\},$$

so the second term in the last integral above gives at most

$$|h|^2 \int_0^1 (1 - t)E\, e^{\zeta(z+th)}\left[|\nabla\zeta(z + th)|^2 + \sum_{i,j} |D_2\zeta(z + th)|_{ij}\right] dt = O(|h|^2)$$

by Lemma 1 and the fact that $\nabla\zeta(z)$ and $D_2\zeta(z)$ are in L_2 uniformly on compact z-sets. This justifies the formula for ∇v; continuity is shown by applying an analogous argument to the integral formula for ∇v.

LEMMA A.6. *If $g \in C_b^{0,3}$, then $v \in C^2$ as a function of z, with*

$$D_2 v(\tau, z) = E\, e^{\zeta}(D_2 k(z_\tau) + 2\nabla k(z_\tau)\nabla\zeta(z)' + k(z_\tau)D_2\zeta(z)),$$

$$(D_2 v)_{ij} = O(\exp \varepsilon |z|^2) \quad \textit{uniformly in } \tau \textit{ for any } \varepsilon > 0.$$

Proof. The proof is very similar to that of Lemma A.5, using the orders of ∇k and $D_2 k$, the L_2-integrability of $\nabla\zeta$ and $D_2\zeta$, and of course Lemma A.1.

LEMMA A.7. *Let g and h be nonanticipating Brownian functionals such that $E\int_0^1 |g|^2\,dt < \infty$, $E\int_0^1 |h|^2\,dt < \infty$, and $E\exp \alpha\zeta_0^1(h) < \infty$ for some $\alpha > 1$. Then*

$$E\int_0^1 g\,dw \exp \zeta_0^1(h) = E\int_0^1 g'h\, e^{\xi_0^s(h)}\,ds.$$

Proof. Set $\chi_M(s)$ = indicator of the event

$$\sup_{0\leqq u\leqq s} \exp \zeta_0^u(h) < M,$$

and recall that $E\{\exp \zeta_s^1(h)|w_u, 0 \leqq u \leqq s\} = 1$ a.s. Then

$$E\int_0^1 g\,dw\left(1 + \int_0^1 \chi_M(s)h(s)\, e^{\zeta_0^s(h)}\,dw_s\right) = E\int_0^1 g'h\chi_M\, e^{\zeta_0^s(h)}\,ds.$$

The function $g'h\, e^{\zeta_0^s(h)}$ is $ds\,dP$ integrable and dominates $\chi_M g'h\, e^{\zeta(h)}$. Similarly,

$$\int_0^1 g\,dw\left(1 + \int_0^1 \chi_M(s)h(s)\, e^{\zeta_0^s(h)}\,dw\right)$$

is dominated by $\int_0^1 g\,dw \sup_{0\leqq t\leqq 1} e^{\zeta_0^t(h)}$, also integrable because $\exp \zeta_0^s(h)$ is an L_α-martingale, so that

$$E|\sup_{0\leqq t\leqq 1} \exp \zeta_0^t(h)|^\alpha \leqq \frac{\alpha}{\alpha - 1} E\, e^{\alpha\zeta_0^1(h)}.$$

The lemma follows by dominated convergence.

LEMMA A.8. *If $g \in C_b^{1,3}$, then $v(\tau, z)$ defined as*

$$v(\tau, z) = Ek(z_\tau)\exp \zeta_{1-\tau}^1(C(\cdot)^{-1}g(\cdot, z.))$$

belongs to $C^{1,2}$ and satisfies the PDE

$$v(0, z) = k(z),$$

$$v_\tau = \tfrac{1}{2}\operatorname{tr} C(1 - \tau)'D_2vC(1 - \tau) + g(1 - \tau, z)'\nabla v.$$

Proof. It is easy to see that we have the "semigroup" property

$$v(\tau + \delta, z) = Ev(\tau, z + \eta_\delta^\delta)\exp \zeta_{1-\tau-\delta}^{1-\tau}(C(\cdot)^{-1}g(\cdot, z + \eta_\cdot^\delta)), \tag{10}$$

where

$$\eta_s^\delta = \int_0^s C(1 - \tau - \delta + u)\,dw_u.$$

Our method will be to exploit the C^2-property of v in z (Lemma A.6), to expand the right-hand side of (10) into terms that will yield $v(\tau, z)$ plus an elliptic operator acting on v. Itô's lemma gives

$$v(\tau + \delta, z) = v(\tau, z) + E\int_0^\delta \nabla v(\tau, z + \eta_s^\delta)'C(1 - \tau - \delta + s)\,dw_s\, e^\zeta$$

$$+ \frac{1}{2}D\int_0^\delta \operatorname{tr} C(1 - \tau - \delta + s)'D_2v\Big|_{\tau, z+\eta_s^\delta} C(1 - \tau - \delta + s)\,ds\, e^\zeta,$$

where ζ is short for the exponent in (10). By Lemma A.7 we can evaluate the first expectation on the right as

$$E\int_0^\delta \nabla v(\tau, z+\eta_s^\delta)' g(1-\tau-\delta+s, z+\eta_s^\delta)\exp \zeta_{1-\tau-\delta}^{1-\tau-\delta+s}(C(\cdot)g(\cdot, z+\eta^\delta))\,ds$$

$$= \delta \nabla v(\tau, z)' g(1-\tau, z) + E\int_0^\delta \nabla v(\tau, z+\eta_s^\delta)' g(1-\tau-\delta+s, z+\eta_s^\delta)\,ds(e^\zeta - 1)$$

(11)

$$+ E\int_0^1 \chi_{s\leqq\delta}[\nabla v(\tau, z+\eta_s^\delta)' g(1-\tau-\delta+s, z+\eta_s^\delta) - \nabla v(\tau, z)' g(1-\tau-\delta+s, z)]\,ds$$

$$+ \int_0^1 \chi_{s\leqq\delta}\nabla v(\tau, z)[g(1-\tau-\delta+s, z) - g(1-\tau, z)]\,ds.$$

The first expectation on the right of (11) is at most

$$\text{const. } E\int_0^\delta e^{\varepsilon|z+\eta_s^\delta|^2}(1+|z+\eta_s^\delta|^2)^{1/2}\,ds\,|e^\zeta - 1|$$

with ζ again the exponent in (10). There exist, by Lemma A.1, constants α, β with $\alpha^{-1} + \beta^{-1} = 1$ and $\alpha > 1$ such that Hölder's inequality implies that this bound is at most

$$\text{const.}\int_0^\delta E^{1/\beta}\, e^{\beta\varepsilon|z+\eta_s^\delta|^2}(1+|z+\eta_s^\delta|^2)^{(1/2)\beta}\,ds E^{1/\alpha}\,|e^\zeta - 1|^\alpha.$$

For $\delta < 1-\tau$ and $Y = \int_0^\delta |C(\cdot)^{-1} g(\cdot, z+\eta_s^\delta)_{1-\tau-\delta+u}|^2\,du$,

$$E|\xi| \leqq Y^{1/2} + \tfrac{1}{2}Y = o(1) \quad \text{as } \delta \to 0,$$

so that $e^\zeta - 1 \to 0$ in measure as $\delta \to 0$. Since $e^{\alpha\zeta}$ for various δ are by Lemma 1 uniformly integrable, they approach 1 in L_α. Hence the first expectation on the right of (11) is $o(\delta)$.

The second expectation on the right of (11) equals, by Taylor,

$$E\int_0^1 \chi_{s\leqq\delta}\int_0^1 (\eta_s^\delta)'\nabla(\nabla v(\tau, z+t\eta_s^\delta)' g(1-\tau-\delta+s, z+t\eta_s^\delta))\,dt\,ds.$$

The first or outer gradient in the integrand is

$$D_2 v(\tau, z+t\eta_s^\delta) g(1-\tau-\delta+s, z+t\eta_s^\delta) + J(1-\tau-\delta+s, z+t\eta_s^\delta)\nabla v(\tau, z+t\eta_s^\delta).$$

Applying the growth bound on g, the boundedness of J, and the orders of $D_2 v$ and ∇v we can bound the second expectation by a constant times

$$E\int_0^1 \chi_{s\leqq\delta}|\eta_s^\delta|(1+|z+\eta_s^\delta)^2)^{1/2}\int_0^1 \exp\varepsilon|z+t\eta_s^\delta|^2\,dt\,ds. \tag{12}$$

Hölder's inequality and the fact that for ε small enough

$$\sup_{0\leqq\delta\leqq 1}\sup_{0\leqq s\leqq\delta}\int_0^1 E|\eta_s^\delta|^2 \exp\varepsilon|z+\eta_s^\delta|^2\,dt < \infty$$

show that (12) $= o(\delta)$. The third term on the right of (11) is clearly $o(\delta)$, since the difference in the integrand is $O(s)$ uniformly. Similar procedures for the trace term give the PDE.

In the next and final lemma of this Appendix we shall suppose that $g_n(t, z)$ are functions of the form

$$g_n(t, z) = g(t, z, u_n(t, z)), \quad u_n \in \mathscr{A}, \quad n = 0, 1, 2, \cdots,$$

with g Lip in its 3rd argument, and satisfying the growth condition

$$|g(t, z, u)|^2 \leqq \kappa(1 + |z|^2).$$

The lemma establishes continuity of the functional

$$J[u_n] = Ek(z_1) \exp \zeta_{1-\tau}^1(g_n(\cdot, z_\cdot)) \quad (k \geqq 0)$$

in the L_2-topology of $\mathscr{A}$.

LEMMA A.9. $u_n \in \mathscr{A}$, $\|u_n - u_0\| \to 0 \Rightarrow J[u_n] \to J[u_0]$.

Proof. Clearly $u_0 \in \mathscr{A}$. We give the argument for $\tau = 1$; for other values it is exactly analogous. Fix z and set

$$\Delta_t^n = \exp \zeta_0^t(g_0(\cdot, z_\cdot)) - \exp \zeta_0^t(g_n(\cdot, z_\cdot)),$$

$$\varphi_N(t) = \text{indicator of } \sup_{0 \leqq u \leqq t} |z_u| < \kappa^{-1} \log N,$$

$$\psi_M(t) = \text{indicator of } \sup_{0 \leqq u \leqq t} \exp \zeta_0^u(g_0(\cdot, z_\cdot)) < M,$$

$$\chi_{MN}(t) = \varphi_N(t)\psi_M(t).$$

Then from $k(z) = O(\exp \kappa|z|)$ we find

$$\begin{aligned} |J[u_0] - J[u_n]| &\leqq Ek(z_1)|\Delta_1^n| \\ &\leqq NE\varphi_N(1)|\Delta_1^n| + \text{const. } E\, e^{\kappa|z_1|}|\Delta_1^n|[1 - \chi_{MN}(1)] \\ &\leqq NE\chi_{MN}(1)|\Delta_1^n| + NE|\Delta_1^n|[1 - \psi_M(1)] \\ &\quad + \text{const. } E\, e^{\kappa|z_1|}|\Delta_1^n|[1 - \chi_{MN}(1)]. \end{aligned}$$

The second and third terms go to zero uniformly in n as *first* $N \uparrow \infty$ and *then* $M \uparrow \infty$, by Hölder's inequality because Lemma A.1 implies $\sup_n E \exp \alpha\zeta_0^1 \cdot (g_n(\cdot, z_\cdot)) < \infty$. Using, for simplicity,

$$\zeta_0^s(g_n) = \zeta_0^s(g_n(\cdot, z_\cdot)), \qquad g_n = g_n(u, z_u),$$

it can be seen that the inequality

$$|\chi_{MN}(s)\Delta_s^n| \leqq \left| \int_0^s (e^{\zeta_0^u(g_0)} g_0 - e^{\zeta_0^u(g_n)} g_n)\chi_{MN}(u)\, dw_u \right|$$

is valid; for if $\chi_{MN}(s) = 0$, the left side is 0, and if $\chi_{MN}(s) = 1$ the equality holds.

Whence for $C = \kappa(1 + \kappa^{-2}\log^2 N)$,

$$|\chi_{MN}(s)\Delta_s^n| \leqq \left| \int_0^s \chi_{MN}(u)\, e^{\zeta_0^u(g_0)}(g_0 - g_n)\, dw_u \right| + \left| \int_0^s \chi_{MN}\Delta_u^n g_n\, dw_u \right|,$$

$$E\chi_{MN}(s)|\Delta_s^n|^2 \leqq 2ME\int_0^s |g_0 - g_n|^2\, du + 2C\int_0^s E\chi_{MN}(u)|\Delta_u^n|^2\, du.$$

By Gronwall's inequality it is enough to show that the first expectation on the right goes to 0 for $s = 1$ as $n \uparrow \infty$. Since g is uniformly Lip in its 3rd argument, we have, with

$$Q(t) = \int_0^t C(u)C(u)'\, du,$$

$$E\int_0^1 |g_0 - g_n|^2\, du \leqq \text{const.}\, E\int_0^1 |u_0(s, z_s) - u_n(s, z_s)|^2\, ds$$

$$\leqq \text{const.} \int_1^\eta \int_\eta^1 |u_0 - u_n|_{s,z+y}^2 \frac{\exp -\frac{1}{2}y'Q(s)y}{(2\pi)^{d/2}\det{}^{1/2} Q(s)^{-1}}\, ds.$$

Choose first η so that the first integral is less than ε; then pick m so that $n > m$ implies

$$\frac{\text{const.}}{(2\pi)^{d/2} \inf\limits_{\eta \leqq s \leqq 1} \det{}^{1/2} Q(s)^{-1}} \|u_0 - u_n\|^2 < \varepsilon.$$

Appendix B. Let $(\xi_t, \mathscr{F}_t)$ be a Wiener process, and $\varphi(t, \omega)$ an $\mathscr{F}_t$-adapted measurable process, with $\int_0^1 |\varphi|^2\, dt < \infty$ a.s. The next portion of this appendix is devoted to this apparently knotty question: When does the (Wald? Girsanov?) identity

$$E \exp \zeta_0^1(\varphi) = 1 \tag{13}$$

actually hold? E. J. McShane [12] has stressed that the answer to this question is very relevant to applications of Girsanov's theorem to estimation and control, especially to the approach we use here. Several methods for establishing (13) are known; however, several are of limited usefulness, and others have been called in question in point of clarity and rigor [12].

Girsanov [1] proved (13) for bounded φ, attributing the result to Maruyama. For a few simple φ, direct integration over Wiener space by Kac's method will prove (13). Lipster and Shiryaev [13] have considered the problem with the upper limit 1 in ζ_0^1 replaced by a Markov time τ of $\{\mathscr{F}_t\}$, or by $+\infty$, and they quote arguments of Novikov [14] suggesting that the sufficient condition he proves,

$$E \exp \frac{1}{2}\int_0^\tau |\varphi|^2\, dt < \infty, \tag{14}$$

is, in the absence of other properties, close to being necessary. Condition (14) will not cover the physically interesting case of linear growth of $|\varphi(t, \omega)|^2$ with $|\xi_t|^2$ or

$\sup_{0 \leqq s \leqq t} |\xi_t|^2$, unless this growth is sufficiently slow. So although it is interesting and nearly necessary in general, (14) is to this extent unsatisfactory.

Girsanov himself [1], in his currently unsettled Lemma 7, tried to give sufficient conditions for (13) based on what are essentially growth and weak uniqueness hypotheses. His Lemma 7 was used by the author [2], and elegantly by Duncan and Varaiya [15], to prove (13) for applications to control theory. It has been suggested, by the referee, by E. J. McShane, and others, that Girsanov's proof of Lemma 7 is at best sloppy. We therefore include a version of this lemma, using (insofar as possible or necessary) the notation, hypotheses, and concepts of Girsanov's paper in its translated form [1]. Page and equation references are to this work.

On page 297 Girsanov introduced a certain weak sense of uniqueness for solutions of stochastic equations such as his (3.1): he called the solution unique if all processes related to some Wiener process by the equation[1] induce the same measure on C_n. One can also formulate a pointwise almost sure, or strong uniqueness, such as would be assured by Lipschitz conditions on A and B, but such a sense is not explicitly used by Girsanov, nor is it needed. Indeed, much of the difficulty people have had with Lemma 7 arises from what seems to be Girsanov's own subsequent imprecise use of his seminal concepts. These have since been developed and expounded, best perhaps in Lipster's and Shiryaev's book [13], into the two notions of strong and weak solutions, each with its own sense of uniqueness. Nor does Girsanov indicate how one might prove weak uniqueness without actually proving the strong form; his remarks about finding solutions of (3.1) refer one to standard works where Lipschitz conditions are used. We shall examine carefully how and where weak uniqueness notions can be used in proving forms of Lemma 7.

Some discussion will precede the statement and proof of the lemma. We shall use the notion of an *Itô process*, and that of a *process of diffusion type*, exactly as does Girsanov, and shall take it for granted that such processes induce measures on the space C_n of continuous R^n-valued functions over $[0, 1]$. With a process $x(t, \omega)$ of diffusion type there can be associated a diffusion matrix $B(\cdot, \cdot)$: $[0, 1] \times C_n \to R^n$, R^n and a drift (or as Girsanov's translator calls it, a vector of transfer) $A(\cdot, \cdot)$: $[0, 1] \times C_n \to R^n$, each "causal" in that their values for $(t, x) \in [0, 1] \times C_n$ do not depend on values of x after t, and a Wiener process $\xi(t, \omega)$ in R^n, such that if $x(\cdot)_\omega$ is the function $\{x(t, \omega): 0 \leqq t \leqq 1\}$, then almost surely for $t \in [0, 1]$,

$$x(t, \omega) = \int_0^t A(s, x(\cdot)_\omega)\, dt + \int_0^t B(s, x(\cdot)_\omega)\, d\xi(s, \omega). \tag{15}$$

Here we of course assume that the entries of $B(\cdot, x(\cdot)_\omega)$ are of integrable square over $[0, 1]$ almost surely, so that the indicated stochastic integral is defined, and that the components of $A(\cdot, x(\cdot)_\omega)$ are integrable almost surely. In this situation we say with Girsanov that $x(t, \omega)$ is a process of diffusion type with drift $A(\cdot, x(\cdot)_\omega)$

[1] There is virtually no loss of generality, and some gain in simplicity, in assuming all initial conditions to be 0; extension to the usual assumption of independent initial conditions is immediate.

and diffusion $B(\cdot, x(\cdot)_\omega)$, with respect to $\xi(t, \omega)$, started at $x(0, \omega)$. We also define

$$U_t = \sigma\{x : x(s) \in A,\ A \text{ Borel in } R^n,\ 0 \leqq s \leqq t\}.$$

This is the σ-algebra of C_n generated by the past up to t.

The role of uniqueness in proving Lemma 7 is elucidated by the following concepts. Let $x(t, \omega)$ be a stochastic process with continuous sample paths, and let τ be a Markov time of $x(\cdot, \cdot)$, i.e., a nonnegative random variable such that any event $\{\tau \geqq t\}$ is measurable on the σ-algebra generated by $\{x(s)_\omega, s \leqq t\}$. Then there is a causal functional $T:[0, 1] \times C_n \to \{0, 1\}$, which we call the *kernel* of τ, such that $T(t, \cdot)$ is U_t-measurable and

$$\chi_{\{\tau \geqq t\}} = T(t, x(\cdot)_\omega).$$

Alternatively, there is a nonincreasing system of Borel sets $B_t \in U_t$ with $\{\tau \geqq t\} = \{\omega : x(\cdot)_\omega \in B_t\}$ and $B_t = \{f \in C_n : T(t, f) = 1\}$. By extension, any such functional defines a Markov time and is called a kernel.

We say that $x(t, \omega)$ is a solution stopped at τ if and only if (15), or equation (3.1) of Girsanov, perhaps holds only up to τ, i.e., almost surely $t \leqq \tau$ implies

$$x(t, \omega) = \int_0^t A(s, x(\cdot)_\omega)\, ds + \int_0^t B(s, x(\cdot)_\omega)\, d\xi_s.$$

A solution stopped at τ is weakly unique if and only if for any process $y(\cdot, \cdot)$ with continuous sample paths, not necessarily defined on the same probability space, and any Markov time κ of $y(\cdot, \cdot)$ such that κ and τ have the same kernel T, the measures induced by (x, τ) and (y, κ) are the same, i.e., if A is Borel and $A \cap \{T(t, y) = 1\} \in U_t$, then $P\{x(\cdot)_\omega \in A, \tau \geqq 1\} = P\{y(\cdot)_\omega \in A, \kappa \geqq t\}$.

We now offer the following modified version of Girsanov's lemma.

LEMMA 7 (after Girsanov). *Let $x(t, \omega)$ be a process of diffusion type with drift $A(\cdot, x(\cdot)_\omega)$ and diffusion $B(\cdot, x(\cdot)_\omega)$ with respect to the Wiener process $(\xi_t, \mathscr{F}_t)$, constituting a solution of equation* (3.1). *Let $\varphi(t, \omega)$ be adapted to $\mathscr{F}_t$ and of integrable square almost surely, and let $y(t, \omega)$ be an Itô process with drift*

$$A(\cdot, y(\cdot)_\omega) - B(\cdot, y(\cdot)_\omega)\varphi(\cdot, \omega)$$

and diffusion $B(\cdot, y(\cdot)_\omega)$ with respect to $(\xi_t, \mathscr{F}_t)$. Suppose that for each $\varepsilon > 0$ there exists $N = N(\varepsilon) < \infty$, and a nonincreasing system (of Borel sets of C_n) $C_N(t)$, $t \in [0, 1]$, such that $C_N(t) \in U_t$ and $s < t \Rightarrow C_N(t) \subset C_N(s)$, and

(a) $D_N(t) \equiv \{y(\cdot)_\omega \in C_N(t)\} \in \mathscr{F}_t$,

(b) $P\{x(\cdot)_\omega \in C_N(1)\} > 1 - \varepsilon$,

(c) $|\varphi(t, \omega)| < N$ *if* $y(\cdot)_\omega \in C_N(t)$,

(d) $f \in C_N(s)$, $f \notin C_N(t)$, $s < t \Rightarrow \tau s < \tau < tf \notin C_N(\tau)$ *and yet* $f \in C_N(v)$ *for* $v < \tau$,

(e) *with T_N the kernel defined by the system $C_N(\cdot)$, the solution $x(t, \omega)$ stopped at $T_N(\cdot, x(\cdot)_\omega)$ is weakly unique.*

Then $E \exp \zeta_0^1(\varphi) = 1$.

Proof. Define, with Girsanov, $\varphi_N(t, \omega) = \varphi(t, \omega)\chi_{D_N(t)}$, and

$$y_N(t, \omega) = y(t, \omega) - \int_0^t B(s, y(\cdot)_\omega)\varphi_N(s, \omega)\, ds. \tag{16}$$

We first prove the set identity stated at the bottom of page 297:

$$\{\omega : y_N(\cdot)_\omega \in C_N(t)\} = D_N(t).$$

To see this note that $\omega \in D_N(t)$ implies $\varphi_N(s,\omega) = \varphi(s,\omega)$ for $0 \leqq s \leqq t$, whence by (16) and monotonicity of $D_N(\cdot)$,

$$y_N(s,\omega) = y(s,\omega) \quad \text{for } 0 \leqq s \leqq t.$$

Since $C_N(t) \in U_t$, we find $y_N(\cdot)_\omega \in C_N(t)$, because by definition $D_N(t) = \{\omega : y(\cdot,\omega) \in C_N(t)\}$. Conversely if $y_N(\cdot)_\omega \in C_N(t)$, then (monotonicity)

$$y_N(\cdot)_\omega \in C_N(s), \qquad 0 \leqq s \leqq t.$$

Now suppose that $y(\cdot,\omega) \notin C_N(t)$. There are two cases according as or not $y(\cdot,\omega) \in C_N(0)$. Since $y(0,\omega) = y_N(0,\omega)$, and $y(\cdot,\omega) \in C_N(0)$ is a condition on $y(0,\omega)$ only, it is clear that $y(\cdot,\omega) \in C_N(0)$ if and only if $y_N(\cdot)_\omega \in C_N(0)$. So if $t = 0$ there is nothing to prove. If $t > 0$, in the former case $y(\cdot,\omega) \in C_N(0)$, and by property (d) there is a τ with $0 < \tau \leqq t$ such that

$$y(\cdot,\omega) \in C_N(s), \qquad 0 \leqq s < \tau,$$
$$y(\cdot,\omega) \notin C_N(\tau).$$

This implies that $\omega \in D_N(s)$ for $0 \leqq s < \tau$ so that also

$$\varphi_N(s,\omega) = \varphi(s,\omega), \qquad 0 \leqq s < \tau,$$
$$y_N(s,\omega) = y(s,\omega), \qquad 0 \leqq s < \tau,$$

and by continuity, $y_N(\tau,\omega) = y(\tau,\omega)$, and thus $y_N(\cdot)_\omega \in C_N(\tau)$, contradicting the hypothesis. In the latter case $y(\cdot,\omega) \notin C_N(0)$, whence $y_N(\cdot)_\omega \notin C_N(s)$ for all $s \geqq 0$, again contradicting $y_N(\cdot)_\omega \in C_N(t)$.

Returning now to the main line of proof, we see that since φ_N is bounded, $E \exp \zeta_0^1(\varphi_N) = 1$. Hence by Girsanov's Theorem 1 the functions

$$\xi(t,\omega) - \int_0^t \varphi_N(s,\omega)\, ds = \tilde{\xi}_N(t,\omega)$$

under the measure $d\tilde{P}_N = \exp \zeta_0^1(\varphi_N)\, dP$ form a Wiener process. Under the same measure, the pair $\{y_N(\cdot,\cdot), T_N(\cdot, y(\cdot)_\omega)\}$ form a solution of (3.1) stopped at $T_N(\cdot, y(\cdot)_\omega)$. Hence by the uniqueness of these stopped solutions,

$$\begin{aligned} P\{x(\cdot)_\omega \in C_N(1)\} &= \tilde{P}_N\{y_N(\cdot)_\omega \in C_N(1)\} \\ &= \tilde{P}_N\{y(\cdot)_\omega \in C_N(1)\} \\ &= \tilde{P}_N\{D_N(1)\} \\ &= \int_{D_N(1)} \exp \zeta_0^1(\varphi) P(d\omega). \end{aligned}$$

The probability on the left above can be made arbitrarily close to one by a sufficiently large choice of N, so that $\tilde{P}(\Omega) = 1$.

We believe that the above arguments show that Girsanov's proof, although untidy and, in its translated form, beset by typos, was basically correct except for

the portion in lines 13–16 on page 298 using uniqueness, which seems to be confused and incomplete. The point, which I owe to the referee, is that $y(t, \omega)$ under $\tilde{P}_N$ is known to satisfy the desired equation with respect to the Wiener process $\tilde{\xi}_N$ only while $y(\cdot)_\xi \in C_N(t)$. It is perhaps plausible, but nevertheless as we see it not obvious nor an immediate sequitur from Girsanov's assumptions, that

$$\tilde{P}_N\{y(\cdot)_\omega \in C_N(1)\} = P\{x(\cdot)_\omega \in C_N(1)\}.$$

This equality is assured by formulating the uniqueness hypothesis in terms of the kernels T_N associated with $C_N(\cdot)$.

Remark 1. In many applications, including that of this paper, the conditions $A = 0$, $B(t, f) = B(t)$, $\det B(t) \neq 0$ obtain; these obviate the uniqueness assumption (e) on the stopped processes.

Remark 2. Varaiya and Duncan [15] have given a proof of (13) from the at most linear growth of $|\varphi(t, \omega)|^2$ with $|\xi_t|^2$, using Girsanov's Lemma 7. It is also possible to extend the random time change argument of Kailath and Zakai [16] to the vector case; when φ is of linear growth their uniform boundedness assumption is gratuitous. It has seemed to us that in this physically motivated special case there should be a simple proof. Such a proof is now sketched.

With $T(t) = \int_0^t |\varphi|^2 \, ds$ and new Markov times

$$\tau_N = \inf\{t : \zeta_0^t \vee T(t) > N\} \wedge 1$$

set $\alpha(t) = \exp \zeta_0^t(\varphi)$, $\alpha = \alpha(1)$, $\alpha_N = \alpha(\tau_N)$. It can be seen that $\alpha_N \to \alpha$ in probability, and $E\alpha_N = 1$. By a modification of the argument of Lemma A1 herein, we can show that if there is a constant β such that a.s. for every $t \in [0, 1]$,

$$|\varphi(t, \omega)|^2 \leqq \beta(1 + \sup_{0 \leqq s \leqq t} |\xi_s|^2),$$

then there exists $\lambda > 1$ such that

$$\sup_N E\alpha_N^\lambda < \infty.$$

It follows that α_N are uniformly integrable functions tending to α in measure, whence also in the mean, i.e., $E|\alpha - \alpha_N| \to 0$, so $E\alpha = 1$.

Appendix C. Finally, in conversation, A. V. Balakrishnan has asked on how large a set of time points the process $\{s(t)'z_t, \tilde{P}\}$ can vanish (notation as in § 3). This question is very natural, in view of the optimal law (2). The answer is the expected one, namely, that these zeros are with probability one a set of Lebesgue measure zero. This result can be proved in several ways, the easiest of which just reduces it to the same property (well known) for one-dimensional Brownian motion. For simplicity we assume that the diffusion matrix $C(\cdot)$ is C^1.

Let $h : [s, 1] \to R^d$ and $f : [s, 1] \to R$ be C^1 curves, with $h(\cdot)$ not passing through the origin, and consider, in the notation of § 3, a process $z_t = z + \int_s^t C(u) \, dw_u$, $1 \geqq t \geqq s$, under the measure $d\tilde{P} = \exp \zeta \, dP$ with

$$\zeta = \int_s^1 C(u)^{-1} g(u, z_u) \, dw_u - \frac{1}{2} \int_s^1 |C(u)^{-1} g(u, z_u)|^2 \, du.$$

We show that

$$\tilde{P}\{\text{meas}\,(u \in [s, 1] : h(u)'z_u = f(u))3 = 0\} = 1.$$

By Ito's lemma,

$$z_t = z + C(t)w_t - C(s)w_s + \int_s^t \dot{C}(u)w_u\,du,$$

so the zeros in question are the same as those of

$$\frac{h(t)'C(t)w_t}{|h(t)'C(t)|} + \frac{\int_s^t \dot{C}(u)w_u\,du + h(t)'[z - C(s)w_s - f(t)]}{|h(t)'C(t)|},$$

since we are assuming as before that $CC' > 0$. This is of the form $b_t - B(t, \omega)$ with $B(\cdot, \omega)$ a w-nonanticipating C_b^1 function, and b a Wiener process in one dimension under P. Now let for $\varepsilon \geqq 0$,

$$A_\varepsilon = \{w : \text{meas}\,(u \in [s, 1] : h(u)'z_u = f(u)) > \varepsilon\}.$$

By Girsanov's theorem [1] $\{b_\cdot - B(\cdot, \omega), P\}$ is equivalent to Wiener's process, so $P(A_\varepsilon) = 0$ for $\varepsilon \geqq 0$. However, by Lemma A.1 and Hölder's inequality, for some $\alpha > 1$ and constant K,

$$\tilde{P}(A_\varepsilon) \leqq \left(\int_{A_\varepsilon} e^{\alpha\zeta}\,dP\right)^{1/\alpha} P^{\alpha/(\alpha-1)}(A_\varepsilon)$$

$$\leqq K \exp\frac{|z|^2}{\alpha} P^{\alpha/(\alpha-1)}(A_\varepsilon).$$

Thus $\tilde{P}(A_0) = 0$.

To apply this result to $s(t)'z_t$ we have only to take $f \equiv 0$, and to note that $k \neq 0$ implies that $s(t) \neq 0$.

REFERENCES

[1] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theor. Probability Appl., 5 (1960), pp. 285–301.

[2] V. E. BENEŠ, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–72.

[3] M. H. A. DAVIS AND P. P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.

[4] V. E. BENEŠ, *Girsanov functionals, and optimal bang-bang laws for final value stochastic control*, Stochastic Processes and their Appls., 2 (1974), pp. 127–140.

[5] R. C. DAVIS, *Stochastic final-value control systems with a fuel constraint*, J. Math. Anal. Appl., 21 (1968), pp. 62–78.

[6] A. VAN GELDER, J. DUNN AND J. MENDELSOHN, *The final value optimal stochastic control problem with bounded controller*, Proc. 1966 JACC, pp. 441–449.

[7] H. P. MCKEAN, *Stochastic Integrals*, Academic Press, New York, 1969.

[8] M. WONHAM, *On the separation theorem of stochastic control*, this Journal, 6 (1968), pp. 312–326.

[9] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, N.J., 1964.

[10] H. J. LANDAU AND L. A. SHEPP, *On the supremum of a Gaussian process*, Sankhyā Ser. A, 32 (1970), pp. 369–378.

[11] X. FERNIQUE, *Intégrabilité des vecteurs Gaussiens*, C. R. Acad. Sci. Paris, 270 (1970), pp. 1698–99.

[12] E. J. MCSHANE, personal communication, and in a talk presented at A.M.S. meeting, Cleveland, Ohio, 1972.

[13] R. Sh. Liptser and A. N. Shiryaev, *Statistics of Random Processes*, Izd. Nauka, Moscow, 1974 (In Russian).

[14] A. A. Novikov, *On an identity for stochastic integrals*, Theory. Probability Appl., 17 (1972), pp. 717–720.

[15] T. E. Duncan and P. P. Varaiya, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.

[16] T. Kailath and M. Zakai, *Absolute continuity and Radon–Nikodym derivatives for certain measures relative to Wiener measure*, Ann. Math. Statist., 42 (1971), pp. 130–140.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

THE EXISTENCE OF VALUE IN STOCHASTIC DIFFERENTIAL GAMES*

ROBERT ELLIOTT†

Abstract. Using the techniques of Davis and Varaiya [3], [4] a two-person zero sum differential game is considered, whose dynamics are interpreted using the Girsanov measure transformation method. If the Isaacs condition holds it is shown that the upper and lower values of the game are equal and there is a saddle point in feedback strategies. The central point of the mathematics is that analogues of the time derivative and gradient of the upper value function are constructed using martingale methods; because the Hamiltonian satisfies a saddle condition at each point these then also give the lower value.

1. Introduction. The following is an extension to differential games of the work of Davis and Varaiya [3], [4]. In particular upper and lower values for two-person zero sum games are introduced and it is shown that if the Isaacs condition holds then the upper and lower values are equal and there is a saddle point in feedback strategies. This result is stronger than the saddle-point result established in [5] and is probably the best possible, because Lemma 4.4 shows that the Isaacs condition must be satisfied at all relevant points. Solutions of the stochastic dynamical equations are defined using the Girsanov measure transformation method, and martingale decomposition results are quoted from [8] and [4] to obtain the analogue of the Hamiltonian.

We suppose the evolution of the system is described by a stochastic functional differential equation of the form

$$dx_t = f(t, x, y, z)\,dt + \sigma(x, t)\,dB_t. \tag{1.1}$$

Here $t \in [0, 1]$ and B is an m-dimensional Brownian motion. Write $\mathscr{C}$ for the space of continuous functions from $[0, 1]$ to R^m. x denotes a member of $\mathscr{C}$ and x_t denotes the value of x at t. We wish to consider a solution of (1.1) which at time 0 has an initial value $x_0 \in R^m$. The drift term f depends at time t on the past $\{x_s : s \leqq t\}$ of the process. The payoff is of the form

$$P(y, z) = E\left\{ g(x(1)) + \int_0^1 h(t, x, y, z)\,dt \right\}, \tag{1.2}$$

where

(i) g and h are real-valued,

(ii) $0 \leqq g \leqq k$ and $0 \leqq h \leqq k$ for some constant k,

(iii) g and h satisfy the measurability properties described below.

A player J_1 chooses a feedback control $y(t, x)$ with values in a compact metric space Y with the object of maximizing the payoff and a player J_2 chooses a feedback control $z(t, x)$ with values in a compact metric space Z with the object of minimizing the payoff. At time t the controls are allowed to depend on the past of the process.

* Received by the editors September 24, 1974, and in revised form December 9, 1974.

† Department of Pure Mathematics, University of Hull, Hull, England.

2. Notation. The situation treated below is similar to that of Davis and Varaiya [3], [4] so we continue a description of their notation, slightly modified.

Write $\mathscr{F}_t$ for the σ-field of $\mathscr{C}$ generated by $\{x_s : x \in \mathscr{C}, s \leqq t\}$. We suppose the m-dimensional Brownian motion B_t is separable and defined on an underlying measure space $(\Omega, \mathscr{A}, \mu)$. Write $\mathscr{D}$ for the σ-field of $[0, 1] \times \mathscr{C}$ consisting of subsets D which have the property that $D \cap \{\{t\} \times \mathscr{C}\} \in \mathscr{F}_t$ for each $t \in [0, 1]$ and $D \cap \{[0, 1] \times \{x\}\}$ is Lebesgue measurable. Beneš [1] proves that a function is $\mathscr{D}$ measurable if and only if $f(t, \cdot)$ is $\mathscr{F}_t$ measurable for each t and $f(\cdot, x)$ is Lebesgue measurable for each x.

The $m \times m$ matrix $\sigma = (\sigma_{ij})$ satisfies

(i) for $1 \leqq i, j \leqq m$, $\sigma_{ij} : [0, 1] \times \mathscr{C} \to R$ is measurable with respect to $\mathscr{D}$,

(ii) $\sigma(t, x)$ is nonsingular,

(iii) each σ_{ij} satisfies a uniform Lipschitz condition in x.

The equation

$$dx_t = \sigma(t, x)\, dB_t, \qquad x(0) = x_0 \in R^m$$

then has a unique solution x_t and it induces a measure P_0 on its sample space $(\mathscr{C}, \mathscr{F}_1)$ according to the formula

$$P_0 A = \mu\{\omega : x(\omega) \in A\}, \qquad A \in \mathscr{F}_1.$$

Write Φ for the set of functions $\phi : [0, 1] \times \mathscr{C} \to R^m$ which are measurable with respect to $\mathscr{D}$ and which satisfy

$$|\phi(t, x)| \leqq M(1 + \|x\|).$$

Write a_t for the matrix $\sigma(t, x)\sigma'(t, x)$ and for $\phi \in \Phi$ write

$$\zeta(\phi) = \int_0^1 \phi_t \cdot a_t^{-1}\, dx_t - \tfrac{1}{2} \int_0^1 \phi_t \cdot a_t^{-1} \phi_t\, dt,$$

where

$$\phi_t = \phi(t, x).$$

Define the measure P_ϕ on $(\mathscr{C}, \mathscr{F}_1)$ by: $P_\phi A = \int_A \exp(\zeta(\phi))\, dP_0$, $A \in \mathscr{F}_1$.

Then we can quote the following results from Girsanov [7] and Beneš [2].

LEMMA 2.1.

(i) *P_ϕ is a probability measure,*

(ii) *P_ϕ is mutually absolutely continuous with respect to P_0,*

(iii) *$\{\omega_t, t \in [0, 1]\}$ is a Brownian motion under P_ϕ, where*

$$d\omega_t = dB_t - \sigma^{-1}(t, x)\phi(t, x)\, dt$$

$$= \sigma^{-1}(t, x)(dx_t - \phi(t, x))\, dt.$$

$\mathscr{Y}$ (resp. $\mathscr{Z}$) is the σ-field of Borel sets of Y (resp. Z).

An admissible feedback control for J_1 is a measurable function

$$y : ([0, 1] \times \mathscr{C}, \mathscr{D}) \to (Y, \mathscr{Y})$$

and an admissible feedback control for J_2 is a measurable function

$$z : ([0, 1] \times \mathscr{C}, \mathscr{D}) \to (Z, \mathscr{Z}).$$

Write $\mathcal{M}_1$ (resp. $\mathcal{M}_2$) for the admissible controls for J_1 (resp. J_2). R^m is always supposed given the Borel σ-field $\mathcal{R}^m$ and the drift function f is supposed to satisfy:

(i) $f : [0, 1] \times \mathcal{C} \times Y \times Z \to R^m$ is measurable with respect to the σ-field $\mathcal{D} * \mathcal{Y} * \mathcal{Z}$.

(ii) for each $(t, x) \in [0, 1] \times \mathcal{C}$, $f(t, x, \cdot\cdot\cdot)$ is continuous on $Y \times Z$.

(iii) there exists a constant K such that for all $(t, x, y, z) \in [0, 1] \times \mathcal{C} \times Y \times Z$,

$$|f(t, x, y, z)| \leqq K(1 + \|x\|),$$

where $\|\cdot\|$ is the uniform norm in $\mathcal{C}$.

For $y \in \mathcal{M}_1$ and $z \in \mathcal{M}_2$ and $(t, x) \in [0, 1] \times \mathcal{C}$ write

$$f^{yz}(t, x) = f(t, x, y(t, x), z(t, x)),$$

$$h^{yz}(t, x) = h(t, x, y(t, x), z(t, x)).$$

We see $f^{yz} \in \Phi$, so writing $P_{f^{yz}}$ as P_{yz}, Lemma 2.1 can be used to say that under measure P_{yz},

$$dx_t = f(t, x, y(t, x), z(t, x))\, dt + \sigma(t, z)\, dB_t,$$

where $\{B_t\}$ is a Brownian motion. Lemma 2.1, therefore, enables a solution of the dynamical equations (1.1) to be interpreted under very general hypotheses on f and σ.

Suppose E_{yz} denotes the expectation with respect to P_{yz}. Then the payoff corresponding to $y \in \mathcal{M}_1$ and $z \in \mathcal{M}_2$ is

$$P(y, z) = E_{yz}\left(g(x(1)) + \int h(t, x, y(t, x), z(t, x))\, dt\right).$$

3. Upper and lower values. Suppose J_2 has chosen $z \in \mathcal{M}_2$. Then for any $y \in \mathcal{M}_1$ the expected remaining payoff from time $t \in [0, 1]$ is

$$\psi_{yz}(t) = E_{yz}\left(g(x(1)) + \int_t^1 h^{yz}(s, x)\, ds | \mathcal{F}_t\right).$$

Now $\{\psi_{yz} : y \in \mathcal{M}_1\}$ is a subset of $L^\infty(\mathcal{C}, \mathcal{F}_t, P_0)$ bounded above by $2k$. By Theorem 1V.8.23 of [6], $L^\infty(\mathcal{C}, \mathcal{F}_t, P_0)$ is a complete lattice so the supremum

$$W_t^z = \bigvee_{y \in \mathcal{M}_1} \psi_{yz}(t)$$

exists in L^∞. Note that

$$W_1^z = g(x(1)) \quad \text{a.s.}$$

Define

$$P_z^* = W_0^z = \sup_{y \in \mathcal{M}_1} P(y, z).$$

The results of [4] can be adapted (we are now working with a supremum instead of an infimum) to deduce the following.

LEMMA 3.1. *For each* $z \in \mathcal{M}_2$:

(i) *there exist processes* $\{\wedge W_t^z\}, \{\nabla W_t^z\}$ *with values in* R *and* R^m *such that*

$$\int_0^1 |\nabla W_t^z|^2 \, dt < \infty \quad a.s.\ (P_0),$$

$$E \int_0^1 | \wedge W_t^z | \, dt < \infty$$

and

$$W_t^z = P_z^* + \int_0^t \wedge W_s^z \, ds + \int_0^t \nabla W_s^z \, dx_s \quad a.s.\ (P_0),$$

(ii) *for any* $y \in \mathcal{M}_1$,

$$\wedge W_t^z + \nabla W_t^z \cdot f(t, x, y_t, z_t) + h(t, x, y_t, z_t) \leqq 0$$

for almost all (t, x), *and* $y_z^* \in \mathcal{M}_1$ *is the optimal reply to* $z \in \mathcal{M}_2$ *if and only if equality holds almost everywhere in the above when* $y = y_z^*$.

As in [3] the optimal reply to $z \in \mathcal{M}_2$ can then be shown to be

$$y_z^*(t, x) = y_z^*(t, x, \nabla W_t^z(t, x)),$$

where $y_z^*(t, x, p)$ is the measurable function from $([0, 1] \times \mathcal{C} \times R^m, \mathcal{D} * \mathcal{R}^m)$ to $(Y, \mathcal{y})$ maximizing

$$p \cdot f(t, x, y, z(t, x)) + h(t, x, y, z(t, x)).$$

Consequently, we can conclude as follows.

LEMMA 3.2. *For each* $z \in \mathcal{M}_2$, J_1 *has an optimal reply* $y_z^* \in \mathcal{M}_1$ *such that*

$$P(y_z^*, z) = \sup_{y \in \mathcal{M}_1} P(y, z).$$

Now consider J_2 who in the "upper game" that we are considering must choose his control $z \in \mathcal{M}_2$ first. The problem is: can J_2 choose $z \in \mathcal{M}_2$ to attain

$$\inf_{z \in \mathcal{M}_2} \sup_{y \in \mathcal{M}_1} P(y, z) = \inf_{z \in \mathcal{M}_2} P(y_z^*, z)?$$

For any $z \in \mathcal{M}_2$ and $t \in [0, 1]$, if we assume J_1 plays his optimal reply, the remaining payoff from time t onwards is

$$\psi_z = E_{y_z^*, z}\left(g(x(1)) + \int_t^1 h^{y_z^* z}(s, x) \, ds | \mathcal{F}_t \right).$$

Again, because $L^\infty(\mathcal{C}, \mathcal{F}_t, P_0)$ is a complete lattice, the infimum

$$W_t^+ = \bigwedge_{z \in \mathcal{M}_2} \psi_z(t)$$

exists in $L^\infty(\mathcal{C}, \mathcal{F}_t, P_0)$.

DEFINITION 3.3. W_t^+ is the *upper value function of the differential game.* Notice that $W_1^+ = g(x(1))$ a.s. (P_0) and define $P^+ = W_0^+ = \inf_{z \in \mathcal{M}_2} P(y_z^*, z)$.

LEMMA 3.4. *For each* $z \in \mathcal{M}_2$, $\delta > 0$ *and* $t \in [0, 1]$,

$$W_t^+ \leqq E_{y_{z,z}^*, z}\left[\int_t^{t+\delta} h^{y_{z,z}^*, z}\, ds | \mathcal{F}_t\right] + E_{y_{z,z}^*, z}[W_{t+\delta}^+ | \mathcal{F}_t] \quad a.s.(P_0).$$

This result is proved by modifying the method of Theorem 3.1 of [4].

LEMMA 3.5. W_t^+ *can be expressed as the difference of a martingale and an absolutely continuous increasing process.*

Proof. The proof is adapted from Lemma 5.1 and Theorem 5.2 of [4].

Choose a sequence $\{z_n\} \subset \mathcal{M}_2$ such that $P(y_{z_n}^*, z_n) = \psi_{z_n}(0)$ is monotonic decreasing to $W_0^+ = P^+$. Then $f^{y_{z_n, z_n}} \in \Phi$ for each n. As in Theorem 2.2 of [4], the set $\{\exp \xi(\phi) : \phi \in \Phi\}$ is weakly compact in $L^1(\mathcal{C}, \mathcal{F}, P_0)$ so there is a subsequence, again denoted by $\{z_n\}$, and a $\psi \in \Phi$ such that $\exp \xi(f^{y_{z_n, z_n}^*})$ converges to $\rho^* = \exp \xi(\psi)$ weakly in $L^1(\mathcal{C}, \mathcal{F}, P_0)$.

Define P^* by putting $dP^* = \rho^*\, dP$.

The proofs of Lemma 5.1 and Theorem 5.2 of [4] then go through to show W_t^+ has a right continuous modification, which we suppose is the version taken. Also $\{W_t - E^*[W_1 | \mathcal{F}_t], \mathcal{F}_t, P^*\}$ is a potential and so, from Theorem VII T 29 of Meyer [8], $W_t^+ - E^*[W_1 | \mathcal{F}_t]$ can be expressed as

$$E^*[A_1 | \mathcal{F}_t] - A_t,$$

where $A_t = \int_0^t \alpha_s\, ds$ for a process $\alpha_s \in L^1(P^*)$.

Further, it is known that the martingale $E^*[W_1 + A_1 | \mathcal{F}_t]$ can be expressed as a stochastic integral of the Brownian motion

$$d\omega = \rho^{-1}(dx - \psi_t\, dt),$$

and so, as in [4] we have the following representation.

LEMMA 3.6.

(i) *There are processes* $\{\bigwedge W_t^+\}$, $\{\nabla W_t^+\}$ *taking values in* R *and* R^m *respectively, adapted to* $\mathcal{F}_t$, *such that*

$$\int_0^1 |\nabla W_t^+|^2\, dt < \infty \quad a.s.\ (P_0),$$

$$E \int_0^1 |\bigwedge W_t^+|\, dt < \infty,$$

and $W_t^+ = P^+ + \int_0^t \bigwedge W_s^+\, ds + \int_0^t \nabla W_s^+\, dx_s \quad a.s.\ (P_0)$.

(ii) *For any* $z \in \mathcal{M}_2$,

$$\bigwedge W_t^+ + \nabla W_t^+ \cdot f(t, x, y_z^*(t, x), z(t, x)) + h(t, x, y_z^*(t, x), z(t, x)) \geqq 0$$

for almost all (t, x). $z^* \in \mathcal{M}_2$ *is optimal if and only if equality holds in the above with* $z = z^*$.

For $(t, x, p) \in [0, 1] \times \mathcal{C} \times R^m$ we introduce the Hamiltonian:

$$H(t, x, p; y, z) = p \cdot f(t, x, y, z) + h(t, x, y, z).$$

Then for fixed (t, x, p), H is continuous on $Y \times Z$. Suppose S (resp. T) is a countable dense subset of Y (resp. Z). Then for $z \in Z$,

$$H(t, x, p; y_z^*, z) = \max_{y \in Y} H(t, x, p; y, z) = \sup_{y \in S} H(t, x, p; y, z)$$

is continuous in z.

Further, for fixed (t, x, p) and $z \in Z$,

$$\{(t, x, p) : \max_{y \in Y} H(t, x, p; y, z) < a\} = \bigcup_{y \in S} \{(t, x, p); H(t, x, p; y, z)\} < a$$

and so, $\max_{y \in Y} H(t, x, p; y, z)$ is measurable with respect to $\mathscr{D} * \mathscr{R}^m$ in (t, x, p).

Now for fixed (t, x, p)

$$\min_{z \in Z} \max_{y \in Y} H(t, x, p; y, z) = \inf_{z \in T} \sup_{y \in S} H(t, x, p; y, z),$$

so

$$\{(t, x, p) : \min_{z} \max_{y} H(t, x, p; y, z) < a\} = \bigcup_{z \in T} \{(t, x, p) : \max_{y} H(t, x, p; y, z) < a\},$$

and so by Lemma 1 of Beneš [1] there is a measurable function

$$z^* : ([0, 1] \times \mathscr{C} \times R^m, \mathscr{D} * \mathscr{R}^m) \to (Y, \mathscr{Y})$$

such that

$$H(t, x, p; y_{z^*}^*(t, x, p), z^*(t, x, p)) = \min_{z} \max_{y} H(t, x, p; y, z)$$

for all (t, x, p).

If J_2 is to choose his feedback control first, therefore, the best he can do is to play $z^*(t, x) =)z^*(t, x, \nabla W^+(t, x))$ because then, as in Theorem 1 of Davis [3] it can be shown that

$$\bigwedge W_t^+ + W_t^+ \cdot f(t, x, y_{z^*}^*(t, x), z^*(t, x)) + h(t, x, y_{z^*}^*(t, x), z^*(t, x)) = 0$$

and

$$P(y_{z^*}^*, z^*) = \inf_{z \in \mathscr{M}_2} \sup_{y \in \mathscr{M}_1} P(y, z).$$

Therefore, we can summarize the above by stating the following result.

THEOREM 3.7. *Consider a two-person zero sum stochastic differential game whose dynamics are described by* (1.1) *and whose payoff is given by* (1.2). *If the minimizing player J_2 must choose a feedback control first, then the players can choose controls $z^* \in \mathscr{M}_2$, $y_{z^*}^* \in \mathscr{M}_1$ which attain the "upper value"*

$$\inf_{z \in \mathscr{M}_2} \sup_{z \in \mathscr{M}_2} P(y, z) = P(y_{z^*}^*, z^*).$$

Remarks 3.8. If the maximizing player J_1 must choose his feedback control first, then there are controls $y^* \in \mathcal{M}_1$, $z^*_{y^*} \in \mathcal{M}_2$ which attain the lower value

$$\sup_{y \in \mathcal{M}_2} \inf_{z \in \mathcal{M}_1} P(y, z) = P(y^*, z^*_{y^*}).$$

4. The Isaacs condition. $H[t, x, p; y, z]$ is the Hamiltonian defined in § 3.

DEFINITION 4.1. We say the *Isaacs condition holds* if for $(t, x, p) \in [0, 1] \times \mathcal{C} \times R^m$,

$$\max_{y \in Y} \min_{z \in Z} H(t, x, p; y, z) = \min_{z \in Z} \max_{y \in Y} H(t, x, p; y, z).$$

For a fixed (t, x, p) let y^*_z be such that $\max_y H[t, x, p; z] = H[t, x, p; y^*_z, z]$ and let z^* be such that

$$\min_z H(t, x, p; y^*_z, z) = H(t, x, p; y^*_{z^*}, z^*) = \min_z \max_y H(t, x, p; y, z).$$

Similarly $z^*_{y^*}$ and y^* are such that

$$\max_y H(t, x, p; y, z^*_y) = H(t, x, p; y^*, z^*_{y^*})$$

$$= \max_y \min_z H(t, x, p; y, z).$$

Consequently

$$H(t, x, p; y^*, z^*) \leqq \min_z \max_y H(t, x, p; y, z)$$

and $H(t, x, p; y^*, z^*) \geqq \max_y \min_z H(t, x, p; y, z)$, so if the Isaacs condition holds, (y^*, z^*) is a saddle point for the function $H(t, x, p; y, z)$.

Now in the discussion of the upper value in § 3, the control $z \in \mathcal{M}_2$ was optimal for J_2 playing first if and only if

$$\bigwedge W_t^+ + \nabla W_t^+ \cdot f(t, x, y^*_{z^*}(t, x), z^*(t, x)) + h(t, x, y^*_{z^*}(t, x), z^*(t, x))$$

$$= \min_{z \in \mathcal{M}_2} \left(\bigwedge W_t^+ + \nabla W_t^+ \cdot f(t, x, y^*_z(t, x), z(t, x)) + h(t, x, y^*_z(t, x), z(t, x)) \right)$$

$$= \min_{z \in \mathcal{M}_2} \max_{y \in \mathcal{M}_1} \left(\bigwedge W_t^+ + \nabla W_t^+ \cdot f(t, x, y(t, x), z(t, x)) + h(t, x, y(t, x), z(t, x)) \right)$$

$$= 0.$$

We recall that z^* had the form $z^*(t, x, \nabla W_t^+)$, where $z^*(t, x, p)$ is the measurable function that attained the minimum of

$$\max_y (p \cdot f(t, x, y, z) + h(t, x, y, z)).$$

Suppose now that $y^* \in \mathcal{M}_1$ is the feedback control $y^*(t, x) = y^*(t, x, \nabla W_t^+)$, where $y^*(t, x, p)$ is the measurable function that attains the maximum of

$$\min_{z \in Z} (p \cdot f(t, x, y, z) + h(t, x, y, z)).$$

From the remarks above we see that if the Isaacs condition holds the pair of feedback controls $(y^*, z^*) \in \mathcal{M}_1 \times \mathcal{M}_2$ is a saddle point for the Hamiltonian $H(t, x, \nabla W_t^+; y, z)$ for almost all (t, x).

Consequently for any other $(y, z) \in \mathcal{M}_1 \times \mathcal{M}_2$ we have,

$$\begin{aligned} \bigwedge W_t^+ + H(t, x, \nabla W_t^+; y(t, x), z^*(t, x)) &\leqq \bigwedge W_t^+ + H(t, x, \nabla W_t^+; y_{z^*}^*(t, x), z^*(t, x)) \\ &= \bigwedge W_t^+ + H(t, x, \nabla W_t^+; y^*(t, x), z^*(t, x)) \\ &= 0 \\ &\leqq \bigwedge W_t^+ + H(t, x, \nabla W_t^+; y^*(t, x), z(t, x)). \end{aligned}$$

We now quote Theorem 5.1 of [4] in a form adapted to our differential game.

THEOREM 4.2. *The admissible control $z^* \in \mathcal{M}_2$ is optimal for J_2 in reply to $y^* \in \mathcal{M}_1$ if there is a constant J^* and processes $\{\eta_t\} \subset R$ and $\{\xi_t\} \subset R^m$ adapted to $\mathcal{F}_t$ and satisfying:*

(i) $\int_0^1 |\xi_t|^2 \, dt < \infty$ *a.s.* (P_0),

(ii) $E \int_0^1 \xi_t \, dx_t = 0$,

(iii) $\chi(1) = g(x(1))$ *a.s. where* $\chi(t) = J^* + \int_0^t \eta_s \, ds + \int_0^t \xi_s \, dx_s$,

(iv) $\eta_t + \xi_t \cdot f_t^{y^*, z} + h_t^{y^*, z} \geqq 0 = \eta_t + \xi_t \cdot f_t^{y^*, z^*} + h_t^{y^*, z^*}$, *for almost all* (t, x) *and each* $z \in \mathcal{M}_2$. *Then* $\inf_{z \in \mathcal{M}_2} \psi_{y^*, z}(t) = \chi(t)$ *and* $P(y^*, z^*)$ *is the minimum payoff in reply to* $y^* \in \mathcal{M}_1$.

We can now state our main result.

THEOREM 4.3. *If the Isaacs condition holds, then there is a pair of admissible feedback controls $(y^*, z^*) \in \mathcal{M}_1 \times \mathcal{M}_2$ which give a saddle point for the payoff*

$$P(y, z^*) \leqq P(y^*, z^*) \leqq P(y^*, z).$$

Consequently, the upper value function of the differential game is almost surely equal to the lower value function.

Proof. We observe that, taking $J^* = P^+$ the processes $\{\bigwedge W_t^+\}$, $\{\nabla W_t^+\}$ satisfy the hypotheses of Theorem 4.2 and so $P(y^*, z^*) = \inf_{z \in \mathcal{M}_2} P(y^*, z)$.

We already know that

$$P(y^*, z^*) = \sup_{y \in \mathcal{M}_1} P(y, z^*)$$

and so the result is proved.

Finally we prove what is almost the converse to the above theorem.

LEMMA 4.4. *Suppose the upper value function equals the lower value function for almost all* (t, x), *that is,*

$$W_t^+ = \inf_{z\in\mathcal{M}_2}\sup_{y\in\mathcal{M}_1}\psi_{yz}(t) = \sup_{y\in\mathcal{M}_1}\inf_{z\in\mathcal{M}_2}\psi_{yz}(t) = W_t^-.$$

Then for almost all $(t, x)\in[0, 1]\times\mathscr{C}$,

$$\min_{z\in Z}\max_{y\in Y} H(t, x, \nabla W_t^+; y, z) = \max_{y\in Y}\min_{z\in Z} H(t, x, \nabla W_t^+; y, z).$$

Proof. The constructive method of obtaining the optimal controls described in § 3 implies that the respective infima and suprema are attained, so the hypothesis implies there are admissible controls $(y^*, z^*)\in\mathcal{M}_1\times\mathcal{M}_2$ such that

$$\psi_{y^*,z^*}(t) = W_t^+ = W_t^-.$$

Now as in Lemma 3.6,

$$W_t^+ = P^+ + \int_0^t \wedge W_s^+\, ds + \int_0^t \nabla W_s^+ \cdot dx_s$$

and

$$\wedge W_t^+ + \nabla W_t^+ \cdot f_t^{y^*,z} + h_t^{y^*,z} \geqq 0 = \wedge W_t^+ + \nabla W_t^+ \cdot f_t^{y^*,z^*} + h_t^{y^*,z^*} \tag{4.1}$$

for almost all (t, x) and each $z\in\mathcal{M}_2$.

Now for any $y\in\mathcal{M}_1$ and $\delta>0$,

$$W_t^- - E_{y,z^*}[W^-(t+\delta)|\mathscr{F}_t] \geqq E_{y,z^*}\left[\int_t^{t+\delta} h_s^{y,z^*}\, ds|\mathscr{F}_t\right]$$

with equality if and only if y is optimal. Because $W_t^+ = W_t^-$,

$$W_t^- - E_{y,z^*}(W^-[t+\delta]|\mathscr{F}_t) = -E_{y,z^*}\left[\int_t^{t+\delta}[\wedge W_s^+ + \nabla W_s^+ \cdot f_s^{y,z^*}]\, ds|\mathscr{F}_t\right]$$

and so

$$E_{y,z^*}\left[\int_t^{t+\delta}(\wedge W_s^+ + \nabla W_s^+ \cdot f_s^{y,z^*} + h_s^{y,z^*})\, ds|\mathscr{F}_t\right] \leqq 0.$$

Taking the product of this expression with any $\theta\in L^\infty(\mathscr{C}, \mathscr{F}_t, P_0)$, dividing by δ and letting $\delta\to 0$ we can conclude as in [4, p. 246] that

$$\wedge W_s^+ + \nabla W_s^+ \cdot f_s^{y,z^*} + h_s^{y,z^*} \leqq 0, \qquad t\in[0, 1].$$

Combining this inequality with (4.1) above we see that if $W_t^+ = W_t^-$, then

$$\min_z\max_y H(t, x, \nabla W_t^+; y, z) = \max_y\min_z H(t, x, W_t^+; y, z)$$

for almost all (t, x). That is, the Isaacs condition holds at almost all "relevant" points.

5. Final remarks. By multiplying h by the characteristic function $I_{t \leqq \tau}$ of time up to some stopping time $\tau \leqq 1$, nonfixed time games are included in our treatment. If the Isaacs condition does not hold, relaxed controls can be introduced.

Acknowledgment. The author is indebted to Professor Varaiya for stimulating discussions.

REFERENCES

[1] V. E. BENEŠ, *Existence of optimal strategies based on specific information for a class of stochastic decision problems*, this Journal, 8 (1970), pp. 179–188.
[2] ———, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–475.
[3] M. H. A. DAVIS, *On the existence of optimal policies in stochastic control*, this Journal, 11 (1973), pp. 587–594.
[4] M. H. A. DAVIS AND P. P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.
[5] T. E. DUNCAN AND P. P. VARAIYA, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.
[6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Part I, Wiley—Interscience, New York, 1958.
[7] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theor. Probability Appl., 5 (1960), pp. 285–301.
[8] P. A. MEYER, *Probability and Potentials*, Blaisdell, Waltham, Mass., 1966.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

UNIFORM CONVERGENCE OF THE POTENTIAL FUNCTION ALGORITHM*

LLOYD FISHER† AND S. J. YAKOWITZ‡

Abstract. The identification problem of concern here is the estimation of a real function $f(x)$ by means of noisy observations $\{(X_i, f(X_i)+\eta_i(X_i))\}$ of its pairs, the X_i's being chosen independently according to some fixed law P. The approach taken for estimation is the "potential function" method (its sources are referenced herein), to wit: Choose f_0 arbitrarily and define the sequence $\{f_n\}$ by the recursive relation $f_{n+1}(x)=f_n(x)+\gamma_n(f(X_{n+1})+\eta(X_{n+1})-f_n(X_{n+1}))K(X_{n+1}, x)$, K being a positive symmetric kernel. From earlier publications it is known that under certain mild restrictions $E[\|f_n-f\|^2]\to 0$ in the $L_2(p)$-norm. Rates of convergence have been obtained in the restrictive case that $K(x, y)=\sum_{i=1}^{N}\lambda_i^2\phi_i(x)\phi_i(y)$ and $f(x)\in$ span $\{\phi_i, 1\leqq i\leqq N\}$. The contribution of this paper is to prove that while no uniform bounds exist in the $L_2(p)$-norm (we prove this) if $\{\phi_i\}$ is an infinite set, we do have $E[\|f-f_n\|_K^2]<C_n(\|f\|)$ for the norm $\|g\|_K^2=\iint g(x)g(y)K(x, y)p(x)p(y)\,dx\,dy$ and $\{C_n(r)\}$ a sequence converging to 0 for each positive r. A final result concerns the rate at which increasing finite-dimensional projections of f_n-f converge to 0 in the $L_2(p)$-norm. From our methods it is seen that if $f\notin V=$ span $(\{\phi_i\})$, then f_n converges in the mean to the projection of f on V.

1. Introduction. Let $\{X_i\}$ denote an independent sequence of observations of the probability experiment $(\mathscr{X}, \mathscr{A}, P)$ and f a real-valued function defined on $\mathscr{X}$. Sequentially the pairs $(X_n, f(X_n)+\eta_n(X_n))$ are made known, $\eta_n(X_n)$ being a random variable independent of $(X_1, X_2, \cdots, X_{n-1})$ and having a variance uniformly (in n and X_n) bounded by the positive number V. The problem confronted here is how to approximate f by f_n, f_n being determined by the pairs $\{(X_i, f(X_i)+\eta_i(X_i)), i\leqq n\}$ in such a way that $f_n\to f$ at some rate depending only on a norm of f.

The only approach to the above identification problem in this generality which the authors have found in their survey of the literature is the "potential function method" which is reviewed in Aizerman et al. [1]. The research results reported here concern the following version of the potential function method. Assume P in the probability experiment has density p with respect to some measure μ, and let $\{\phi_i\}$ denote some orthonormal sequence in $L_2(p)$ (the space of functions with inner product $(f, g)=\int f(x)g(x)p(x)\mu(dx)$). Let $K(x, y)=\sum\lambda_i^2\phi_i(x)\phi_i(y)$, where the λ_i's are chosen so as to assume that K (called the "potential function") is the kernel of a positive Hermitian operator on $L_2(p)$. With K so defined, the potential function method is to form a sequence $\{f_n\}$ of functions by the iterative formula

$$f_{n+1}(x)=f_n(x)+\gamma_n(f(X_{n+1})+\eta_{n+1}(X_{n+1})-f_n(X_{n+1}))K(X_{n+1}, x),$$

$\{\gamma_i\}$ being a sequence of positive numbers which sum to infinity, but such that $\sum\gamma_i^2<\infty$. f_0 is selected arbitrarily. From Aizerman et al. [2], it is known that if $f\in L_2(p)$ is in the span of the ϕ_i's, then $E[\|f_n-f\|^2]\to 0$ in P-probability. There are

* Received by the editors October 23, 1973, and in revised form December 19, 1974.

† Biostatistics Department, University of Washington, Seattle, Washington 98195. The work of this author was supported in part by the National Science Foundation under Grant GK 39751.

‡ Department of Systems and Industrial Engineering, University of Arizona, Tucson, Arizona 85721. The work of this author was supported in part by the National Science Foundation under Grant 35915.

no bounds supplied concerning the rate of this convergence, and, in fact, it is proven here that regardless of f_0,

$$\sup_{\|f\| \leq r} E[\|f_n - f\|^2] \geq r^2 \quad \text{for every } n. \tag{1.1}$$

Braverman and Pjatnichiĭ [4] have supplied a rate of convergence under the very restrictive assumption that the set $\{\phi_i\}$ be finite.

The main result of this study is to show that although (1.1) is true, if we define the slightly different norm $\|g\|_K$ to be

$$\|g\|_K^2 = \int\int K(x, y)g(x)g(y)p(x)p(y)\mu(dx)\mu(dy),$$

then for every positive number r, one may compute a sequence $\{c_n(r)\}$ converging to 0 such that

$$E[\|f_n - f\|_K^2] < C_n(r) \tag{1.2}$$

whenever $\|f\| < r$ and f_0 is taken to be the 0 function. Let us compare the two norms: If $g = \sum c_i\phi_i$, then $\|g\|^2 = \sum c_i^2$ and $\|g\|_K^2 = \sum \lambda_i^2 c_i^2$.

It appears that the potential function method may be useful, for example, in finding the shape of ore bodies, aquifers and, as described in [1], [3] and [6], performing supervised learning in pattern recognition problems.

Although we have not seen mention of it in the literature, the potential function method appears to be particularly suited to identifying a line or surface u which is known to be a solution of a differential equation

$$Lu = f, \tag{1.3}$$

where L is a given self-adjoint operator on $L_2(p)$, but the forcing function f in $L_2(p)$ is not known. (Such a situation arises in studying the aquifer in the Tucson basin, for example. Hydrologists believe they know the equation for the pressure head, but the aquifer recharge from rain and underground sources cannot be measured.) Under these circumstances, one may conclude (from (4.29) in [5], for example) that the Greens function $G(x, y)$ for L has the representation

$$G(x, y) = \sum_i \beta_i^{-1}\phi_i(x)\phi_i(y),$$

where the ϕ_i's are eigenfunctions which are orthogonal with respect to the $L_2(p)$ inner product and the β_i's are the associated eigenvalues. If the β_i's are positive, then G itself is a positive symmetric kernel and therefore suitable as a potential function. Otherwise, one may be assured that the function

$$K(x, y) = \int G(x, z)G(z, y)p(z)\mu(dz) = \sum_i \beta_i^{-2}\phi_i(x)\phi_i(y)$$

is positive symmetric and that u is in the span of the ϕ_i's. As explained and demonstrated in [5] and elsewhere, it is often relatively easy to find $G(x, y)$.

If the operator L in (1.4) is an integral operator with positive symmetric $L_2(p)$ kernel of the form

$$I(u) = \int u(x)K(x, y)p(x)\mu(dx)$$

whose spectrum is discrete, then K itself is a suitable potential function for the potential function algorithm. For we may represent K as $K(x, y) = \sum_i \beta_i \phi_i(x)\phi_i(y)$, where the ϕ_i's are orthonormal eigenfunctions for the integral operator and β_i's the associated eigenvalues.

While we have not found reports on numerical experiments using the potential function method, our own studies (which we are preparing to submit for publication) indicate that the method works well in comparison to alternative heuristic schemes. We report a typical experiment. Define $\mathcal{X}$ to be the finite set $\{0.00, 0.01, 0.02, \cdots, 0.99, 1.00\}$. We take the target function $f(x)$ to be sin (4x), $x \in \mathcal{X}$. The noise $\{\eta_j\}$ is a sequence of independent observations uniformly distributed on $[-1/2, 1/2]$. The X_i's are chosen uniformly on $\mathcal{X}$. The potential function is $K(x, y) = \exp(-10(x-y)^2)$. From Theorem 14 of [1], this is a potential function, and its orthonormal functions ϕ_i are complete in $L_2([0, 1])$. The weights $\{\gamma_n\}$ are determined by $\gamma_n = (20 + n^{1/2})^{-1}$. For purposes of comparison, we chose as a benchmark a heuristic nonparametric interpolation function which averages when it can and otherwise interpolates linearly. Specifically, this interpolation function f'_n is defined on $\mathcal{X}$ by the rule: (i) $f_n(x)' = \text{average } \{f(x_j) + \eta_j(x_j) : x_j = x, j \leqq n\}$. If the set in (i) is empty, (ii) $f_n(x)'$ is gotten by linearly interpolating between average values at x_i and x_j, where x_i and x_j are the nearest points on each side of x which have been sampled by the nth iteration. (iii) If one side of x hasn't been sampled, $f_n(x)'$ is simply the average value at the point closest to x which has been sampled. We have found that this "averaging function" works relatively well in the noisy observation case. In Table 1, we have also given the error associated with Lagrange and cubic spline interpolation. In the table, we have given the rms error, where

$$\text{rms error} \equiv \left[\sum_x (f(x) - f_n(x))^2\right]^{1/2} \div 100.$$

TABLE 1

A comparison of interpolation methods for noisy sine samples

Number of Samples	*100*	*200*	*300*	*400*
Potential function	0.217	0.157	0.116	—
Averaging function	0.259	0.231	0.209	—
Lagrange interpolation*	0.292	32×10^5	—	0.173
Cubic spline interpolation*	3418.000	0.284	—	0.164

* with averaging at multiply-sampled points

2. Principal results. Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a sigma-finite measure space. Let $p(x)$ be a probability density with respect to μ. Let $X_1, X_2, \cdots$, be a sequence of independent, p-distributed random elements of $\mathcal{X}$ and η_i (conditionally on X_i) a sequence of independent mean-zero random variables with variance bounded by a constant $V < \infty$. Let $\phi_1, \phi_2, \cdots, \psi_1, \psi_2, \cdots$, be an orthonormal basis for $L_2(p)$, that is, with the inner product $(g, h) = \int g(x)h(x)p(x)\mu(dx)$; $\|g\|^2 \equiv (g, g)$.

Let $K(x, y) = \sum_i \lambda_i^2 \phi_i(x)\phi_i(y)$, where $|K| \leqq R$ and $\sum_i \lambda_i^2 < \infty$. Let

$$f = h + \phi + \psi,$$

where

(i) $\int h^2(x)p(x)\mu(dx) = 0$,
(ii) $\phi + \psi$ is in $L_2(p)$,
$\phi = \sum c_i\phi_i$ and $\psi = \sum d_j\psi_j$.

Let $\gamma_i > 0$, $\sum_i \gamma_i = +\infty$, $\sum_i \gamma_i^2 < \infty$ and $Y_i = f(X_i) + \eta_i$.

The sequence (Y_1, X_1), (Y_2, X_2), $\cdots$, is observed, and f is estimated as follows. Let f_0 be a fixed element of $L_2(\{\phi_1, \phi_2, \cdots\})$. Recursively define f_n by

$$f_{n+1}(x) = f_n(x) - \gamma_n(f_n(X_{n+1}) - Y_{n+1})K(X_{n+1}, x). \tag{2.1}$$

For $f, g \in L_2(p)$, let

$$(f, g)_K = \int\int f(x)K(x, y)g(y)p(x)p(y)\mu(dx)\mu(dx)$$

and

$$\|f\|_K^2 = (f, f)_K.$$

THEOREM 1. *Under the above assumptions*:

(i) $E[(f_n - \phi, f_n - \phi)_K] \to_P 0$,

(ii) $E[(f_n - \phi, f_n - \phi)_K] \leqq C_n(\|f\|)$,

where $\lim_n C_n(\|f\|) = 0$ *and* $C_n(r)$ *may be chosen to be nondecreasing in* r *and nonincreasing in* n.

Proof. (i) follows from (ii) so that it is sufficient to prove (ii). The proof follows the lines of Lemma 1 of Aizerman, Braverman and Rozonoer [2] and is similar to many of the proofs in the area of stochastic approximation.

Let

$$\alpha_i = \|f_i - f\|^2 \quad \text{and} \quad \beta_i = \|f_i - f\|_K^2.$$

Let $F_n = \{(Y_1, X_1), \cdots, (Y_n, X_n)\}$. We first show that $E(\alpha_i)$ is bounded.

$$\begin{aligned} E[\alpha_{n+1}|F_n] &= E\left[\int (f_{n+1}(x) - f_n(x) + f_n(x) - f(x))^2 p(x)\mu(dx)|F_n\right] \\ &= \alpha_n + 2E_{X,\eta_{n+1}}\left[\int (f_{n+1}(x) - f_n(x))(f_n(x) - f(x))p(x)\mu(dx)|F_n\right] \\ &\quad + E_{X,\eta_{n+1}}\left[\int (f_{n+1}(x) - f_n(x))^2 p(x)\mu(dx)|F_n\right]. \end{aligned} \tag{2.2}$$

The middle term of (2.2) is equal to

$$\begin{aligned} 2\gamma_n \int\int\int &(h(X) + \phi(X) + \psi(X) + \eta_{n+1}(X) - f_n(X)) \sum \lambda_i^2 \phi_i(X)\phi_i(x) \\ &\cdot (f_n(x) - (h(x) + \phi(x) + \psi(x)))p(x)p(X)\, dF(\eta_n|X)\mu(dX)\mu(dx). \end{aligned}$$

Using

(a) For each X, $\int \eta\, dF(\eta|X) = 0$,
(b) $h(X) = 0$ a.e. $(p(x)\mu(dx))$,
(c) $\int \psi(x)\phi_i(x)p(x)\mu(dx) = 0$ for all i,

we see that the middle term reduces to

$$-2\gamma_n\beta_n = -2\gamma_n \int\int (\phi(X)-f_n(X))K(X,x)(\phi(x)-f_n(x))p(x)p(X)\mu(dx)\mu(dX)$$
$$= -2\gamma_n \sum_i \lambda_i^2(\Delta c_n^i)^2,$$

where $f_n = \sum_i c_i^n \phi_i$, $\phi = \sum_i c_i\phi_i$ and $\Delta c_i^n = c_i - c_i^n$ (the difference in the ith Fourier coefficients).

Here and hereafter, sometimes $\eta_j(X_j)$ is simply denoted by η_j.

We turn our attention to the third term of (2.2). This is equal to

$$\gamma_n^2 E_{X,\eta_{n+1}}\left(\int [(\phi(X)+\psi(X)+\eta_{n+1}-f_n(X))K(X,x)]^2 p(x)\mu(dx)\right)$$
$$\leqq \gamma_n^2 R^2 E_{X,\eta_{n+1}}[(\phi(X)-f_n(X))^2+\psi(n)^2+\eta_{n+1}^2+2(\phi(X)-f_n(X))\psi(X)$$
$$+2(\phi(X)-f_n(X))\eta_{n+1}+2\psi(X)\eta_{n+1}]$$
$$\leqq \gamma_n^2 R^2(\alpha_n+D+V),$$

where in the last inequality, we set $D \equiv \sum d_j^2$ and used the fact that $f_n \in L_2(\{\phi_1, \phi_2, \cdots\})$, and thus all cross-product terms have expected value zero.

Combining all this and letting $\tilde{\alpha}_n = E[\alpha_n]$, $\tilde{\beta}_n = E[\beta_n]$, we have

$$E(\alpha_{n+1}|F_n) \leqq \tilde{\alpha}_n - 2\gamma_n\tilde{\beta}_n + \gamma_n^2 R^2(\tilde{\alpha}_n + D + V),$$

and taking the expectation over F_n,

$$\begin{aligned} E(\alpha_{n+1}) &\leqq \tilde{\alpha}_n - 2\gamma_n\tilde{\beta}_n + \gamma_n^2 R^2(\tilde{\alpha}_n + D + V) \\ &\leqq \tilde{\alpha}_n(1+\gamma_n^2 R^2) + \gamma_n^2(D+V), \end{aligned} \tag{2.3}$$

since $\gamma_n > 0$ and $\beta_n \geqq 0$.

From (2.3), a recursive argument shows that for all n,

$$E(\alpha_{n+1}) \leqq \prod_{i=1}^{n} (1+\gamma_i^2(R^2+D+V)) \max(\alpha_0, 1), \tag{2.4}$$

where $\alpha_0 = \|f-f_0\|^2$. Clearly this holds for $n=0$. Then inductively,

$$\begin{aligned} E(\alpha_{n+1}) &\leqq (1+\gamma_n^2 R^2)\alpha_n + \gamma_n^2(D+V) \\ &\leqq (1+\gamma_n^2 R^2) \prod_{i=1}^{n-1} (1+\gamma_i^2(R^2+D+V)) \max(\alpha_0, 1) \\ &\quad + \gamma_n^2(D+V) \prod_{i=1}^{n-1} (1+\gamma_i^2(R^2+D+V)) \max(\alpha_0, 1) \\ &= \prod_{i=1}^{n} (1+\gamma_i^2(R^2+D+V)) \max(\alpha_0, 1). \end{aligned}$$

Since $\sum_i \gamma_i^2 < \infty$, the infinite product converges to a finite limit. Thus let

$$B = \prod_{i=1}^{\infty} (1+\gamma_i^2(R^2+D+V)) \max(\alpha_0, 1).$$

For all n,

$$E(\alpha_n) \leqq B < \infty. \tag{2.5}$$

We now turn to consideration of the β_n's.

$$E(\beta_{n+1}|F_n) = E_{X_{n+1},\eta_{n+1}}\Big[\int\int\int (f_{n+1}(x)-f(x))K(x, y) \cdot (f_{n+1}(y)-f(y))p(x)p(y)\mu(dx)\mu(dy)|F_n\Big]. \tag{2.6}$$

Replacing $f_{n+1}-f$ by $f_{n+1}-f_n+f_n-f$ and expanding (2.6), it is found that

$$E(\beta_{n+1}|F_n) = \beta_n + 2E_{X_{n+1},\eta_{n+1}}((f_{n+1}-f_n, f_n-f)_K|F_n) + E_{X_{n+1},\eta_{n+1}}(\|f_{n+1}-f_n\|_K^2|F_n). \tag{2.7}$$

It is now shown that the second term is nonpositive. It is equal to

$$-2\gamma_n \int\int\int\int (\phi(X)+\psi(X)+\eta_{n+1}-f_n(X))K(X, x)K(X, y)(\phi(y)+\psi(y)-f_n(y)) \cdot p(X)p(x)p(y)\, dF(\eta|X)\mu(dX)\mu(dx)\mu(dy).$$

The conditional η integration eliminates the η_{n+1} term, and the x integration changes the kernel to $\sum_i \lambda_i^4\phi_i(X)\phi_i(y)$. When the integration is completed, the term becomes

$$-2\gamma_n \sum_i \lambda_i^4(\Delta c_i^n)^2 \leqq 0.$$

The last term on the right-hand side of (2.7) is given by

$$\begin{aligned}
&\gamma_n^2 \int\int\int\int (\phi(X)+\psi(X)+\eta_{n+1}-f_n(X))K(X, x)K(x, y)K(y, X) \\
&\quad \cdot (\phi(X)+\psi(X)+\eta_{n+1}-f_n(X))p(X)p(x)p(y)\, dF(\eta_{n+1}|X)\mu(dX)\mu(dx)\mu(dy) \\
&= \gamma_n^2 \sum_i \lambda_i^6 \int\int (\phi(X)+\psi(X)+\eta_{n+1}-f_n(X))^2\phi_i^2(X)\, dF(\eta_{n+1}|X)p(X)\mu(dX) \\
&\leqq \gamma_n^2 R \sum_i \lambda_i^4 \int\int (\phi(X)+\psi(X)+\eta_{n+1}-f_n(X))^2\, dF(\eta_{n+1}|X)p(X)\mu(dX) \\
&= \gamma_n^2 R \sum_i \lambda_i^4 \Big\{\int (\phi(X)-f_n(X))^2 p(X)\mu(dX) + \int \psi^2(X)p(x)\mu(dX) \\
&\quad + \int\int \eta_{n+1}^2\, dF(\eta_{n+1}|X)p(X)\mu(dX)\Big\} \\
&\leqq \gamma_n^2 R\Big[\Big(\sum \lambda_i^4\Big)(B+\|f\|^2+V)\Big] = \gamma_n^2 Q.
\end{aligned}$$

Thus, $E(\beta_{n+1}|F_n) \leqq \beta_n - 2\gamma_n \sum_i \lambda_i^4(\Delta c_i^n)^2 + \gamma_n^2 Q$, and taking expectations,

$$E(\beta_{n+1}) \leqq E(\beta_n) - 2\gamma_n \sum_i \lambda_i^4 E(\Delta c_i^n)^2 + \gamma_n^2 Q. \tag{2.8}$$

Now it is shown that (2.8) implies condition (iii) of the theorem. Without loss of generality, assume $\lambda_1 \geqq \lambda_2 \geqq \lambda_3 \geqq \lambda_4 \geqq \cdots \searrow 0$.

$$E(\beta_{n+1}) = E\left(\int (f_n(x) - \phi(x))K(x, y)(f_n(y) - \phi(y))p(x)p(y)\mu(dx)\mu(dy) \right)$$

$$= \sum_i \lambda_i^2 E[(\Delta c_i^n)^2] \leqq \lambda_1 B = S$$

for all n.

To show (ii), it is enough to show that for each $\varepsilon > 0$, an $N(\varepsilon)$ can be found such that for $n \geqq N(\varepsilon)$, $E(\beta_n) < \varepsilon$, whenever $\|f\| \leqq H$ (a constant). Note that the bound (2.5) is uniform over $\{f : \|f\| \leqq H\}$. Choose N_1 such that $\sum_{i=N_1}^{\infty} \gamma_n^2 Q < \varepsilon/2$. Then as the middle term in (2.8) is nonpositive if $E(\beta_m) < \varepsilon/2$ and $m \geqq N_1$ (m fixed), (2.8) implies $E(\beta_n) < \varepsilon$ for all $n \geqq m$. Choose N_2 such that $\lambda_{N_2+1}^2 B < \varepsilon/4$. Choose N_3 such that $\frac{1}{4}\sum_{i=N_1}^{N_3} \gamma_i \lambda_{N_2}^2 \varepsilon > S + \sum_i \gamma_i^2 Q + 1$. By contradiction it will be shown that $E(\beta_n) < \varepsilon/2$ for some n with $N_1 \leqq n \leqq N_3$, implying that $E(\beta_n) < \varepsilon$ for all $n' > n \geqq N_3 = N(\varepsilon)$.

$$\sum_i \lambda_i^4 E((\Delta c_i^n)^2) \geqq \sum_{i=1}^{N_2} \lambda_i^4 E((\Delta c_i^n)^2)$$

$$\geqq (\lambda_{N_2})^2 \sum_{i=1}^{N^2} \lambda_i E((\Delta c_i^n)^2)$$

$$= (\lambda_{N_2})^2 (E(\beta_n) - \sum_{i=N_2+1}^{\infty} \lambda_i^2 E((\Delta c_i^n)^2)$$

$$\geqq (\lambda_{N_2})^2 (E(\beta_n) - \lambda_{N_2+1} B).$$

Thus (2.8) yields

$$E(\beta_{n+1}) \leqq E(\beta_n) - 2\gamma_n \max(\lambda_{N_2}^2 (E(\beta_n) - \lambda_{N_2+1}^2 B), 0) + \gamma_n^2 Q. \tag{2.9}$$

If $E(\beta_n) \geqq \varepsilon/2$, (2.9) gives

$$E(\beta_{n+1}) \leqq E(\beta_n) - \frac{2\gamma_n \lambda_{N_2} \varepsilon}{4} + \gamma_n^2 Q. \tag{2.9'}$$

If $\beta_n \geqq \varepsilon/2$ for $N_1 \leqq n \leqq N_3$, repeated application of (2.9) gives

$$E(\beta_{N_3+1}) \leqq E(\beta_{N_2}) - \tfrac{1}{2} \sum_{n=N_2}^{N_3} \gamma_n \lambda_{N_2}^2 \varepsilon + \sum_{n=N_2}^{N_3} \gamma_n^2 Q$$

$$\leqq S - \left(S + \sum_i \gamma_i^2 Q + 1 \right) + \sum_i \gamma_i^2 Q = -1.$$

But $\beta_{N_3+1} \geqq 0$, giving the desired contradiction.

COROLLARY 1. *Let $F \subseteq L_2(p)$ and let there exist a constant H such that $f \in F \Rightarrow \|f\| \leqq H$. Then if the orthogonal projection of f onto $\{\phi_1, \phi_2, \phi_3, \cdots\}$ with respect to $\|\cdot\|$ is denoted by Pf, then $\|Pf - f_n\|_K^2 \to 0$ uniformly for $f \in F$.*

On the other hand, it is easy to see that one cannot hope for uniform convergence in the $\|\cdot\|$-norm.

THEOREM 2. *Let* $\{\phi_1, \phi_2, \cdots\}$ *be an infinite set. Then* $F = \{f : \|f\| \leqq 1, f = \sum_{i=1}^{N(f)} c_i\phi_i\}$ *is such that* $E[\|f - f_n\|] \to 0$ *but* $\sup_{f\in F} E[\|f - f_n\|] \geqq 1$ *for each n.* (In other words, there is no uniform bound on the rate of convergence.)

Proof. Convergence is proved in [2].

It is easy to see that

$$c_i^{n+1} = c_i^n + \gamma_n\lambda_i^2[f(X_{n+1}) + \eta_{n+1} - f_n(X_{n+1})]\phi_i(X_n) \tag{2.10}$$

or

$$E(c_i^{n+1} - c_i^n) = \gamma_n\lambda_i^2 E_{X_{n+1}}((f(X_{n+1}) - f_n(X_{n+1}))\phi_i(X_n)).$$

Thus

$$\begin{aligned}|E(c_i^{n+1} - c_i^n)| &\leqq \gamma_n\lambda_i^2[E_{X_{n+1}}((f(X_{n+1}) - f_n(X_{n+1}))^2 E_{X_{n+1}}(\phi_i^2(X_i)))]^{1/2} \\ &= \gamma_n\lambda_i^2\alpha_n \leqq \gamma_n\lambda_i^2 B.\end{aligned}$$

Now

$$|E(c_i^{n+1} - c_i)| \geqq |E(c_i^0 - c_i)| - \sum_{j=1}^{n} |E(c_i^j - c_i^{j+1})|.$$

Thus

$$|E(c_i^{n+1} - c_i)| \geqq |E(c_i^0 - c_i)| - \lambda_i^2 \sum_{j=1}^{n} \gamma_j B.$$

Fix n and $\varepsilon > 0$. Since $f_0 \in L_2(p)$, choose i_0 such that $|c_{i_0}| < \varepsilon/2$ for $i \geqq i_0$. As $\sum \lambda_i^2 < \infty$, choose $i' > i_0$ such that $\lambda_{i'}^2 \sum_{j=1}^{n} \gamma_j B < \varepsilon/2$. Let $f = \phi_{i'}$. Then

$$\|f - f_n\| = \sqrt{\sum (c_i^n - c_i)^2} \geqq |c_{i'}^n - c_{i'}| \geqq 1 - \varepsilon/2 - \varepsilon/2 = 1 - \varepsilon.$$

The following presents two approaches to the problem of getting uniform convergence in the $\|\cdot\|$-norm. The first is to look at the convergence in the finite subspace $\{\phi_1, \cdots, \phi_m\}$ which will be uniform and then to let $m(n) \to \infty$ as the sample size n approaches ∞. The second approach is to require f_0 to be close enough to f so that all the coefficients converge at an appropriate rate.

THEOREM 3. *Under the above assumptions,*

(i) *Let* P_n *be the* $\|\cdot\|$ *projection onto* $\{\phi_1, \cdots, \phi_n\}$ *and* $\|\cdot\|_n$ *be the norm on* $\{\phi_1, \cdots, \phi_n\}$, *that is,* $\|P_n f\| = \|f\|_n$. *Let* $|\phi_i| \leqq \delta_i < \infty$ *for each i. Suppose* $F = \{f : \|f\| \leqq H\}$, *where* $H < \infty$ *is fixed. There exists a sequence* $m(n)$ *such that* $\lim_n \sup_{f\in F} E(\|f - f_n\|_{m(n)}^2) = 0$.

(ii) *Let* $|\phi_i| \leqq T < \infty$ *for all i. Given* $H > 0$, *let*

$$F = \left\{f : f = \sum c_i\phi_i + \psi_i + h, \sum_i \frac{(c_i^0 - c_i)^2}{\lambda_i^2} \leqq H\right\},$$

where $f_0 = \sum c_i^0\phi_i$. *Then*

$$\limsup_n \sup_{f\in F} E(\|f - f_n\|^2) = 0.$$

Proof. (i) Standard stochastic approximation techniques (for example, see Schmetterer [7]) or arguments similar to those of the proof of Theorem 1,

show that for each fixed m, $\|f-f_n\|_m \to 0$ uniformly over F. The selection of the sequence $m(n)$ then presents no problems.

(ii) Using the fact that $(c_i^0 - c_i)^2 \leqq H\lambda_i^2$, (2.10) and an argument analogous to that used in proving (2.5), one can show that

$$E((\Delta c_i^n)^2) \leqq \lambda_i^2 W$$

for all i (where W is constant depending on T, H, and V). As $E\|\phi - f_n\|^2 = E(\sum (\Delta c_i^n)^2)$, choose i_0 such that $\sum_{i \geqq i0} \lambda_i^2 W < \varepsilon/2$. Then

$$E(\|\phi - f_n\|^2) \leqq E(\|P_m\phi - f_n\|_{i_0}) + \sum_{i>i_0} \lambda_i^2 W \leqq E(\|\phi - f_n\|_{i_0}) + \varepsilon/2.$$

Uniform convergence with respect to $\|\cdot\|_{i_0}$ (as in (i)) allows the choice of n_0 such that for $n \geqq n_0$, $E(\|\phi - f_n\|_{i_0}) < \varepsilon/2$, so that for $n \geqq n_0$, $\sup_{f \in F} E(\|\phi - f_n\|^2) < \varepsilon$.

In summary, in this paper it is seen (under suitable regularity conditions) that:

1. f converges to the projection on the subspace spanned by the eigenfunctions of K.

2. The convergence is uniform in the $\|\cdot\|_K$-norm for bounded sets in the $\|\cdot\|$-norm.

3. To get uniform convergence in the $\|\cdot\|$-norm of a set F of functions, all the functions must be close to the starting function in the sequential approximation process.

REFERENCES

[1] M. A. AIZERMAN, E. M. BRAVERMAN AND L. I. ROZONOER, *Extrapolative problems in automatic control and the method of potential functions*, Amer. Math. Soc. Transl., 87 (1970), pp. 281–303. (Translation of International Congress of Mathematicians in Moscow, 1966, pp. 619–711.)

[2] ———, *Theoretical foundations of the potential function method in pattern recognition learning*, Automat. Remote Control, 25 (1964), pp. 1546–1556.

[3] B. G. BATCHELOR, *A comparison of the decision surfaces of nearest neighbor and potential function classifiers*, Information Sci., 5 (1973), pp. 171–178.

[4] E. M. BRAVERMAN AND E. S. PJATNICHIĬ, *Estimation of the rate of convergence of algorithms based on the potential function method*, Automat. Remote Control, 27 (1966), pp. 80–100.

[5] B. FRIEDMAN, *Principles and Techniques of Applied Mathematics*, John Wiley, New York, 1956.

[6] W. MEISEL, *Potential functions in mathematical pattern recognition*, IEEE Trans. Computers, C-18(1969), pp. 911–918.

[7] L. SCHMETTERER, *Multidimensional stochastic approximation*, Multivariate Analysis II, P. R. Krishnaiah, ed., Academic Press, New York, 1969, pp. 443–460.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

CONTROLS AND GOALS IN ECONOMIC EQUILIBRIUM*

RODRIGO A. RESTREPO†

Abstract. This paper generalizes some results of Gale and Debreu related to results of Bohnenblust and Karlin, commonly used to prove the existence of equilibrium in competitive economies. Applied to economics, these results allow further latitude in the choice of admissible price vectors, and in the conditions defining equilibrium. These results are also applicable in more general contexts where, instead of prices, one considers a set of controls or instruments of policy and where, instead of equilibrium, another goal is sought. These goals are related to the appropriate instruments through the concept of dual convex cones. Some examples are provided.

Introduction and results. In their study of economic equilibrium, Arrow and Debreu [1], Arrow and Hahn [2], Debreu [4], [6], Nikaido [10] and other authors, observing that each price vector p in some set P determines a set $Z(p)$ of excess demands for goods, have determined conditions for the existence of $\hat{p} \in P$ and $\hat{z} \in Z(\hat{p})$ with $\hat{z} \leqq 0$. Thus at price $\hat{p}$, demand can be satisfied.

The proofs of the existence of such $\hat{p}$ and $\hat{z}$ are usually based on properties of convex cones established by Bohnenblust and Karlin [3], Gale [8] and Debreu [5]. This paper provides a further generalization of these results using, as does Debreu, the concept of dual convex cones. Though the motivation comes from economic theory, the results are applicable to situations where instead of a price simplex P one considers more general sets C of controls or instruments of policy which, through an appropriate point-to-set mapping, determine a set of outcomes $Z(c)$ for each $c \in C$. The condition $z \leqq 0$ can then be replaced by other conditions in the manner indicated below.

In what follows, the inner product of two vectors z and c will be denoted by $z'c$, and the Cartesian product of two sets C_1 and C_2 will be denoted by $C_1 \times C_2$. With each nonempty convex set C will be associated a set C^*, called the dual convex cone, defined by

$$C^* = \{z \mid z'c \leqq 0, \text{ for all } c \in C\}.$$

Using these concepts, the following theorem will be established.

THEOREM. *Let $C = C_1 \times \cdots \times C_k$ and $Z = Z_1 \times \cdots \times Z_k$ where, for each i, C_i and Z_i are nonempty, convex, compact subset of the Euclidean space R^{n_i}. If to each $c \in C$ is associated a set $Z(c)$ such that*

(a) *for each $c \in C$, $Z(c)$ is a nonempty, closed convex subset of Z,*

(b) *if $c = (c_1, \cdots, c_k) \in C$ and $z = (z_1, \cdots, z_k) \in Z(c)$, then $z_i' c_i \leqq 0$ for $i = 1, \cdots, k$,*

(c) *the mapping $c \to Z(c)$ is upper-semicontinuous.*

then there exist $\hat{c} \in C$ and $\hat{z} = (\hat{z}_1, \cdots, \hat{z}_k) \in Z(\hat{c})$ such that, for each i, $\hat{z}_i \in C_i^*$.

The following examples motivate the theorem. In economic equilibrium theory, [1], [2], [5], C is the standard price simplex P, and then $C^* = P^* = \{z \mid z \leqq 0\}$. The more general case considered by Debreu [5] corresponds to $k = 1$

* Received by the editors March 7, 1974, and in revised form December 9, 1974.

† Department of Economics, Harvard University, Cambridge, Massachusetts. Now at Department of Mathematics, University of British Columbia, Vancouver 8, British Columbia, Canada.

in the preceding theorem. This case is applicable to markets with commodities possessing multiple exhaustive uses (e.g., electricity for lighting, for heating, for cooking and for communications) required to have the same prices. Then C is a proper subset of P, and C^* stipulates only that the total joint excess demand for these multiple uses be nonpositive, in agreement with intuition. The general case, with $k \geqq 1$ in the theorem, can be applied to models involving several epochs, with the a priori requirement that not all goods be free on each epoch. For such models, $C = P \times P \times \cdots \times P$. The arguments used in the proof of the theorem may be applied also to multinational markets. Then one obtains equilibria with sufficient quantities of goods to satisfy demand, and enough foreign exchange to finance required imports in each country.

Proof of the Theorem. Consider first the case where each set C_i is a convex polytope with vertices $v_{i1}, \cdots, v_{im_i}$. For each $c_i \in C_i$, $z_i \in Z_i$, let

$$h_i(c_i, z_i) = \frac{c_i + \sum_{j=1}^{m_i} \max\{0, z_i' v_{ij}\} v_{ij}}{1 + \sum_j \max\{0, z_i' v_{ij}\}}. \tag{1}$$

Clearly, $h_i(c_i, z_i)$ is a vector in C_i, and the mapping $(c_i, z_i) \to h_i(c_i, z_i)$ is continuous on $C_i \times Z_i$. This continuity, together with assumption (c), implies that the mapping $(c, z) \to W(c, z)$, defined by

$$W(c, z) = \{(\gamma, \zeta) \in C \times Z \mid \gamma_i = h_i(c_i, z_i), \text{ all } i; \zeta \in Z(c)\},$$

is a point-to-set, upper-semicontinuous map of $C \times Z$ into its subsets. Furthermore, each image set $W(c, z)$ is nonempty, closed and convex. Thus, the Kakutani fixed point theorem [9] is applicable, showing that there exists $(\hat{c}, \hat{z}) \in C \times Z$, such that $(\hat{c}, \hat{z}) \in W(\hat{c}, \hat{z})$. Then, in particular,

$$\hat{z} \in Z(\hat{c}), \tag{2}$$

and also, $h_i(\hat{c}_i, \hat{z}_i) = \hat{c}_i$ for each i; that is,

$$\hat{c}_i + \sum_j [\max(0, \hat{z}_i' v_{ij})] v_{ij} = \hat{c}_i \left[1 + \sum_j \max(0, \hat{z}_i v_{ij}) \right].$$

Simplifying and multiplying both sides of the preceding equation by $\hat{z}_i$, one obtains that

$$\sum_j [\max(0, \hat{z}_i' v_{ij})] \hat{z}_i' v_{ij} = \hat{z}_i' \hat{c}_i \sum_j \max(0, \hat{z}_i' v_{ij}). \tag{3}$$

In (3), the left-hand side is the sum of nonnegative terms, while the right-hand side is nonpositive by assumption (b), since $\hat{z} \in Z(\hat{c})$. Thus, both sides must be zero, and so must be each term on the left; that is,

$$[\max(0, \hat{z}_i' v_{ij})] \hat{z}_0' v_{ij} = 0, \quad \text{all } i, j.$$

This implies that $\hat{z}_i' v_{ij} \leqq 0$, all i, j, and therefore $\hat{z}_i' c_i \leqq 0$ for $c_i \in C_1$. That is, $\hat{z}_i \in C_i^*$, as desired, under the assumption that each C_i is a polytope.

To extend the result to general nonempty, convex, compact sets C_i, observe that for each such C_i there exists an increasing sequence $\{C_i^m | m = 1, 2, \cdots\}$ of convex polytopes such that $C_i^m \subset C_i$ for all m, i and rel int $C_i \subset \bigcup_m C_i^m$. Applying the result already established to $C^m = C_i^m \times \cdots \times C_i^m$ and Z with the mapping $c \to Z(c)$ restricted to $c \in C^m$, one obtains $\hat{c}^m$ and $\hat{z}^m$ such that

$$\hat{c}^m \in C^m, \quad \hat{z} \in Z(c^m), \quad z_i^m \in (C_i^m)^*, \quad \text{all } i. \tag{4}$$

By compactness, there exists some subsequence of $\{\hat{c}^m, \hat{z}^m\}$ converging to some $(\hat{c}, \hat{z}) \in C \times Z$, and $z \in Z(\hat{c})$ by (4) and the upper-semicontinuity of the map $c \to Z(c)$; and also by (4) and the construction of $\{C_i^m\}$, one must have $\hat{z}_i' c_i \leqq 0$ for all $c_i \in$ rel int C_i, and therefore, $\hat{z}_i' c_i \leqq 0$ for all $c_i \in C_i$, as desired.

The preceding proof incorporates a simplification suggested to the author by Professor R. T. Rockafellar.

REFERENCES

[1] K. J. ARROW AND G. DEBREU, *Existence of an equilibrium for a competitive economy*, Econometrica, 22 (1954), pp. 265–290.

[2] K. J. ARROW AND F. H. HAHN, *General Competitive Analysis*, Holden–Day, San Francisco, Calif., 1971.

[3] H. F. BOHNENBLUST AND S. KARLIN, *On a theorem of Ville*, Annals of Math. Studies, vol. 24, Princeton University Press, Princeton, N.J., 1950, pp. 155–161.

[4] G. DEBREU, *New concepts and techniques for equilibrium analysis*, Internat. Econom. Rev., 3 (1962), pp. 257–273.

[5] ———, *Market equilibrium*, Proc. Nat. Acad. Sci. U.S.A., 42 (1956), pp. 876–878.

[6] ———, *Theory of Value*, John Wiley, New York, 1959.

[7] S. EILENBERG AND D. MONTGOMERY, *Fixed point theorems for multivalued transformations*, Amer. J. Math., 69 (1946), pp. 214–222.

[8] D. GALE, *The law of supply and demand*, Math. Scand., 3 (1955), pp. 155–169.

[9] S. KAKUTANI, *A generalization of Brower's fixed point theorem*, Duke Math. J., 8 (1941), pp. 457–459.

[10] N. NIKAIDO, *Convex Structures and Economic Analysis*, Academic Press, New York, 1968.

[11] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

NORMAL SYMMETRIC DYNAMICAL SYSTEMS*

ROGER W. BROCKETT AND PAUL A. FUHRMANN†

Abstract. In this paper we establish a form of the state space isomorphism theorem for linear differentiable dynamical systems in a Hilbert space and make some application of these results. The methods used are based on spectral representations and suggest a close connection between the state space isomorphism theorem and certain classical representation theorems in analysis. We also give a class of counterexamples which illuminate the difficulties in extending the finite-dimensional theory thus justifying, in part, the stronger hypothesis used here.

1. Introduction. Recently there has been great progress in extending the main results of finite-dimensional linear system theory to the context of systems with infinite-dimensional state spaces (e.g., [1], [5], [7]–[10]). A great part of this work uses shift operators as models for the internal structure of systems. In this paper we try to give a detailed study for symmetric systems, that is, systems with self-adjoint or normal generators and identical input and output operators. We shall characterize the weighting patterns realizable by such systems, prove the spectral minimality theorem for this class of system as well as a version of the state space isomorphism theorem which generalizes to this context. Applications to stability questions are considered and finally, by means of a counterexample, we indicate how, what seems to be a slight relaxation of the assumptions in the state space isomorphism theorem is enough to make the conclusion false.

To fix terminology we review some of the standard definitions. We consider an $m \times n$ matrix-valued function γ defined on $[0, \infty)$ to which we refer as a weighting pattern. It characterizes the input/output relations by means of a convolution type integral

$$y(t) = \int_0^t \gamma(t-\tau)u(\tau)\, d\tau. \tag{1.1}$$

The Laplace transform Γ of γ is assumed to exist in some half-plane $\{\lambda | \operatorname{Re} \lambda > \omega_0\}$ and is called the transfer function of the system. A triple $\{A, B, C\}$ of operators with A being the infinitesimal generator of a strongly continuous semigroup $T(t)$ in some Hilbert space H and $B : \mathbb{C}^n \to H$ and $C : H \to \mathbb{C}^m$ is called a realization of the impulse response function γ if

$$\gamma(t) = CT(t)B \quad \text{for } t > 0 \tag{1.2}$$

or equivalently

$$\Gamma(z) = C(I-A)^{-1}B \quad \text{for } z \in \rho_0(A), \tag{1.3}$$

where $\rho_0(A)$ denotes the principal connected component of $\rho(A)$, the resolvent set of A, that is the connected component of $\rho(A)$ that includes the half-space

* Received by the editors August 27, 1974.

† Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts 02138. This work was sponsored by the U.S. Office of Naval Research under the Joint Services Electronics Program by Contract N00014-67-A-0298-0006.

$\{\lambda | \text{Re } \lambda > \omega_0\}$. In case A is bounded this is the connected component of A which includes the point ∞. Two realizations $\{A, B, C\}$ and $\{A_1, B_1, C_1\}$ in Hilbert spaces H and H_1 are isomorphic if there exists a bounded and boundedly invertible transformation R from H to H_1 for which Fig. 1 is commutative. Two realizations are unitarily equivalent if the operator R is actually a unitary operator from H to H_1. A realization $\{A, B, C\}$ in a Hilbert space is controllable if $\bigcap_{t\geq 0} \ker B^*T(t)^* = \{0\}$ and observable if $\bigcap_{t\geq 0} \ker CT(t) = \{0\}$. A realization which is both controllable and observable is called a canonical realization. A state space isomorphism theorem is a statement about the relation between different canonical realizations of the same transfer function.

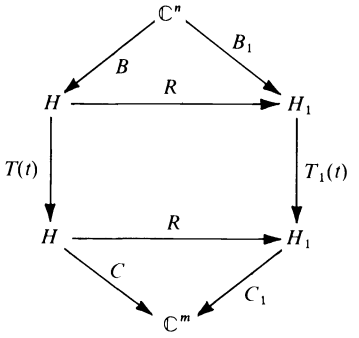


FIG. 1.

A system is self-adjoint if the infinitesimal generator is a (possibly unbounded) self-adjoint operator and $C = B^*$. In particular $\gamma(t)$ is a pointwise self-adjoint matrix-valued map. Similarly a system will be called normal symmetric if A is normal and $C = B^*$.

Whereas there is no inherent difficulty in working directly with unbounded operators, it is still technically simpler to reduce the problem to bounded operators. It is clear that given an infinitesimal generator A of a strongly continuous semigroup of normal operators, then for sufficiently large $\lambda_0 > 0$, $A - \lambda_0 I$ is the infinitesimal generator of a semigroup of normal contraction operators. Replacing A by $A - \lambda_0 I$ has the effect of multiplying the weighting pattern by $e^{-\lambda_0 t}$. So, without loss of generality, we may, as far as the state space isomorphism theorem is concerned, assume that the realizations are by contractive semigroups. To an infinitesimal generator of a strongly continuous contraction semigroup we associate a contraction T defined as the Cayley transform of A. Thus $T = (A + I)(A - I)^{-1}$. T will be called the cogenerator of the semigroup. T will be self-adjoint or normal if the semigroup is of self-adjoint or normal operators. For a treatment of cogenerators we refer to [6], [13]. Now T may be considered as the generator of a discrete system $\{T, B, C\}$ for which controllability and observability are defined by $\bigcap \ker B^*T^{*n} = \{0\}$, $\bigcap_{n\geq 0} \ker CT^n = \{0\}$, respectively. It turns out that the continuous time system $\{A, B, C\}$ is canonical if and only if the discrete time system $\{T, B, C\}$ is [9]. Since two continuous time systems $\{A, B, C\}$ and $\{A_1, B_1, C_1\}$ are unitary equivalent if and only if the discrete time

systems $\{T, B, C\}$ and $\{T_1, B_1, C_1\}$ are, it suffices to prove the state isomorphism in the later context.

For the case of normal systems the notions of controllability and observability have weaker counterparts which we call bilateral controllability and bilateral observability respectively. A discrete time normal system $\{A, B, C,\}$ is bilaterally controllable if $\bigcap_{n,m\geqq 0} \ker B^*A^{*n}A^{*m} = \{0\}$ and similarly bilateral observability is equivalent to $\bigcap_{n,m\geqq 0} \ker CA^nA^{*m} = \{0\}$. Clearly controllability implies bilateral controllability. For continuous time systems bilateral controllability is equivalent to $\bigcap_{t,\tau\geqq 0} \ker B^*e^{A^*t}e^{A^*\tau} = \{0\}$ and similarly for bilateral observability.

2. Spectral minimality. We want to study in this section the relation between the singularities of the transfer function and the spectrum of the generator in a realization of the transfer function. Let $\{A, B, C\}$ be a realization and Γ the transfer function of the system as defined in § 1. Γ is defined a priori only in some half-plane of the form $\{\lambda | \operatorname{Re} \lambda > \omega_0\}$. Being an analytic function Γ has an analytic continuation to $\rho_0(A)$, the continuation being given by (1.3). Let us denote by $\sigma(\Gamma)$ the set of nonanalyticity of the transfer function, continued analytically as above to $\rho_0(A)$. Obviously the relation

$$\sigma(\Gamma) \subset \sigma_0(A) \tag{2.1}$$

holds. We call this the spectral inclusion relation. A realization $\{A, B, C\}$ is spectrally minimal if there exists an analytic continuation of Γ for which $\sigma(\Gamma) = \sigma(A)$. If $\rho(A)$, the resolvent set of A is connected, then actually $\sigma_0(A) = \sigma(A)$ and there are no complications. However if $\sigma(A)$ is not connected, let $\rho_i(A)$ be a connected component of $\rho(A)$ which is not principal. It might turn out that Γ as defined in $\rho_0(A)$ has an analytic continuation to $\rho_i(A)$. On the other hand the function $\Gamma_i(z) = C(z-A)^{-1}B$ defined in $\rho_i(A)$ is certainly analytic. Unhappily Γ and Γ_i might be completely different. To avoid this kind of ambiguity we will restrict ourselves in this section to systems whose generators have connected resolvents. This assumption is of course redundant when dealing with self-adjoint infinitesimal generators. In this case there exists an analytic continuation of Γ for which $\sigma(\Gamma) = \sigma(A)$. We will say a realization $\{A, B, C\}$ of Γ is spectrally minimal if $\sigma(\Gamma) = \sigma(A)$.

THEOREM 2.1. *If $\{A, B, B^*\}$ is a canonical self-adjoint realization of a transfer function Γ, then the realization is spectrally minimal.*

Proof. Since A is self-adjoint then by the spectral theorem [3] there exists a spectral measure $E(\cdot)$ defined on the Borel sets of the real line and for which we have the following integral representation:

$$A = \int \lambda E(d\lambda).$$

Given an open interval (a, b) on the real line we have, limits taken in the strong operator topology [3, p. 920], that

$$E((a, b)) = \lim_{\delta\to 0} \lim_{\varepsilon\to 0} \frac{1}{2\pi i} \int_{a+\delta}^{b-\delta} [R[\lambda - i\varepsilon, A) - R(\lambda + i\varepsilon, A)]\, d\lambda. \tag{2.2}$$

Hence for every vector $\xi \in \mathbb{C}^n$ we have

$$\begin{aligned} \|E((a,b))B\xi\|^2 &= (B^*E((a,b))B\xi, \xi) \\ &= \lim_{\delta+0} \lim_{\varepsilon \to 0} \frac{1}{2\pi i} \int_{a+\delta}^{a-\delta} [(\Gamma(\lambda - i\varepsilon)\xi, \xi) - (\Gamma(\lambda + i\varepsilon)\xi, \xi)]\, d\lambda. \end{aligned} \tag{2.3}$$

Now let (a, b) be an open interval on the real line that is included in the domain of analyticity of Γ. The equality (2.3) implies $E((a, b))B\xi = 0$ for every $\xi \in \mathbb{C}^n$. Since the semigroup $T(t)$ generated by A commutes with the spectral measure $E(\cdot)$, we have $E((a, b))T(t)B\xi = T(t)E((a, b))B\xi = 0$. Now the set of vectors of the form $T(t)B\xi$, $t \geqq 0$, and $\xi \in \mathbb{C}^n$ spans the Hilbert space H by the assumption of controllability and hence it follows that $E((a, b)) = 0$. Thus $(a, b) \subset \rho(A)$ which in turn implies that $\sigma(A) \subset \sigma(\Gamma)$. Taking into account the spectral inclusion property the proof is complete.

Thus it follows that the spectra of two generators in two different canonical self-adjoint realizations of the same transfer function necessarily coincide.

As a by-product of (2.2) we have the following lemma which will be used in the sequel.

LEMMA 2.1. *Let $\{A, B, B^*\}$ and $\{A_1, B_1, B_1^*\}$ be two canonical self-adjoint realizations with transfer functions Γ and Γ_1 and let $E(\cdot)$ and $E_1(\cdot)$ be the spectral measures of A and A_1 respectively. The transfer functions of the two systems coincide if and only if for every Borel set on the real line we have*

$$B^*E(\sigma)B = B_1^*E_1(\sigma)B_1. \tag{2.4}$$

Proof. Assume (2.3) holds. Then

$$\begin{aligned} \Gamma(z) = B^*R(z; A)B = \int \frac{1}{z-\lambda} B^*E(d\lambda)B = \int \frac{1}{z-\lambda} B_1^*E_1(d\lambda)B_1 = B_1^*R(z; A_1) \\ = \Gamma_1(z). \end{aligned}$$

The converse follows from (2.2) for open intervals and hence, by standard measure theoretic technique, for all Borel sets.

Theorem 2.1 can be generalized to the case of normal symmetric systems. Let A be a bounded normal operator and let E be the spectral measure associated with it. For each vector x in H we let μ_x denote the positive measure defined by $\mu_x(\sigma) = (E(\sigma)x, x)$ for all Borel sets σ.

THEOREM 2.2. *Let $\{A, B, B^*\}$ be a canonical, finite input, normal symmetric system with $\rho(A)$ being connected, which realizes a transfer function Γ. Then the realization is spectrally minimal.*

Proof. Let σ be an open set in the domain of analyticity of Γ. Since for each $\xi \in \mathbb{C}^n$ we have

$$(\Gamma(z)\xi, \xi) = (B^*(z-A)^{-1}B\xi, \xi) = \int (z-\lambda)^{-1}(E(d\lambda)R\xi, R\xi) = \int (z-\lambda)^{-1}\, d\mu_{B\xi},$$

it follows that $(\Gamma(z)\xi, \xi)$ is the Cauchy transform of the measure $\mu_{B\xi}$. By Theorem 8.2 in [4], we have

$$\mu_{B\xi}(\sigma) = \|E(\sigma)B\xi\|^2 = 0$$

and hence also $E(\sigma)B\xi = 0$. Since the normal operator A commutes with its

associated spectral measure E we have $E(\sigma)A^m B\xi = 0$ for all $\xi \in \mathbb{C}^n$ and $m \geqq 0$. By the controllability assumption $E(\sigma) = 0$. Therefore, as in Theorem 2.1, we conclude that the realization is spectrally minimal.

We remark that the conclusion of Lemma 2.1 holds just as well for normal symmetric systems.

3. Spectral representations and controllability. Whereas abstract Hilbert spaces provide us with a very general setting, the solution of specific problems requires frequently more structure. The essence of spectral theory is to study an operator through some representation, i.e., its image under a unitary transformation in a different space. In particular, function spaces turn out to be most useful. For our particular problem the canonical or ordered spectral representation of normal operators is the natural candidate. This has been recognized by Fattorini [5] who used the ordered spectral representation to give necessary and sufficient conditions for the controllability of a system with a self-adjoint generator by means of a finite input controller. We review the main ideas concerning spectral representations following [2] and refer to [3] for a more complete account of spectral representations and multiplicity theory.

Let $\mu_1, \mu_2, \cdots, \mu_\infty$ be a collection of mutually singular nonnegative measures and K_p, $p \geqq 1$, be p-dimensional Hilbert spaces and let K_∞ be a separable Hilbert space. We consider the spaces $L^2(\mu_p; K_p)$ of K_p-valued measurable μ_p-square integrable functions on the complex plane. In $L^2(\mu_p; K_p)$ we consider the normal operator Λ_p defined by $(\Lambda_p f)(\lambda) = \lambda f(\lambda)$. The operator Λ_p is bounded if μ_p has compact support. Next we consider the direct sum $\oplus_{p=1}^{\infty} L^2(\mu_p; K_p)$ and the operator $\Lambda = \oplus_{p=1}^{\infty} \Lambda_p$. Every normal operator A in a separable Hilbert space is unitarily equivalent to such an operator Λ. The unitary map $U : H \to \oplus L^2(\mu_p; K_p)$ for which $UAU^{-1} = \Lambda$ is called the canonical spectral representation of A. In the canonical spectral representation the measures μ_p are unique up to measure equivalence. The support of μ_p, i.e., the complement of the largest open set where μ_p vanishes, is the set of multiplicity p. The normal operator A has finite multiplicity p_0 if $\mu_{p_0} \neq 0$ and $\mu_p = 0$ for $p > p_0$.

An equivalent spectral representation is the ordered spectral representation. Let us choose K_p so that $K_1 \subset K_2 \subset \cdots \subset K_\infty$, $K_\infty = \bigvee_{p=1}^{\infty} K_p$ and let $\mu = \sum \mu_p$. Let χ_p be the characteristic function of supp (μ_p), the support of μ_p. Then $\chi_{p_1}\chi_{p_2} = 0$ for $p_1 \neq p_2$ and $\mu_p = \chi_p \mu$. If $f_p \in L^2(\mu_p; K_p)$ and $\sum \|f_p\|^2 < \infty$ we let $f = \sum \chi_p f_p$. Thus $f \in L^2(\mu; K_\infty)$ and we have a unitary map of $\oplus L^2(\mu_p; K_p)$ onto a closed subspace of $L^2(\mu; K_\infty)$. If A has finite multiplicity m then we consider $L^2(\mu; K_m)$ as the space in which we have the ordered spectral representation.

Now a self-adjoint system is controllable if and only if it is observable and hence characterization of controllability is at the same time a characterization of canonical self-adjoint systems. This is no longer the case for normal symmetric systems. However if we make the assumption that the resolvent set of the normal operator A is connected and the spectrum of A has no interior then it follows, by an application of Mergelyan's theorem [4] that a normal symmetric system is controllable if and only if it is observable. So let us assume that the normal operator A is already given in its canonical spectral representation. Thus $H = \oplus L^2(\mu_p, K_p)$, $A = \Lambda$ and $B : \mathbb{C}^n \to H$. For convenience we identify K_p with $\mathbb{C}^p$ and

hence with respect to the standard orthonormal basis $\{e_i\}_{i=1}^m$ in $\mathbb{C}^m$ we have a convenient matrix representation for B. Thus $Be_i = \sum (Be_i)_p$ with $(Be_i)_p \in L^2(\mu_p; \mathbb{C}^p)$, and $(Be_i)_p = (\beta_{ji}^{(p)})_{j=1,\cdots,p}$. We will denote by $B^{(p)}$ the $p \times n$ matrix-valued function $(\beta_{ji}^{(p)})$.

Thus Fattorini's theorem [5] can be stated in the following form.

THEOREM 3.1. *Let A be a normal operator with a connected resolvent set and spectrum with no interior acting in a separable Hilbert space H. Let $B : \mathbb{C}^n \to H$ be a linear operator. Then the normal symmetric system $\{A, B, B^*\}$ is canonical if and only if the spectral multiplicity m of A is less than or equal to n and the conditions*

$$\operatorname{rank} (\beta_{ji}^{(p)}) = p, \qquad j = 1, \cdots, p, \quad i = 1, \cdots, n,$$

are satisfied μ_p – a.e. for $p = 1, \cdots, m$.

4. The state space isomorphism theorem. We proceed now to the proof of the main result of this paper, namely the state space isomorphism theorem for normal symmetric systems.

THEOREM 4.1. *Let $\{A, B, B^*\}$ and $\{A_1, B_1, B_1^*\}$ be two canonical normal realizations in Hilbert spaces H and H_1 respectively and assume the generators A and A_1 have connected resolvent sets. A necessary and sufficient condition that the two systems realize the same transfer function is that the systems are unitarily equivalent.*

Proof. The sufficiency part is trivial. To prove necessity we assume that the two systems realize the same transfer function. By Lemma 2.1, the equality (2.4) holds for all Borel sets in the complex plane. Since we are interested in unitary equivalence of the system we may, without loss of generality, assume that both systems are given in their canonical spectral representation. Let $x \in H$. Then we write $x = \sum_i x^{(i)}$ with $x^{(i)} \in L^2(\mu_i, K_i)$, the direct sum decomposition arising from the canonical spectral representation of A. For every Borel subset σ of the complex plane we have

$$(E(\sigma)x)^{(i)} = \chi_\sigma x^{(i)},$$

where χ_σ is the characteristic function of σ. Consider now the ordered representation and let $B(\lambda) = \sum \chi_p(\lambda) B^{(p)}(\lambda)$, where χ_p is the characteristic function of the support of μ_p. Also let $\mu = \sum \mu_p$ and we make similar definitions for the system $\{A_1, B_1, B_1^*\}$. Equality (2.3) implies that for each Borel set σ,

$$\int_\sigma B(\lambda)^* B(\lambda)\, d\mu = \int B_1(\lambda)^* B_1(\lambda)\, d\mu^{(1)}$$

holds. This in turn implies the scalar equality

$$\int_\sigma \operatorname{tr} B(\lambda)^* B(\lambda)\, d\mu = \int \operatorname{tr} B_1(A)^* B_1(\lambda)\, d\mu^{(1)}. \tag{4.1}$$

Since $\operatorname{rank} B(\lambda) \geqq 1$ a.e. with respect to μ and $\operatorname{rank} B_1(\lambda) \geqq 1$ a.e. with respect to $\mu^{(1)}$ and the supports of μ and $\mu^{(1)}$ are the same coinciding with the spectra of the generators, the above trace functions are positive. This implies that μ and $\mu^{(1)}$

are equivalent measures, i.e., each is absolutely continuous with respect to the other. Let

$$\psi^2 = \frac{d\mu}{d\mu^{(1)}} \tag{4.2}$$

be the Radon–Nykrodym derivative of μ with respect to $\mu^{(1)}$. Then

$$\psi(\lambda)^2 B(\lambda)^* B(\lambda) = B_1(\lambda)^* B_1(\lambda) \tag{4.3}$$

and in particular the equality

$$\text{rank } B(\lambda) = \text{rank } B_1(\lambda)$$

holds a.e. with respect to μ or $\mu^{(1)}$. Thus the multiplicity sets of the two normal operators A and A_1 are essentially the same. If we let $\psi_p = \chi_p \psi$, then we have also $\mu_p \simeq \mu_p^{(1)}$,

$$\psi_p^2 = \frac{d\mu_p}{d\mu_p^{(1)}} \tag{4.4}$$

and

$$\psi_p(\lambda)^2 B^{(p)}(\lambda)^* B^{(p)}(\lambda) = B_1^{(p)}(\lambda)^* B_1^{(p)}(\lambda). \tag{4.5}$$

Next we construct the unitary map U that intertwines the two systems. Let $\beta_i^{(p)}$ and $\beta_{j,i}^{(p)}$ be the columns of $B^{(p)}$ and $B_1^{(p)}$, respectively. By Theorem 3.1, $\{\beta_i^{(p)} | i = 1, \cdots, n\}$ and $\{\beta_{1,i}^{(p)} | i = 1, \cdots, n\}$ each space $\mathbb{C}^p$ a.e. with respect to μ and $\mu^{(1)}$ respectively. Define a map $U_p(\lambda) : \mathbb{C}^p \to \mathbb{C}^p$ by

$$U_p(\lambda)\beta_i^{(p)} = \beta_{1,i}^{(p)}. \tag{4.6}$$

From the rank conditions, rank $B^{(p)}(\lambda) = \text{rank } B_1^{(p)}(\lambda) = p$ a.e., it follows that $U_p(\lambda)$ is invertible. Moreover from (4.5) it follows that a.e.$(1/\psi_p(\lambda))U_p(\lambda)$ is unitary and hence, by Theorem 4.5.b in [2], U_p is a unitary map of $L^2(\mu_p; \mathbb{C}^p)$ onto $L^2(\mu^{(1)}, \mathbb{C}^p)$. Since it is a pointwise multiplication operator it clearly intertwines Λ_p and $\Lambda_p^{(1)}$, the multiplication by λ operators in $L^2(\mu_p; \mathbb{C}^p)$ and $L^2(\mu_p^{(1)}; \mathbb{C}^p)$, respectively. That is, $U_p \Lambda_p = \Lambda_p^{(1)} U_p$. Next define U by $U = \oplus U_p$. Then U is a unitary map of $\oplus L^2(\mu_p; \mathbb{C}^p)$ onto $+L^2(\mu_p^{(1)}; \mathbb{C}^p)$ that intertwines Λ and $\Lambda^{(1)}$ where $\Lambda = \oplus \Lambda_p$ and $\Lambda^{(1)} = \oplus \Lambda_p^{(1)}$. Clearly (4.6) is equivalent to $B = B_1$ and this completes the proof.

5. Realization by stable self-adjoint systems. We characterize in this section those weighting patterns realizable by means of finite input finite output stable self-adjoint systems.

Let $\{A, B, B^*\}$ be a self-adjoint system. Thus we assume that A is a self-adjoint infinitesimal generator of a, necessarily self-adjoint, strongly continuous semigroup $T(t)$ in a Hilbert space H. This implies that A is semibounded from above, i.e., there exists a real number ω such that for all x in the domain of A we have

$$(Ax, x) \leqq \omega \|x\|^2.$$

The implication of this inequality is that the spectrum of A is restricted to

$(-\infty, \omega]$. Let $E(\cdot)$ be the spectral measure of A. Then, by a simple application of the spectral theorem, we have for the weighting pattern γ of the system

$$\gamma(t) = B^*T(t)B = B^*\int_{-\infty}^{\omega} e^{\lambda t}E(d\lambda)B = \int_{-\infty}^{\omega} e^{\lambda t}B^*E(d\lambda)B.$$

If we make the additional assumption that A generates a contraction semigroup, i.e., that the system $\{A, B, B^*\}$ is stable, then the spectrum of A is restricted to the negative half-axis or equivalently $\omega = 0$. In this case $\gamma(t) = \int_{-\infty}^{0} e^{\lambda t}B^*E(d\lambda)B$ and hence for each ξ in $\mathbb{R}^n$,

$$(\gamma(t)\xi, \xi) = \int_{-\infty}^{0} e^{\lambda t}(E(d\lambda)B\xi, B\xi).$$

Since for each x in H the set function $(E(\cdot)x, x)$ is a finite nonnegative Borel measure on the real line it follows that

$$(-1)^n(\gamma^{(n)}(t)\xi, \xi) = \int_{-\infty}^{0} (-1)^n\lambda^n e^{\lambda t}(E(d\lambda)B\xi, B\xi) \geqq 0. \tag{5.1}$$

A scalar function ϕ defined on $[0, \infty)$ is called completely monotonic if ϕ is infinitely differentiable in $(0, \infty)$, continuous in $[0, \infty)$ and satisfies $(-1)^n\phi^{(n)}(t) \geqq 0$ for all $t > 0$ [14]. We extend this definition to Hilbert space operator-valued functions in a natural way. The differentiability assumption is replaced by weak differentiability. Thus a self-adjoint operator-valued function Φ is completely monotonic if for all x in H the function $\phi(t) = (\Phi(t)x, x)$ is completely monotonic. Since scalar completely monotonic functions have analytic extensions to the open right half-plane and since weak and uniform analyticity are equivalent [3] it follows that a completely monotonic function is actually differentiable in the uniform operator topology. Thus from (5.1) it follows that the weighting pattern of a stable self-adjoint system is a completely monotonic function. The converse is also true and we have the following theorem.

THEOREM 5.1. *An $n \times n$ matrix-valued function γ defined on $[0, \infty)$ is the weighting pattern of a stable self-adjoint system if and only if it is completely monotonic.*

Proof. In view of the remarks preceeding the theorem we have to prove only that a completely monotonic function is realizable by a stable self-adjoint system. The proof is based on a representation theorem of S. Bernstein [14] which characterizes a scalar completely monotonic function ϕ as an integral $\phi(t) = \int_{-\infty}^{0} e^{\lambda t}\, d\mu$ of a unique finite nonnegative Borel measure μ.

Now let $\gamma(t)$ be an $n \times n$ matrix-valued completely monotonic function. It follows from Bernstein's theorem that for each ξ in $\mathbb{C}^n$ there exists a finite nonnegative Borel measure μ_ξ such that

$$(\gamma(t)\xi, \xi) = \int_{-\infty}^{0} e^{\lambda t}\, d\mu_\xi.$$

Using polarization it follows that for each ξ and η in $\mathbb{C}^n$ there exists a finite complex Borel measure $\mu_{\xi,\eta}$ for which

$$(\gamma(t)\xi, \eta) = \int e^{\lambda t}\, d\mu_{\xi,\eta}. \tag{5.2}$$

The uniqueness part in Bernstein's theorem implies the uniqueness of the measure $\mu_{\xi,\eta}$ in the representation (5.2). By standard methods of spectral theory the uniqueness of the representing measure and (5.2) imply the existence of a matrix-valued measure $M(\cdot)$ defined on the Borel sets of the real line such that for each Borel set σ and all $\xi, \eta \in \mathbb{C}^n$ we have

$$\mu_{\xi,\eta}(\sigma) = (M(\sigma)\xi, \eta).$$

Since $\mu_{\xi,\xi}$ is a nonnegative measure it follows that $M(\cdot)$ is actually a nonnegative matrix-valued measure and

$$\gamma(t) = \int_{-\infty}^{0} e^{\lambda t} M(d\lambda).$$

For detailed accounts of matrix-valued measures we refer to [3].

To get the required realization we want to factor $M(\cdot)$ as

$$M(\cdot) = B^* E(\cdot) B, \tag{5.3}$$

where $E(\cdot)$ is some spectral measure in a Hilbert space H and B in a linear map from $\mathbb{C}^n$ to H.

To this end we construct the space $L^2(M)$ consisting of all $\mathbb{C}^n$-valued Borel measurable functions F defined on $(-\infty, 0]$ which satisfy

$$\|F\|^2 = \int_{-\infty}^{0} (M(d\lambda)F(\lambda), F(\lambda)) < \infty.$$

As usual we identify functions differing by null functions, i.e., functions whose norm vanishes. We introduce in $L^2(M)$ an inner product by means of the definition $(F, G) = \int (M(d\lambda)F(\lambda), G(\lambda))$.

With this inner product $L^2(M)$ becomes a Hilbert space [3, Chap. XIII]. In $L^2(M)$ we define the operator A by

$$(AF)(\lambda) = \lambda F(\lambda). \tag{5.4}$$

The domain of A is the set of all F in $L^2(M)$ for which the function $\lambda F(\lambda)$ is in $L^2(M)$. Clearly A is self-adjoint. Let $E(\cdot)$ be the spectral measure of A. For each Borel set σ we have

$$(E(\sigma)f)(\lambda) = \chi_\sigma(\lambda) f(\lambda),$$

where χ_σ is the characteristic function of the set σ. Next we define a map $B : \mathbb{C}^n \to L^2(M)$ by $(B\xi)(\lambda) = \xi$, i.e., a vector ξ in $\mathbb{C}^n$ is mapped into the constant function ξ. Now for every Borel set σ we have

$$(B^* E(\sigma) B\xi, \xi) = (E(\sigma)B\xi, B\xi) = \int_\sigma (M(d\lambda)\xi, \xi) = (M(\sigma)\xi, \xi).$$

This implies the factorization (5.3). Since the spectrum of A is supported on $(-\infty, 0]$, the generated semigroup $T(t)$ is contractive and has the representation

$$T(t) = \int_{-\infty}^{0} e^{\lambda t} E(d\lambda)$$

and hence

$$B^* T(t) B = B^* \int_{-\infty}^{0} e^{\lambda t} E(d\lambda) B = \int_{-\infty}^{0} e^{\lambda t} B^* E(d\lambda) B = \int_{-\infty}^{0} e^{\lambda t} M(d\lambda) = \gamma(t).$$

This completes the proof.

We wish to remark that the theorem holds true also for a self-adjoint system with infinite input and output, that is, for the case where $B : H_1 \to H$ is a bounded operator from a Hilbert space H_1 to H. This follows as a corollary to Naimark's theorem concerning unitary dilations of positive definite functions defined on groups [6], [13]. In fact given any set-valued function $M(\cdot)$ defined on the Borel subsets of the real line with values that are positive operators in H_1 satisfying $M(\sigma) \geqq I$, then there exists a larger Hilbert space $H \supset H$, and a spectral measure $E(\cdot)$ there for which

$$M(\sigma) = PE(\sigma)|H_1$$

for all Borel sets σ. Here P is the orthogonal projection of H onto H_1. Thus obviously $M(\sigma) = PE(\sigma)P$ and we have the factorization (5.3) as required.

The circle of ideas developed above can be used to yield some more system theoretic results. Mainly we will be concerned with skew adjoint systems (A, B, B^*), where $A = iA_0$, and A_0 is a, not necessarily bounded, self-adjoint operator in a Hilbert space H and B a bounded linear operator. The operator A is the infinitesimal generator of a group of unitary operators. The analytical tools in this case are the theorem of Bochner concerning the integral representation of positive definite functions on $\mathbb{R}$ [3] and the related Stone representation theorem for groups of unitary operators.

Without going into the details of the proof we state the following.

THEOREM 5.2. *An operator-valued weighting pattern $\gamma(t)$, $t \geqq 0$, is realizable by a skew adjoint system if and only if $\hat{\gamma}(t)$ defined on $\mathbb{R}$ by*

$$\hat{\gamma}(t) = \begin{cases} \gamma(t), & t \geqq 0, \\ \gamma(-t)^*, & t < 0, \end{cases} \tag{5.5}$$

is a positive definite function.

We recall that a Hilbert space-valued function $\hat{\gamma}(t)$ defined on $\mathbb{R}$ is positive definite if for all finite sets $t_1, \cdots, t_n \in \mathbb{R}$ and $\xi_1, \cdots, \xi_n \in H$ we have

$$\sum_{i,j=1}^{n} (\hat{\gamma}(t_i - t_j)\xi_i, \xi_j) \geqq 0.$$

A special class of functions which permits skew adjoint realizations is the class of completely monotonic functions. This follows, by way of Theorem 5.1, from the following lemma.

LEMMA 5.1. *Let $\gamma(t)$, $t \geqq 0$, be a completely monotonic operator-valued function in a Hilbert space H_1. Let $\hat{\gamma}$ be defined by* (5.5). *Then it is positive definite.*

Proof. By Theorem 5.1, we have for $t \geqq 0$ the factorization $\gamma(t) = B^* T(t) E$, where $T(t)$ is a contraction semigroup in a Hilbert space H. The semigroup $T(t)$ can be extended to a positive definite function on $\mathbb{R}$ by letting

$$\hat{T}(t) = \begin{cases} T(t), & t \geqq 0, \\ T(-t)^*, & t < 0. \end{cases} \tag{5.6}$$

The proof of this fact can be found in [13, p. 30], and it immediately implies the positive definiteness of $\hat{\gamma}$.

6. External and internal stability properties of self-adjoint systems. In the case of infinite-dimensional systems, knowledge about external stability properties of the system, even assuming controllability and observability, does not imply corresponding results about internal stability of a given realization. Moreover the lack of a general state space isomorphism theorem precludes us from dealing with all canonical realizations simultaneously. In fact we may have different canonical realizations of the same weighting pattern with one realization stable and another unstable [8]. However when we restrict ourselves to the class of self-adjoint systems those results become easily accessible.

Let $\Sigma: \{A, B, B^*\}$ be a canonical self-adjoint system in a Hilbert space H. We will say Σ is state stable (output stable) if for each $x \in H$ there exists an M_x such that $\|T(t)x\| \leqq M_x (\|B^* T(t)x\| \leqq M_x)$ for all $t \geqq 0$, Σ is asymptotically state stable (asymptotically output stable) if for each x, $\lim T(t)x \to 0$ ($\lim B^* T(t)x \to 0$) as $t \to \infty$, Σ is bounded input/bounded state stable (bounded input/bounded output stable) if there exists an $M > 0$ such that for $\|u(t)\| \leqq 1$ and all $\omega \geqq 0$ we have $\|\int_0^\omega T(t)Bu(\tau)\, d\tau\| \leqq M (\|\int_0^\omega B^* T(t)Bu(\tau)\, d\tau\| \leqq M)$. We will refer to these stability notations as s., a.s., a.s.s., a.o.s., b.i.b.s. and b.i.b.o. stability respectively. Obviously the following implications hold: s.s. $\Rightarrow$ o.s., a.s.s. $\Rightarrow$ a.o.s. and b.i.b.s. stability $\Rightarrow$ b.i.b.o. stability. We are interested in the converse implications.

THEOREM 6.1. *Let $\{A, B, B^*\}$ be a canonical, self-adjoint, finite input, finite output system. Then*

(i) *a.s. $\Rightarrow$ s.s.,*

(ii) *a.o.s. $\Rightarrow$ a.s.s.,*

(iii) *b.i.b.o. stability $\Rightarrow$ a.s.s.*

Proof. By the state space isomorphism theorem we may as well assume that the system is given in the spectral representation.

(i) To prove s.s. it suffices to show that the spectrum of A is restricted to the negative half-axis. Since the realization is o.s., by assumption we have for each x in H, $\|B^* T(t)x\| \leqq M_x$, and hence for each $\xi \in \mathbb{C}^n$ there exists an M_ξ such that $(B^* T(t)B\xi, \xi) \leqq M_\xi$ or $\|T(t/2)B\xi\|^2 \leqq M_\xi$. Since this expression remains bounded for all ξ we have $(B(\lambda)^* B(\lambda)\xi, \xi) = 0$ a.e. with respect to μ for $\lambda > 0$. Thus for $\lambda > 0$, rank $R(\lambda) = 0$ a.e. with respect to μ. Hence by Fattorini's result, $\mu((0, \infty)) = 0$.

(ii) Assume a.o.s., this implies o.s. and hence by (i) s.s.. We will show that $0 \in \sigma_p(A)$ is impossible. Assume $0 \in \sigma_p(A)$ and let $E_0 = E(\{0\})$ where $E(\cdot)$ is the

spectral measure of A. Thus $E_0 \neq 0$. By controllability there exists a $\xi \in \mathbb{C}^n$. such that $E_0 B\xi \neq 0$. By a.o.s., $B^*T(t)E_0B\xi \to 0$ and hence also $(B^*T(t)E_0B\xi, \xi) \to 0$. But $T(t)E_0B\xi = E_0B\xi$ and hence $(B^*T(t)E_0B\xi, \xi) = \|E_0B\xi\|^2 > 0$, a contradiction. Hence $0 \notin \sigma_p(A)$ and $E_0 = 0$. Let x be in H. To prove $T(t)x \to 0$ it suffices to show by self-adjointness that $(T(t)x, x) \to 0$. Now $(T(t)x, x = \int_{-\infty}^{0} e^{\lambda t}(E(d\lambda)x, x)$. Since $E_0 = 0$ the measure $(E(\cdot)x, x)$ of $\{0\}$ is zero. The result follows now by Lebesgue's dominated convergence theorem.

(iii) Assume the system is b.i.b.o. stable. This implies $B^*T(t)B$ is an $n \times n$ matrix with $L^1(0, \infty)$ entries. Since $B^*T(t)B$ is continuous we have $(B^*T(t)B\xi, \xi) \to 0$ for all $\xi \in \mathbb{C}^n$. By previous arguments this implies a.s.s.

7. A counterexample. Consider the scalar input/scalar output system defined by

$$\begin{bmatrix} \dot{x}_n(t) \\ \dot{x}_{n+1}(t) \end{bmatrix} = \frac{1}{n}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} x_n(t) \\ x_{n+1}(t) \end{bmatrix} + \frac{1}{n}\begin{bmatrix} 0 \\ 1 \end{bmatrix}u(t), \qquad n = 1, 3, 5, \cdots,$$

$$y(t) = \sum_{k=1}^{\infty} \frac{1}{(2k-1)} x_{2k}(t).$$

This system is clearly controllable and observable and realizes the weighting pattern

$$w(t-\sigma) = \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} \cos \frac{(t-\sigma)}{2k-1}.$$

This system is of the form (A, b, b) with A skew-adjoint. Let α_n, $n = 1, 3, 5, \cdots$, be a real sequence and consider also the system

$$\begin{bmatrix} \dot{z}_n(t) \\ \dot{z}_{n+1}(t) \end{bmatrix} = \frac{1}{n}\begin{bmatrix} \sinh 2\alpha_n & 1 + 2\sinh^2 \alpha_n \\ -1 - 2\sinh^2 \alpha_n & -\sinh 2\alpha_n \end{bmatrix}\begin{bmatrix} z_n(t) \\ z_{n+1}(t) \end{bmatrix} + \frac{1}{n}\begin{bmatrix} -\sinh \alpha_n \\ \cosh \alpha_n \end{bmatrix}u(t), \qquad n = 1, 3, 5, \cdots,$$

$$y(t) = \sum_{k=1}^{\infty} \frac{1}{(2k-1)}[+(\sinh \alpha_n) z_{2k}(t) + (\cosh \alpha_n) z_{2k+1}(t)].$$

This system is controllable and observable, realizes the same weighting pattern as the previous system, but is of the form $(A, b, \Sigma b)$ where Σ is self-adjoint, idempotent, and $\Sigma A \Sigma = A^*$. In this case the transformation relating x and z is

$$\begin{bmatrix} \cosh \alpha_n & -\sinh \alpha_n \\ -\sinh \alpha_n & \cosh \alpha_n \end{bmatrix}\begin{bmatrix} x_n \\ x_{n+1} \end{bmatrix} = \begin{bmatrix} z_n \\ z_{n+1} \end{bmatrix}, \qquad n = 1, 3, 5, \cdots.$$

This is a bounded map if and only if the sequence α_n is bounded. Thus we see that the given weighting pattern admits a realization of the form $(A, b, \Sigma b)$ with $\Sigma A \Sigma = A^*$ which is not similar to the normal symmetric realization displayed above. Moreover, we see that any two sequences $\{\alpha_n\}_{n=1}^{\infty}$ and $\{\beta_n\}_{n=1}^{\infty}$ such that $\{(\alpha_n - \beta_n)\}$ is not bounded generate realizations of the form $(A, b, \Sigma b)$; $\Sigma A \Sigma = A^*$ which are not similar.

REFERENCES

[1] J. S. BARAS AND R. W. BROCKETT, *H^2-functions and infinite-dimensional realization theory*, this Journal, 1(1975), pp. 221–241.

[2] R. BEALS, *Topics in Operator Theory*, Univ. of Chicago Press, Chicago, Ill., 1972.

[3] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, vol. 1, 2, Interscience, New York, 1957, 1963.

[4] T. W. GAMELIN, *Uniform Algebras*, Prentice-Hall, Englewood Cliffs, N.J., 1969.

[5] H. O. FATTORINI, *On complete controllability of linear systems*, Differential Equations, (1967), pp. 391–402.

[6] P. FILLMORE, *Notes on Operator Theory*, Van Nostrand, New York, 1970.

[7] P. A. FUHRMANN, *On weak and strong reachability and controllability of infinite dimensional linear systems*, J. Optimization Theory Appl., 9 (1972).

[8] ——— *On realization of linear systems and applications to some questions of stability*, Math. Systems Theory, 8 (1974), pp. 132–141.

[9] ———, *Exact controllability and observability and realization theory in Hilbert space*, J. Math. Anal. Appl., to appear.

[10] J. W. HELTON, *Discrete time systems, operator models and scattering theory*, J. Funct. Anal., 16 (1974), pp. 15–38.

[11] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, N.J., 1962.

[12] M. G. KREIN, *Introduction to the geometry of indefinite J-spaces and to the theory of operators in those spaces*, Amer. Math. Soc. Transl., 93 (1970), pp. 103–176.

[13] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

[14] D. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, N.J., 1946.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

LINEAR QUADRATIC DIFFERENTIAL GAMES IN A HILBERT SPACE*

AKIRA ICHIKAWA†

Abstract. We consider linear quadratic games in a Hilbert space. The system equation is linear and involves an unbounded operator which generates a strongly continuous evolution operator (or semigroup). We show that the existence of a solution to a Riccati integral equation implies the existence of a saddle point for the closed-loop game, and that the former is guaranteed if there exists a unique open-loop saddle point. We also consider quadratic games on an infinite interval.

1. Introduction. Let H_i, $i = 1, 2, 3$, be real Hilbert spaces. Consider a linear differential system

$$\dot{x} = A(t)x + B(t)u + C(t)v, \tag{1.1}$$

$$x(t_0) = x_0 \in H_1, \tag{1.2}$$

and a payoff functional

$$J(u, v) = (Fx(t_1), x(t_1)) + \int_{t_0}^{t_1} [(Wx, x) + (Uu, u) + (Vv, v)]\, dt. \tag{1.3}$$

The inner product of the space H_i will be denoted by $(\cdot, \cdot)$ and the norm by $|\cdot|$, while $x(t)$ represents the state of the system in H_1, and u, v are control functions with values in H_2, H_3, respectively. $A(t)$ is a closed linear unbounded operator whose domain $D(A(t))$ is dense in H_1. We assume that $A(t)$ generates a strongly continuous evolution operator (or two-parameter semigroup) $S(t, s)$, $t \geqq s \geqq 0$, on H_1. The operators $B(t)$: $H_2 \to H_1$, $C(t)$: $H_3 \to H_1$ are linear and uniformly bounded on $[t_0, t_1]$. The operators F, $W(t)$ on H_1 are self-adjoint and nonnegative definite. $U(t)$, $U^{-1}(t)$ on H_2 are self-adjoint and positive definite, while $V(t)$, $V^{-1}(t)$ on H_3 are self-adjoint and negative definite. The controller u is the minimizer of $J(u, v)$, and the controller v is the maximizer.

We define a solution of (1.1), (1.2) corresponding to locally Bochner integrable functions $u(t)$, $v(t)$ by

$$x(t) = S(t, t_0)x_0 + \int_{t_0}^{t} S(t, \tau)[B(\tau)u(\tau) + C(\tau)v(\tau)]\, d\tau. \tag{1.4}$$

Here the integral is in the sense of Bochner (see [5]). We also define a solution corresponding to closed-loop controls $u = \phi(t, x)$, $v = \psi(t, x)$ by the solution of the integral equation

$$x(t) = S(t, t_0)x_0 + \int_{t_0}^{t} S(t, \tau)[B(\tau)\phi(\tau, x(\tau)) + C(\tau)\psi(\tau, x(\tau))]\, d\tau. \tag{1.5}$$

Let $I = [t_0, t_1]$ be a fixed interval, and let $L_2(I; H_i)$ denote the space of strongly

* Received by the editors May 17, 1974, and in revised form December 12, 1974.

† Department of Mathematics, University of British Columbia, Vancouver 8, British Columbia, Canada. This work is based on the author's doctoral dissertation in Applied Mathematics at the State University of New York at Stony Brook. It was supported in part by the National Research Council of Canada under Grant A8051.

measurable functions $y(t) \in H_i$ such that

$$\int_I |y(t)|^2 \, dt < \infty.$$

Then $L_2(I; H_i)$ is a real Hilbert space with an inner product $\langle \cdot, \cdot \rangle$ defined by

$$\langle y, z \rangle = \int_I (y(t), z(t)) \, dt.$$

The norm in $L_2(I; H_i)$ will be denoted by $\| \cdot \|$. We take admissible controls for the open-loop game to be $L_2(I; H_i)$-functions, and for the closed-loop game to be $\{\phi(t, x), \psi(t, x)\}$, such that (1.5) has a unique solution. Our problem is to seek a saddle point (or an optimal pair) $\bar{u}$, $\bar{v}$ which satisfies

$$J(\bar{u}, v) \leqq J(\bar{u}, \bar{v}) \leqq J(u, \bar{v}) \tag{1.6}$$

for any admissible controls u, v. The number $J(\bar{u}, \bar{v})$ is called the value of the game, if it exists.

The optimal control problems with quadratic cost in a Hilbert space were studied by several authors [4], [6], [7]. R. Temam [9] and Curtain and Pritchard [3] considered Riccati equations in an infinite-dimensional space, which are related to optimal control and filtering. A. Benssousan [1] studied differential games in a Hilbert space. His system is very general, and the results are similar to the present paper. But our approach is different from his and based on [4], [7], [3]. Our results are more general than [1], in the sense that the operator $K(t)$ (given in (2.1)) characterizes the saddle point and the value of the game, and that we do not have to solve a decoupled system of equations. The results on quadratic games in a finite-dimensional space are given, for example, in [8]. N-person quadratic games are discussed in a recent paper [2].

2. Quadratic games with closed-loop control. We consider the system (1.1), (1.2) and the payoff functional (1.3). Admissible controls are closed-loop control laws $\phi(t, x)$, $\psi(t, x)$ which give a unique solution to (1.5).

THEOREM 2.1. *Suppose that there exists a unique strongly continuous linear self-adjoint operator $K(t) \geqq 0$, $t \in I$, satisfying*

$$K(t) = S_K^*(t_1, t) F S_K(t_1, t) + \int_t^{t_1} S_K^*(\tau, t)[W(\tau) + K(\tau)D(\tau)K(\tau)]S_K(\tau, t) \, d\tau, \tag{2.1}$$

or, equivalently,

$$K(t) = S^*(t_1, t) F S(t_1, t) + \int_t^{t_1} S^*(\tau, t)[W(\tau) - K(\tau)D(\tau)K(\tau)]S(\tau, t) \, d\tau. \tag{2.1$'$}$$

Here $D(t) = B(t)U^{-1}(t)B^*(t) + C(t)V^{-1}(t)C^*(t)$, and $S_K(s, t)$ is an evolution operator generated by $A(s) - D(s)K(s)$. Then there exists a unique optimal pair given by

$$\begin{aligned} \bar{u}(t) &= -U^{-1}(t)B^*(t)K(t)x, \\ \bar{v}(t) &= -V^{-1}(t)C^*(t)K(t)x. \end{aligned} \tag{2.2}$$

Moreover, the value of the game is given by $(K(t_0)x_0, x_0)$, and the optimal trajectory

$\bar{x}(t)$ is expressed by

$$\bar{x}(t) = S_K(t, t_0)x_0.$$

Remark 2.1. Known results on evolution operators are summarized in [3].

Remark 2.2. The nonnegativity of $K(t)$ is necessary since

$$0 \leqq J(\bar{u}, 0) \leqq J(\bar{u}, \bar{v}),$$

and since the existence of $K(t)$ on some interval implies the existence of an optimal pair on the same interval.

The proof of the theorem involves several steps. First, consider a linear control problem

$$\begin{aligned} \dot{x} &= \tilde{A}(t)x + C(t)v, \\ x(t_0) &= x_0, \end{aligned} \tag{2.3}$$

with

$$J(v) = (Fx(t_1), x(t_1)) + \int_{t_0}^{t_1} [(\tilde{W}(t)x(t), x(t)) + (V(t)v(t), v(t))]\, dt,$$

where $F \geqq 0$, $\tilde{W}(t) \geqq 0$ and $V(t) < 0$. Here v tries to maximize $J(v)$. Let $T(t, s)$ be the evolution operator generated by $\tilde{A}(t)$, and let $L(t)$ be a strongly continuous linear operator.

LEMMA 2.1. *Let $Q(t)$ be a self-adjoint operator defined by*

$$Q(t) = T_L^*(t_1, t)FT_L(t_1, t) + \int_t^{t_1} T_L^*(\tau, t)[\tilde{W}(\tau) + L^*(\tau)V(\tau)L(\tau)]T_L(\tau, t)\, d\tau, \tag{2.4}$$

where $T_L(s, t)$ is generated by $\tilde{A}(s) + C(s)L(s)$. Then $(Q(t_0)x_0, x_0)$ gives the cost $J(v)$ corresponding to the control $v = L(t)x$.

Proof. The unique solution of (2.3) corresponding to $v = L(t)x$ is given by

$$x(t) = T_L(t, t_0)x_0. \tag{2.5}$$

Consider the following:

$$\begin{aligned} (Q(t)x(t), x(t)) &= (FT_L(t_1, t)x(t), T_L(t_1, t)x(t)) \\ &\quad + \int_t^{t_1} ([\tilde{W}(\tau) + L^*(\tau)V(\tau)L(\tau)]T_L(\tau, t)x(t), T_L(\tau, t)x(t))\, d\tau. \end{aligned}$$

Using the identity

$$T_L(s, t)x(t) = x(s),$$

we obtain

$$\begin{aligned} (Q(t)x(t), x(t)) &= (Fx(t_1), x(t_1)) \\ &\quad + \int_t^{t_1} [(\tilde{W}(\tau)x(\tau), x(\tau)) + (V(\tau)L(\tau)x(\tau), L(\tau)x(\tau))]\, d\tau \\ &= (Fx(t_1), x(t_1)) + \int_t^{t_1} [(\tilde{W}(\tau)x(\tau), x(\tau)) + (V(\tau)v(\tau), v(\tau))]\, d\tau. \end{aligned}$$

Setting $t = t_0$, we arrive at our result.

LEMMA 2.2. *Suppose that there exists a strongly continuous linear self-adjoint operator $\tilde{K}(t)$ satisfying the following*:

$$\tilde{K}(t) = T_L^*(t_1, t)FT_L(t_1, t) + \int_t^{t_1} T_L^*(\tau, t)[\tilde{W}(\tau) - \tilde{K}(\tau)D_2(\tau)\tilde{K}(\tau) - L^*(\tau)C^*(\tau)\tilde{K}(\tau) - \tilde{K}(\tau)C(\tau)L(\tau)]T_L(\tau, t)\,d\tau, \tag{2.6}$$

where $D_2(t) = C(t)V^{-1}(t)C^*(t)$. Let $Q(t)$ be defined by (2.4). *Then*

$$\tilde{K}(t) \geqq Q(t) \quad \text{for any } t \in [t_0, t_1].$$

Proof.

$$\tilde{K}(t) - Q(t) = -\int_t^{t_1} T_L^*(\tau, t)[\tilde{K}(\tau)D_2(\tau)\tilde{K}(\tau) + L^*(\tau)C^*(\tau)\tilde{K}(\tau) + \tilde{K}(\tau)C(\tau)L(\tau) + L^*(\tau)VL(\tau)]T_L(\tau, t)\,d\tau$$

$$= -\int_t^{t_1} T_L^*(\tau, t)[L(\tau) + V^{-1}(\tau)C^*(\tau)\tilde{K}(\tau)]^* V(\tau) \times [L(\tau) + V^{-1}(\tau)C^*(\tau)\tilde{K}(\tau)]T_L(\tau, t)\,d\tau$$

$$\geqq 0, \quad \text{since } V(t) < 0.$$

The integral equation (2.1) corresponds formally to the differential equation

$$\begin{aligned} \dot{K}(t) &= -[A(t) - D(t)K(t)]^*K(t) - K(t)[A(t) - D(t)K(t)] - W(t) \\ &\quad - K(t)D(t)K(t), \\ K(t_1) &= F. \end{aligned} \tag{2.7}$$

We can rearrange this into

$$\begin{aligned} \dot{K}(t) &= -A^*(t)K(t) - K(t)A(t) - W(t) + K(t)D(t)K(t), \\ K(t_1) &= F, \end{aligned} \tag{2.7$'$}$$

and

$$\begin{aligned} \dot{K}(t) &= -[A(t) + P(t)]^*K(t) - K(t)[A(t) + P(t)] \\ &\quad - W(t) + K(t)D(t)K(t) + P^*(t)K(t) + K(t)P(t), \\ K(t_1) &= F. \end{aligned} \tag{2.7$''$}$$

Here $P(t)$ is a strongly continuous linear operator. The integral equation (2.1)$'$ corresponds formally to (2.7)$'$. The integral equation corresponding to (2.7)$''$ is given as follows:

$$K(t) = S_P^*(t_1, t)FS_P(t_1, t) + \int_t^{t_1} S_P^*(\tau, t)[W(\tau) - K(\tau)D(\tau)K(\tau) - P^*(\tau)K(\tau) - K(\tau)P(\tau)]S_P(\tau, t)\,d\tau, \tag{2.1$''$}$$

where $S_P(s, t)$ is generated by $A(s) + P(s)$. The equivalence of (2.7)$'$, (2.7)$'$ and (2.7)$''$ is trivial, their meaning aside. We shall show the equivalence of (2.1), (2.1)$'$ and (2.1)$''$.

LEMMA 2.3. *Let $M(t)$ be a strongly continuous linear self-adjoint operator. Define for each $t \in [t_0, t_1]$,*

$$N(t) = \int_t^{t_1} \hat{T}^*(\tau, t) M(\tau) \hat{T}(\tau, t)\, d\tau$$

and

$$\tilde{N}(t) = \int_t^{t_1} \hat{T}_P^*(\tau, t) M(\tau) \hat{T}_P(\tau, t)\, d\tau.$$

Here $\hat{T}(\tau, t)$ is generated by $\hat{A}(t)$, and $\hat{T}_P(\tau, t)$ by $\hat{A}(t) + P(t)$. Then

$$N(t) = \tilde{N}(t) - \int_t^{t_1} \hat{T}_P^*(\tau, t)[P^*(\tau)N(\tau) + N(\tau)P(\tau)]\hat{T}_P(\tau, t)\, d\tau. \tag{2.8}$$

Proof. Let

$$\hat{N}(t) = \int_t^{t_1} \hat{T}_P^*(\tau, t) M(\tau) \hat{T}(\tau, t)\, d\tau.$$

First, we shall show

$$\hat{N}(t) = N(t) + \int_t^{t_1} \hat{T}_P^*(\tau, t) P^*(\tau) N(\tau) \hat{T}(\tau, t)\, d\tau. \tag{2.9}$$

We use the relation [3]

$$\hat{T}_P(\tau, t) = \hat{T}(\tau, t) + \int_t^{\tau} \hat{T}(\tau, t) P(\tau) \hat{T}_P(\tau, t)\, d\tau. \tag{2.10}$$

Then

$$\begin{aligned}
\hat{N}(t) &= \int_t^{t_1} \left[\hat{T}^*(\tau, t) + \int_t^{\tau} \hat{T}_P^*(s, t) P^*(s) \hat{T}^*(\tau, s)\, ds \right] M(\tau) \hat{T}(\tau, t)\, d\tau \\
&= N(t) + \int_t^{t_1} \int_s^{t_1} \hat{T}_P^*(s, t) P^*(s) \hat{T}^*(\tau, s) M(\tau) \hat{T}(\tau, t)\, d\tau\, ds \\
&= N(t) + \int_t^{t_1} \hat{T}_P^*(s, t) P^*(s) \int_s^{t_1} \hat{T}^*(\tau, s) M(\tau) \hat{T}(\tau, s)\, d\tau \hat{T}(s, t)\, ds \\
&= N(t) + \int_t^{t_1} \hat{T}_P^*(s, t) P^*(s) N(s) \hat{T}(s, t)\, ds.
\end{aligned}$$

Here we have used Fubini's theorem for the second equality and the semigroup property $\hat{T}(\tau, t) = \hat{T}(\tau, s)\hat{T}(s, t)$ for the third equality. Now we claim:

$$\begin{aligned}
&\int_t^{t_1} \hat{T}_P^*(\tau, t) P^*(\tau) N(\tau) \hat{T}(\tau, t)\, d\tau + \int_t^{t_1} \hat{T}_P^*(\tau, t) \hat{N}(\tau) P(\tau) \hat{T}_P(\tau, t)\, d\tau \\
&\qquad = \int_t^{t_1} \hat{T}_P^*(\tau, t)[P^*(\tau)N(\tau) + N(\tau)P(\tau)]\hat{T}_P(\tau, t)\, d\tau.
\end{aligned} \tag{2.11}$$

In fact, using (2.9), we have

$$\begin{aligned}
\text{L.H.S.} &= \int_t^{t_1} \hat{T}_P^*(\tau, t)P^*(\tau)N(\tau)\hat{T}(\tau, t)\,d\tau \\
&\quad + \int_t^{t_1} \hat{T}_P^*(\tau, t)\left[N(\tau) + \int_\tau^{t_1} \hat{T}_P^*(s, \tau)P^*(s)N(s)\hat{T}(s, \tau)\,ds\right]P(\tau)\hat{T}_P(\tau, t)\,d\tau \\
&= \int_t^{t_1} \hat{T}_P^*(\tau, t)N(\tau)P(\tau)\hat{T}_P(\tau, t)\,d\tau + \int_t^{t_1} \hat{T}_P^*(s, t)P^*(s)N(s)\hat{T}(s, t)\,ds \\
&\quad + \int_t^{t_1}\int_t^{s} \hat{T}_P^*(s, t)P^*(s)N(s)\hat{T}(s, \tau)P(\tau)\hat{T}_P(\tau, t)\,d\tau\,ds \\
&= \int_t^{t_1} \hat{T}_P^*(\tau, t)N(\tau)P(\tau)\hat{T}_P(\tau, t)\,d\tau \\
&\quad + \int_t^{t_1} \hat{T}_P^*(\tau, t)P^*(s)N(s)\left[\hat{T}(s, t) + \int_t^{s} \hat{T}(s, \tau)P(\tau)\hat{T}_P(\tau, t)\,d\tau\right]ds \\
&= \int_t^{t_1} \hat{T}_P^*(\tau, t)N(\tau)P(\tau)\hat{T}_P(\tau, t)\,d\tau + \int_t^{t_1} \hat{T}_P^*(s, t)P^*(s)N(s)\hat{T}_P(s, t)\,ds \\
&= \text{R.H.S.}
\end{aligned}$$

Here we have used Fubini's theorem and (2.10). Finally, consider

$$\begin{aligned}
N(t) - \tilde{N}(t) &= \left[\int_t^{t_1} \hat{T}^*(\tau, t)M(\tau)\hat{T}(\tau, t)\,d\tau - \int_t^{t_1} \hat{T}_P^*(\tau, t)M(\tau)\hat{T}(\tau, t)\,d\tau\right] \\
&\quad + \left[\int_t^{t_1} \hat{T}_P^*(\tau, t)M(\tau)\hat{T}(\tau, t)\,d\tau - \int_t^{t_1} \hat{T}^*(\tau, t)M(\tau)\hat{T}(\tau, t)\,d\tau\right] \\
&= I_1 + I_2.
\end{aligned}$$

$$\begin{aligned}
I_1 &= \int_t^{t_1} [\hat{T}^*(\tau, t) - \hat{T}_P^*(\tau, t)]M(\tau)\hat{T}(\tau, t)\,d\tau \\
&= -\int_t^{t_1} \left[\int_t^{\tau} \hat{T}_P^*(\tau, t)P^*(s)\hat{T}^*(\tau, s)\,ds\right]M(\tau)\hat{T}(\tau, t)\,d\tau \\
&= -\int_t^{t_1}\int_s^{t_1} \hat{T}_P^*(s, t)P^*(s)\hat{T}^*(\tau, s)M(\tau)\hat{T}(\tau, s)\hat{T}(s, t)\,d\tau\,ds \\
&= -\int_t^{t_1} \hat{T}_P^*(s, t)P^*(s)\left[\int_s^{t} \hat{T}^*(\tau, s)M(\tau)\hat{T}(\tau, s)\,d\tau\right]\hat{T}(s, t)\,ds \\
&= -\int_t^{t_1} \hat{T}_P^*(s, t)P^*(s)N(s)\hat{T}(s, t)\,ds.
\end{aligned}$$

Similarly, we can show

$$I_2 = -\int_t^{t_1} \hat{T}_P^*(\tau, t)\hat{N}(\tau)P(\tau)\hat{T}_P(\tau, t)\,d\tau.$$

Hence

$$N(t)-\tilde{N}(t)=I_1+I_2$$
$$=-\int_t^{t_1}\hat{T}_P^*(\tau,t)P^*(\tau)N(\tau)\hat{T}(\tau,t)\,d\tau$$
$$-\int_t^{t_1}\hat{T}_P^*(\tau,t)N(\tau)P(\tau)\hat{T}_P(\tau,t)\,d\tau$$
$$=-\int_t^{t_1}\hat{T}_P^*(\tau,t)[P^*(\tau)N(\tau)+N(\tau)P(\tau)]\hat{T}_P(\tau,t)\,d\tau.$$

The last equality follows from (2.11). This completes the proof.

Using Lemma 2.3, we can show the equivalence of (2.1)′, (2.1)′ and (2.1)″. We shall prove this only for $F=\theta$. Since $M(t)$, $\hat{A}(t)$, $P(t)$ are arbitrary, we take

$$M(t)=W(t)+K(t)D(t)K(t),$$

and $\hat{T}(t,s)$ to be the evolution operator generated by $\hat{A}(t)=A(t)-D(t)K(t)$. Here $K(t)$ is the solution of (2.1). Then the definition of $N(t)$ gives the relation

$$K(t)=\int_t^{t_1}S_K^*(\tau,t)[W(\tau)+K(\tau)D(\tau)K(\tau)]S_K(\tau,t)\,d\tau. \tag{2.12}$$

The relation (2.8) states

$$K(t)=\int_t^{t_1}S_P^*(\tau,t)[W(\tau)+K(\tau)D(\tau)K(\tau)-P^*(\tau)K(\tau)$$
$$-K(\tau)P(\tau)]S_P(\tau,t)\,d\tau, \tag{2.13}$$

where $S_P(\tau,t)$ is generated by $A(\tau)-D(\tau)K(\tau)+P(\tau)$. If we substitute $D(t)K(t)$ for $P(t)$ in (2.13), we obtain (2.1)′ with $F=\theta$. If we substitute $D(t)K(t)+P(t)$ for $P(t)$, we find (2.1)″ with $F=\theta$.

Remark 2.3. Lukes and Russell [6] considered an integral equation of type (2.1)′, while Curtain and Pritchard [3] constructed a solution to an integral equation of type (2.1). But the one implies the other, and two integral equations are equivalent.

LEMMA 2.4. *Consider the control problem* (2.3) *with* $\tilde{A}(t)$, $\tilde{W}(t)$ *given by*

$$\tilde{A}(t)=A(t)-D_1(t)K(t),$$
$$\tilde{W}(t)=W(t)+K(t)D_1(t)K(t), \tag{2.14}$$

where $K(t)$ *is the solution of* (2.1), *and* $D_1(t)=B(t)U^{-1}(t)B^*(t)$. *Then* $K(t)$ *satisfies the integral equation* (2.6).

Proof. We shall prove this assertion for $F=\theta$. This is an immediate consequence of (2.13) when we set

$$P(t)=D_2(t)K(t)+C(t)L(t).$$

LEMMA 2.5. *Consider the control problem* (2.3) *with* $\tilde{A}(t)$, $\tilde{W}(t)$ *given in* (2.14). *The control law*

$$\bar{v}(t) = -V^{-1}(t)C^*(t)K(t)x$$

is optimal, if $K(t)$ *is the solution of* (2.1).

Proof. Since $K(t)$ satisfies (2.6), we have, in view of Lemma 2.2, that

$$K(t) \geqq Q(t)$$

for each $L(t)$. As in [3], we can construct a sequence of strongly continuous linear Operators $L_n(t)$ such that the corresponding operators $Q_n(t)$ satisfy

$$Q_1(t) \leqq Q_2(t) \leqq \cdots \leqq Q_n(t) \leqq \cdots.$$

Since $Q_n(t) \leqq K(t)$ for each n, $Q_n(t)$ converges to some operator $Q_\infty(t) \leqq K(t)$. But $Q_\infty(t)$ corresponds to a unique optimal control, so that we have

$$Q_\infty(t) \geqq K(t).$$

Hence $Q_\infty(t) = K(t)$.

LEMMA 2.6. *The control law*

$$\bar{u}(t) = -U^{-1}(t)B^*(t)K(t)x$$

is a unique optimal control for the control problem

$$\dot{x} = (A(t) - D_2(t)K(t))x + B(t)u,$$

$$x(t_0) = x_0,$$

and

$$J(u) = (Fx(t_1), x(t_1)) + \int_{t_0}^{t_1} [([W(\tau) + K(\tau)D_2(\tau)K(\tau)]x(\tau), x(\tau)) + (U(\tau)u(\tau), u(\tau))]\, d\tau.$$

This is the exact counterpart of Lemma 2.5, and therefore the proof is omitted.

Proof of Theorem 2.1. The theorem follows directly from the preceding results. In fact, Lemma 2.5 gives

$$J(\bar{u}, v) \leqq J(\bar{u}, \bar{v}),$$

and Lemma 2.6 tells us

$$J(\bar{u}, \bar{v}) \leqq J(u, \bar{v}).$$

The relation $(K(t_0)x_0, x_0) = J(\bar{u}, \bar{v})$ follows, for example, from (2.4) with $L(t) = -D_2(t)K(t)$ and the observation that $K(t) = Q(t)$ for this particular $L(t)$.

3. Quadratic games with open-loop controls. We consider the same differential system (1.1), (1.2) and the payoff functional (1.3). Admissible controls are now $L_2(I; H_i)$-functions. Our solution of the system (1.1), (1.2) is defined by (1.4).

Define operators P, P_1, Q and Q_1 by

$$(Pu)(t)=\int_{t_0}^{t} S(t,\tau)B(\tau)u(\tau)\,d\tau, \qquad P_1u=(Pu)(t_1), \qquad u\in L_2(I;H_2),$$
$$(Qv)(t)=\int_{t_0}^{t} S(t,\tau)C(\tau)V(\tau)\,d\tau, \qquad Q_1v=(Qv)(t_1), \qquad v\in L_2(I;H_3) \tag{3.1}$$

Then, $P, Q\in B[L_2(I;H_i);L_2(I;H_1)]$, $i=2,3$, and $P_1, Q_1\in B[L_2(I;H_i);H_1]$, $i=2,3$. Here $B[X;Y]$ denotes the set of linear bounded operators mapping X into Y. Let $r(t)=S(t,t_0)x_0$, $r_1=r(t_1)$; then $r(t)\in L_2(I;H_1)$. We denote by $x_{u,v}(t)$ the solution of (1.1), (1.2) corresponding to a control pair $\{u,v\}$. Then we have

$$x_{u,v}(t)=r(t)+(Pu)(t)+(Qv)(t) \tag{3.2}$$

and

$$\begin{aligned} J(u,v)&=(F[P_1u+Q_1v+r_1],P_1u+Q_1v+r_1)\\ &\quad+\langle W(Pu+Qv+r),Pu+Qv+r\rangle+\langle Uu,u\rangle+\langle Vv,v\rangle\\ &=\langle(P_1^*FP_1+P^*WP+U)u,u\rangle+2\langle P_1^*F(Q_1v+r_1),u\rangle\\ &\quad+2\langle P^*W(Qv+r),u\rangle+\langle(Q_1^*FQ_1+Q^*WQ+V)v,v\rangle\\ &\quad+2\langle Q_1^*Fr_1+Q^*Wr,v\rangle+\langle Wr,r\rangle+(Fr_1,r_1). \end{aligned} \tag{3.3}$$

Here * denotes the adjoint of an operator.

THEOREM 3.1. *Assume that*

A1: $$V+Q_1^*FQ_1+Q^*WQ<0 \quad \text{on } L_2(I;H_3)$$

holds, then there exists a unique optimal pair $\bar u$, $\bar v$ satisfying the relation

$$\bar u(t)=-U^{-1}(t)B^*(t)\Big[S^*(t_1,t)Fx_{\bar u,\bar v}(t_1)+\int_t^{t_1} S^*(\tau,t)Wx_{\bar u,\bar v}(\tau)\,d\tau\Big],$$
$$\bar v(t)=-V^{-1}(t)C^*(t)\Big[S^*(t_1,t)Fx_{\bar u,\bar v}(t_1)+\int_t^{t_1} S^*(\tau,t)Wx_{\bar u,\bar v}(\tau)\,d\tau\Big], \tag{3.4}$$

where $x_{\bar u,\bar v}(t)$ is the optimal trajectory of (1.1), (1.2).

Proof. Under Assumption A1, $J(u,v)$ is strictly convex and lower semicontinuous in u, and strictly concave and upper semicontinuous in v. Hence there exists a unique saddle point, which is given by the solution of

$$\nabla_1J(u,v)=2(P_1^*FP_1+P^*WP+U)u+2P_1^*F(Q_1v+r_1)+2P^*W(Qv+r)=0,$$
$$\nabla_2J(u,v)=2(Q_1^*FQ_1+Q^*WQ+V)v+2Q_1^*F(P_1u+r_1)+2Q^*W(Pu+r)=0. \tag{3.5}$$

Here $\nabla_1J(u,v)$, $\nabla_2J(u,v)$ are partial Fréchet derivatives of $J(u,v)$ with respect to u, v, respectively. Thus $\{\bar u,\bar v\}$ satisfies the following:

$$\bar u=-U^{-1}[P_1^*F(P_1\bar u+Q_1\bar v+r_1)+P^*W(P\bar u+Q\bar v+r)],$$
$$\bar v=-V^{-1}[Q_1^*F(Q\bar u+Q_1\bar v+r_1)+Q^*W(P\bar u+Q\bar v+r)].$$

Hence we have

$$\begin{aligned}\bar{u}(t) &= -U^{-1}(t)[P_1^* F x_{\bar{u}\bar{v}}(t_1) + P^* W x_{\bar{u},\bar{v}}](t),\\ \bar{v}(t) &= -V^{-1}(t)[Q_1^* F x_{\bar{u},\bar{v}}(t_1) + Q^* W x_{\bar{u},\bar{v}}](t).\end{aligned} \tag{3.6}$$

Since we have relations such as

$$(P_1^* h)(t) = B^*(t)S^*(t_1, t)h, \qquad h \in H_1,$$

$$(P^* y)(t) = \int_t^{t_1} B^*(\tau)S^*(\tau, t)y(\tau)\, d\tau, \quad y(t) \in L_2(I; H_1),$$

we may rewrite (3.6) to obtain (3.4).

Remark 3.1. For the existence of an optimal pair, we require only that

$$V + Q_1^* F Q_1 + Q^* W Q \leqq 0.$$

Let $s \in I$ be arbitrary, and let $I_s = [s, t_1]$. We define operators P_s, P_{1s}, Q_s and Q_{1s} on $L_2(I_s; H_i)$ as in (3.1), with t_0 replaced by s.

LEMMA 3.1. *Assumption* A1 *implies that*

$$V + Q_{1s}^* F Q_{1s} + Q_s^* W Q_s < 0, \qquad s \in I. \tag{3.7}$$

In view of the above lemma, quadratic games on I_s for each $s \in I$ are well-defined and have a unique optimal pair for any initial value $h \in H_1$. Let $x(\cdot, s, h)$ denote the unique optimal trajectory with initial condition h for the quadratic game on I_s. We define an operator $K(s)$, $s \in I$, on H_1 by

$$K(s)h = S^*(t_1, s)Fx(t_1, s, h) + \int_s^{t_1} S^*(\tau, s)W(\tau)x(\tau, s, h)\, d\tau. \tag{3.8}$$

We shall rewrite the optimal pair (3.4) in terms of this operator $K(s)$. Since $x(\cdot, t_0, x_0)$ restricted on I_s is again optimal trajectory on I_s corresponding to the initial value $x(s, t_0, x_0)$, we arrive at the following relation:

$$x(\tau, s, x(s, t_0, x_0)) = x(\tau, t_0, x_0), \qquad \tau \in I_s.$$

Hence (3.4) has an equivalent form

$$\begin{aligned}\bar{u}(t) &= -U^{-1}(t)B^*(t)K(t)x(t, t_0, x_0),\\ \bar{v}(t) &= -V^{-1}(t)C^*(t)K(t)x(t, t_0, x_0).\end{aligned} \tag{3.9}$$

LEMMA 3.2. *$K(t)$, $t \in I$, is a linear bounded self-adjoint operator mapping H_1 into itself. Furthermore, $K(t) \geqq 0$, and for any $h_0, h_1 \in H_1$, $t \in I$,*

$$\begin{aligned}(K(t)h_0, h_1) = (Fx(t_1, t, h_0), x(t_1, t, h_1)) + \int_t^{t_1} [&(W(\tau)x(\tau, t, h_0), x(\tau, t, h_1))\\ &+ (U(\tau)u_0(\tau), u_1(\tau)) + (V(\tau)v_0(\tau), v_1(\tau))]\, d\tau.\end{aligned} \tag{3.10}$$

Here $\{u_i, v_i\}$, $i = 0, 1$, is an optimal pair on I_t for the initial value h_i.

The proof is similar to Theorem 2 in [4]. Summing up, we have the following.

THEOREM 3.2. *Consider the quadratic game* (1.1), (1.2) *and* (1.3). *Let Assumption* A1 *hold. Then there exists a unique optimal pair $\bar{u}$, $\bar{v}$ given by* (3.9),

and

$$J(\bar{u}, \bar{v}) = (K(t_0)x_0, x_0).$$

Below we shall examine the properties of $K(t)$ and show that $K(t)$ satisfies the Riccati integral equation (2.1).

LEMMA 3.3. *Let $\{u_t, v_t\}$ be the optimal pair for the quadratic game on I_t, $t \in I$. Then $\{u_t\}$, $\{v_t\}$, $t \in I$, are uniformly bounded, or, equivalently, $K(t)$ is uniformly bounded in t.*

Proof. Suppose the contrary. Without loss of generality, we can assume the existence of $\{\tau_n\} \downarrow t_0$, $\tau_n \in I$, such that at least one of $\{u_n = u_{\tau_n}\}$, $\{v_n = v_{\tau_n}\}$ tends to $+\infty$ in norm. By the saddle point condition, we have

$$J_n(u_n, \bar{v}) \leqq J_n(u_n, v_n) \leqq J_n(\bar{u}, v_n), \tag{3.11}$$

where $J_n(\cdot\,, \cdot)$ is the payoff functional on $[\tau_n, t_1] = I_n$. We note that $\bar{u}, \bar{v}$ is the optimal pair given in Theorem 3.2, and we understand $\bar{u}, \bar{v}$ in (3.11) to be restrictions of $\bar{u}, \bar{v}$ on $[\tau_n, t_1]$. If $\{u_n\}, \{v_n\}$ are bounded, then so are $J_n(u_n, \bar{v})$, $J_n(\bar{u}, v_n)$. Since $J_n(\bar{u}, v_n) \downarrow -\infty$ as $\|v_n\|_n \to +\infty$, and $J_n(u_n, \bar{v}) \uparrow +\infty$ as $\|u_n\|_n \to +\infty$, we have in either case a contradiction. Here $\|\cdot\|_n$ is the norm in $L_2(I_n; H_i)$.

LEMMA 3.4. *$K(t) \to K(s)$ strongly as $t \uparrow s$.*

Proof. Let $t_0 < \tau_2 < \tau_1 < t_1$, and let $J_i(\cdot\,, \cdot)$, $i = 1, 2$, be the payoff functional on $[\tau_i, t_1]$. The optimal pair on $[\tau_i, t_1]$ will be denoted by $\{u_i, v_i\}$. The controls u_1, v_1 are defined on $[\tau_1, t_1]$, but we extend them on $[\tau_2, t_1]$. Thus expressions $J_2(u_1, v_2)$, $J_2(u_2, v_1)$ are meaningful. Similarly, when we write $J_1(u_1, v_2)$, $J_1(u_2, v_1)$, we understand u_2, v_2 to be restrictions on $[\tau_1, t_1]$. Consider

$$\begin{aligned}
J_2(u_2, v_2) \geqq J_2(u_2, v_1) &= (Fx_{2,1}(t_1, \tau_2), x_{2,1}(t_1, \tau_2)) \\
&\quad + \int_{\tau_2}^{t_1} [(W(\tau)x_{2,1}(\tau, \tau_2), x_{2,1}(\tau, \tau_2)) + (U(\tau)u_2(\tau), u_2(\tau)) \\
&\qquad + (V(\tau)v_1(\tau), v_1(\tau))]\, d\tau \\
&= (Fx_{2,1}(t_1, \tau_2), x_{2,1}(t_1, \tau_2)) \\
&\quad + \int_{\tau_1}^{t_1} [(W(\tau)x_{2,1}(\tau, \tau_2), x_{2,1}(\tau_2)) + (U(\tau)u_2(\tau), u_2(\tau)) \\
&\qquad + (V(\tau)v_1(\tau), v_1(\tau))]\, d\tau \\
&\quad + \int_{\tau_2}^{\tau_1} [(W(\tau)x_{2,1}(\tau, \tau_2), x_{2,1}(\tau, \tau_2)) \quad + (U(\tau)u_2(\tau), u_2(\tau))]\, d\tau,
\end{aligned} \tag{3.12}$$

where $x_{ij}(\cdot\,, \tau_i)$ denotes the trajectory on $[\tau_i, t_1]$ corresponding to the control pair u_i, v_j. Note that, using decomposition as in (3.2), we obtain

$$\begin{aligned}
x_{2,1}(\tau, \tau_2) &= x_{2,1}(\tau, \tau_1) + [S(\tau, \tau_2) - S(\tau, \tau_1)]x_0 + \int_{\tau_2}^{\tau_1} S(\tau, \sigma)B(\sigma)u_2(\sigma)\, d\sigma \\
&\equiv x_{2,1}(\tau, \tau_1) + \alpha(\tau)
\end{aligned}$$

and

$$x_{2,1}(t_1, \tau_2) = x_{2,1}(t_1, \tau_1) + [S(t_1, \tau_2) - S(t_1, \tau_1)]x_0 + \int_{\tau_2}^{\tau_1} S(t_1, \sigma)B(\sigma)u_2(\sigma)\, d\sigma$$

$$\equiv x_{2,1}(t_1, \tau_1) + \beta.$$

Then, from (3.12),

$$\begin{aligned} J_2(u_2, v_2) &\geqq (F[x_{2,1}(t_1, \tau_1) + \beta], x_{2,1}(t_1, \tau_1) + \beta) \\ &\quad + \int_{\tau_1}^{t_1} (W[x_{2,1}(\tau, \tau_1) + \alpha(\tau)], x_{2,1}(\tau, \tau_1) + \alpha(\tau))\, d\tau \\ &\quad + \int_{\tau_1}^{t_1} [(U(\tau)u_2(\tau), u_2(\tau)) + (V(\tau)v_1(\tau), v_1(\tau))]\, d\tau \\ &\quad + \int_{\tau_2}^{\tau_1} [(W(\tau)x_{2,1}(\tau, \tau_2), x_{2,1}(\tau, \tau_2)) + (U(\tau)u_2(\tau), u_2(\tau))]\, d\tau \\ &= (Fx_{2,1}(t_1, \tau_1), x_{2,1}(t_1, \tau_1)) \\ &\quad + \int_{\tau_1}^{t_1} [(Wx_{2,1}(\tau, \tau_1), x_{2,1}(\tau, \tau_1) + (U(\tau)u_2(\tau), u_2(\tau)) \\ &\qquad + (V(\tau)v_1(\tau), v_1(\tau))]\, d\tau + \gamma \\ &= J_1(u_2, v_1) + \gamma \\ &\geqq J_1(u_1, v_1) + \gamma, \end{aligned} \tag{3.13}$$

where

$$\begin{aligned} \gamma &= (F\beta, \beta) + 2(Fx_{2,1}(t_1, \tau_1), \beta) \\ &\quad + \int_{\tau_1}^{t_1} [(W(\tau)\alpha(\tau), \alpha(\tau)) + 2(W(\tau)x_{2,1}(\tau, \tau_1), \alpha(\tau))]\, d\tau \\ &\quad + \int_{\tau_2}^{\tau_1} [(W(\tau)x_{2,1}(\tau, \tau_2), x_{2,1}(\tau, \tau_2)) + (U(\tau)u_2(\tau), u_2(\tau))]\, d\tau. \end{aligned}$$

Since $\{u_i, v_i\}$, $i = 1, 2$, are bounded by Lemma 3.3, we have an estimate

$$\|\alpha\|_1 = O(\tau_1 - \tau_2), \qquad |\beta| = O(\tau_1 - \tau_2),$$

and hence

$$|\gamma| = O(\tau_1 - \tau_2).$$

Similarly, we can show that

$$J_2(u_2, v_2) \leqq J_1(u_1, v_1) + \gamma' \tag{3.14}$$

for some γ' with $|\gamma'| = O(\tau_1 - \tau_2)$. Combining (3.13), (3.14) gives us our result.

LEMMA 3.5. *If the system* (1.1), (1.2) *and* (1.3) *is time-invariant, namely,* $A(t) = A$, $B(t) = B, \cdots$, *then* $K(t)$ *is nonincreasing in* t, *i.e.,* $K(t) \geqq K(s) \geqq 0$ *if* $s > t$.

Proof. Since $K(t)$ depends on t_1, we denote it by $K_{t_1}(t)$ to be precise. Since the system is time-invariant, it is easy to see that $K_{t_1}(t) = K_{t_1-t}(0)$. Hence it is sufficient to show that $t_0 \leqq s_1 < s_2 \leqq t_1$ implies

$$0 \leqq J_1(u_1, v_1) \leqq J_2(u_2, v_2),$$

where $J_i(\cdot, \cdot)$, $i = 1, 2$, is the payoff functional on $[t_0, s_i]$ defined by (1.3) with $t_1 = s_i$, and $\{u_i, v_i\}$ is the optimal pair. As in Lemma 3.4, consider

$$\begin{aligned}
J_2(u_2, v_2) &\leqq J_2(u_2, v_1) \\
&= (Fx_{2,1}(s_2), x_{2,1}(s_2)) + \int_{t_0}^{s_2} [(W(\tau)x_{2,1}(\tau), x_{2,1}(\tau)) + (U(\tau)u_2(\tau), u_1(\tau)) \\
&\qquad + (V(\tau)v_1(\tau), v_1(\tau))]\, d\tau \\
&= (Fx_{2,1}(s_1), x_{2,1}(s_1)) + \int_{t_0}^{s_1} [(W(\tau)x_{2,1}(\tau), x_{2,1}(\tau)) + (U(\tau)u_2(\tau), u_2(\tau)) \\
&\qquad + (Vv_1(\tau), (\tau))]\, d\tau \qquad (3.15) \\
&\quad - (Fx_{2,1}(s_1), x_{2,1}(s_1)) + \Big[(Fx_{2,1}(s_2), x_{2,1}(s_2)) \\
&\qquad + \int_{s_1}^{s_2} [(W(\tau)x_{2,1}(\tau), x_{2,1}(\tau)) + (Uu_2(\tau), u_2(\tau))]\, d\tau \Big] \\
&= J_1(u_2, v_1) - (Fx_{2,1}(s_1), x_{2,1}(s_1)) + J_{12}(u_2) \\
&\geqq J_1(u_1, v_1) - (Fx_{2,1}(s_1), x_{2,1}(s_1)) + J_{12}(u_2).
\end{aligned}$$

Here

$$\begin{aligned}
J_{12}(u_2) &= (Fx_{2,1}(s_2), x_{2,1}(s_2)) \\
&\quad + \int_{s_1}^{s_2} [(W(\tau)x_{2,1}(\tau), x_{2,1}(\tau)) + (U(\tau)u_2(\tau), u_2(\tau))]\, d\tau.
\end{aligned}$$

Consider the control problem

$$\begin{aligned}
\dot{x} &= Ax + Bu, \\
x(s_1) &= x_{2,1}(s_1),
\end{aligned}$$

with a cost functional

$$J_{12}(u) = (Fx(s_2), x(s_2)) + \int_{s_1}^{s_2} [(W(\tau)x(\tau), x(\tau)) + (Uu(\tau), u(\tau))]\, d\tau. \qquad (3.16)$$

It is known [3] that there exists a unique minimizing control given by

$$\bar{u}_{12}(t) = -U^{-1}(t)B^*(t)R(t)x,$$

where $R(t)$ is the solution of

$$R(t) = S_R^*(s_2, t)FS_R(s_2, t) + \int_t^{s_2} S_R^*(\tau, t)[W(\tau) + R(\tau)D_1(\tau)R(\tau)]S_R(\tau, t)\, d\tau.$$

$S_R(\tau, t)$ is the evolution operator generated by $A(\tau) - D_1(\tau)R(\tau)$, and $D_1(\tau)$

$= B(\tau)U^{-1}(\tau)B^*(\tau)$. Note that

$$F \leqq R(t) \leqq R(\tau) \quad \text{if } t > \tau, \qquad t, \tau \in [s_1, s_2].$$

Since

$$J_{12}(\bar{u}_{12}) = \min_u J_{12}(u) \leqq J_{12}(u_2)$$

and

$$J_{12}(\bar{u}_{12}) = (R(s_1)x_{2,1}(s_1), x_{2,1}(s_1)),$$

we conclude

$$\begin{aligned} J_{12}(u_2) &\geqq (R(s_1)x_{2,1}(s_1), x_{2,1}(s_1)) \\ &\geqq (Fx_{2,1}(s_1), x_{2,1}(s_1)). \end{aligned}$$

Thus we have shown through (3.15) that

$$J_2(u_2, v_2) \geqq J_1(u_1, v_1).$$

From Theorem 3.2, the optimal solution $x(t, t_0, x_0)$ is given by the solution of

$$\begin{aligned} &\dot{x} = (A(t) - D(t)K(t))x, \\ &x(t_0) = x_0. \end{aligned} \tag{3.17}$$

Let $S_K(t, s)$ be an evolution operator generated by $A(t) - D(t)K(t)$. Then

$$x(t, t_0, x_0) = S_K(t, t_0)x_0.$$

THEOREM 3.3. *The operator $K(t)$ defined by* (3.8) *satisfies the Riccati integral equation* (2.1):

$$K(t) = S_K^*(t_1, t)FS_K(t_1, t) + \int_t^{t_1} S_K^*(\tau, t)[W(\tau) + K(\tau)D(\tau)K(\tau)]S_K(\tau, t)\, d\tau.$$

Proof. Let $h_0, h_1 \in H_1$. Then

$$x(\tau, t, h_i) = S_K(\tau, t)h_i, \qquad u_i(\tau) = -U^{-1}(\tau)B^*(\tau)K(\tau)S_K(\tau, t)h_i$$

and

$$v_i(\tau) = -V^{-1}(\tau)C^*(\tau)K(\tau)S_K(\tau, t)h_i, \qquad i = 0, 1.$$

Hence (3.10) may be rewritten as

$$\begin{aligned} (K(t)h_0, h_1) &= (FS_K(t_1, t)h_0, S_K(t_1, t)h_1) \\ &\quad + \int_t^{t_1} ([W(\tau) + K(\tau)D(\tau)K(\tau)]S_K(\tau, t)h_0, S_K(\tau, t)h_1)\, d\tau. \end{aligned}$$

Since h_0, h_1 are arbitrary, we obtain (2.1).

Remark 3.2. This theorem tells us that (18) in [4] is essentially equivalent to the Riccati integral equation in [3].

4. A quadratic game on an infinite interval. Consider the time-invariant system

$$\dot{x} = Ax + Bu + Cv, \tag{4.1}$$

$$x(0) = x_0, \tag{4.2}$$

with payoff functional

$$J(u, v) = \int_0^\infty [(Wx, x) + (Uu, u) + (Vv, v)]\, dt. \tag{4.3}$$

We take the sets of admissible control functions to be $L_2(R^+; H_i)$. In general, $J(u, v)$ may not be finite. But, if we impose a strong condition

A2: The semigroup $S(t)$ generated by A is exponentially stable, namely,

$$|S(t)| \leqq Me^{-\alpha t} \quad \text{for some } \alpha > 0,\ M \geqq 1,$$

then $J(u, v)$ is always finite (Lemma 4.1). A sufficient condition for A2 is given by

$$|R(\lambda; A)| \leqq \frac{M}{|\lambda| + \omega}, \qquad \lambda \in S_\phi,$$

for some $M, \omega > 0$. Here $R(\lambda; A)$ is the resolvent of A, and

$$S_\phi = \left\{\lambda;\ \lambda \neq 0, \frac{\pi}{2} - \phi < \arg \lambda < \frac{3}{2}\pi + \phi\right\}, \qquad 0 < \phi < \frac{\pi}{2}.$$

Define mappings P, Q on $L_2(R^+; H_i)$ by

$$\begin{aligned} (Pu)(t) &= \int_0^t S(t-\tau)Bu(\tau)\, d\tau, \\ (Qv)(t) &= \int_0^t S(t-\tau)Cv(\tau)\, d\tau. \end{aligned} \tag{4.4}$$

Then $P, Q \in B[L_2(R^+; H_i); L_2(R^+; H_1)]$, $i = 2, 3$. This follows from Assumption A2 and the following lemma.

LEMMA 4.1. *Let $x(t)$, $k(t)$ be numerically-valued function in $L_2(R^+)$, $L_1(R^+)$, respectively. Then the function defined by*

$$y(t) = \int_0^t k(t-\tau)x(\tau)\, d\tau$$

is in $L_2(R^+)$. Furthermore, $\|y\|_2 \leqq \|k\|_1 \|x\|_2$.

We further assume

A3: $$V + Q^*WQ < 0 \quad \text{on } L_2(R^+; H_3).$$

Then we can derive results analogous to those in § 2. We shall state results without proofs.

THEOREM 4.1. *Let Assumptions* A2, A3 *hold. Then there exist a linear self-adjoint nonnegative time-invariant operator K and a unique optimal pair*

u^∞, v_∞ such that

$$u_\infty = -U^{-1}B^*Kx_\infty,$$

$$v_\infty = -V^{-1}C^*Kx_\infty.$$

Here x_∞ is the optimal trajectory given by

$$x_\infty(t) = S_K(t)x_0,$$

and $S_K(t)$ is a semigroup generated by $A - DK$. Moreover,

$$J(u_\infty, v_\infty) = (Kx_0, x_0).$$

THEOREM 4.2. *The operator K satisfies the Riccati equation*

$$K = \int_t^\infty S_K^*(\tau - t)[W + KDK]S_K(\tau - t)\, d\tau, \tag{4.5}$$

which is independent of t. From (4.5) *we can derive an inner product Riccati equation*

$$(KAh_0, h_1) + (Kh_0, Ah_1) + (Wh_0, h_1) - (KDKh_0, h_1) = 0$$

for any $h_0, h_1 \in D(A)$.

Remark 4.1. Since $h \in D(A)$ *implies* $h \in D(A - DK)$, $x_\infty(t) = S_K(t)x_0$ is differentiable and satisfies

$$\dot{x} = (A - DK)x,$$

$$\dot{x}(0) = x_0,$$

if $x_0 \in D(A)$. Hence $x_\infty(t)$ is a strict solution for each $x_0 \in D(A)$.

Remark 4.2. Assumption A3 implies A1 with $F = \theta$ for each interval $I = [t_0, t_1] \subset R^+$. Hence there exists a unique optimal pair for (1.1), (1.2) and (1.3) with $F = \theta$. Let $I_n = [0, t_n]$ with $t_n \uparrow \infty$, and let $\{u_n, v_n\}$ denote the optimal pair. Then, using Lemma 3.5, we can show that

$$K_{t_n}(t) \to K \quad \text{strongly for each fixed } t,$$

and

$$\left.\begin{aligned} u_n &\to u_\infty \\ v_n &\to v_\infty \end{aligned}\right\} \text{in } L_2(R^+; H_i).$$

Remark 4.3. It is known that differential game theory can be applied for sensitivity design and control problems with uncertainty to obtain an upper bound of the cost. Assumption A2 is very restrictive, but the results of this section may be used to obtain an upper bound of the cost of a regulator problem (at least in a finite-dimensional space).

Acknowledgment. The author would like to thank Professors V. Dolezal and U. Haussman for their encouragement and helpful discussions.

REFERENCES

[1] A. BENSSOUSAN, *Saddle points of convex concave functionals with applications to linear quadratic differential games*, Differential Games and Related Topics, H. W. Kuhn and G. P. Szego, eds., American Elsevier, New York, 1971, pp. 177–199.

[2] ———, *Points de Nash dans le cas de fonctionnelles quadratiques et jeux différentiels linéares à N personnes*, this Journal (1974), pp. 460–499.

[3] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite-dimensional Riccati equation*, J. Math. Anal. Appl., 47 (1974), pp. 43–57.

[4] R. DATKO, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[5] E. HILLE AND R. S. PHILLIPS, *Functional analysis and semi-groups*, Colloquium Publications, vol. 31, American Mathematical Society, Providence, R.I., 1957.

[6] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

[7] A. J. PRITCHARD, *Stability and control of distributed parameter systems governed by wave equations*, IFAC Conference on Distributed Parameter Systems, Banff, 1971.

[8] W. E. SCHMITENDORF, *Existence of optimal open-loop strategies for a class of differential games*, J. Optimization Theory Appl., 5 (1970), pp. 363–375.

[9] R. TEMAM, *Sur l'équation de Riccati associée à des opérateurs non bornés, en dimension infinie*, J. Functional Analysis, 7 (1971), pp. 85–115.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

OPTIMAL CONTROL PROBLEMS IN SOBOLEV SPACES WITH WEIGHTS*

CLAUDIA SIMIONESCU†

Abstract. We consider an optimal control problem in certain Sobolev spaces with weights used systematically by F. Trèves in [1], [2]. The notations and definitions are the same as in [2] and [3].

1. Definitions and background. Let E be a real Hilbert space and let $q(t)$ be a real function with continuous derivatives that satisfies the condition:

(A) there exists a constant $p_0 > 0$ such that $|q'(t)| \geqq p_0$ for every real t.

If k is an integer ($k \in Z$), then $\mathscr{D}^k(q; E)$ denotes the Hilbert space obtained by completing $\mathscr{D}(E)$[1] with respect to the structure defined by the Hermitian product

$$(1) \qquad (\varphi, \psi)_{E;q,k} = (e^{-q(t)}D^k\varphi, e^{-q(t)}D^k\psi)_{L^2(E)} = \int (e^{-q(t)}D^k\varphi, e^{-q(t)}D^k\psi)_E \, dt.$$

Consequently, the norm of $\mathscr{D}^k(q; E)$ is

$$(2) \qquad \|\varphi\|_{E;q,k} = \|e^{-q(t)}D^k\varphi\|_{L^2(E)} = \left[\int \|e^{-q(t)}D^k\varphi\|_E^2 \, dt\right]^{1/2}.$$

In [2] the following results are shown:

(i) If k, $h \in Z$, $h \leqq k$, and $q(t)$ verifies condition (A), then there exists a continuous mapping from $\mathscr{D}^k(q; E) \to \mathscr{D}^h(q; E)$ such that

$$(3) \qquad \|f\|_{E;q,h} \leqq \sup_{t \in R} \frac{1}{|q'(t)|^{k-h}} \|f\|_{E;q,k} \quad \text{for all } f \in \mathscr{D}^k(q; E).$$

(ii) For each integer $k \in Z$ and each real function $q(t)$ verifying (A), $\mathscr{D}^k(q; E)$ is a space of distributions with values in E, that is,

$$\mathscr{D}^k(q; E) \subset \mathscr{D}'(E).$$

(iii) If k is an integer and $k \geqq 0$, then $\mathscr{D}^k(q; E)$ is the space of all (classes of) measurable functions from R into E such that $e^{-q(t)}D^h u(t) \in L^2(E)$ for every $0 \leqq h \leqq k$.

(iv) For each $k \in Z$, $k \geqq 0$, $\mathscr{D}^{-h}(q; E)$ is a space of distributions having the property: if $T \in \mathscr{D}^{-k}(q; E)$, then there exists $k+1$ functions $g_h \in L^2(E)$ such that

$$T = e^q g_0 + D(e^q g_1) + \cdots + D^k(e^q g_k).$$

---

* Received by the editors June 27, 1974.

† Facultatea de Stiinte, Universitatea din Braşov, Braşov–Romania. Results presented in the paper were obtained while the author was an IREX Fellow in the Department of System Science, University of California, Los Angeles, California, 1973–1974.

[1] $\mathscr{D}(E)$ is the space of functions defined on the real line taking values in E, with derivatives of all orders and compact support.

(v) If $A(t) \in \mathcal{B}_t(\mathcal{L}(E, F))$[2], then $f \to A(t)f$ is a continuous linear mapping from $\mathcal{D}^k(q; E) \to \mathcal{D}^k(q; F)$.

(vi) Let us consider two Hilbert spaces E and F, a positive constant p_0 and an operator $A(t) \in \mathcal{E}_t(\mathcal{L}(E, F))$. Then for every $\varepsilon > 0$ and $k \in Z$, there exists a positive function with continuous derivatives $G_{k,\varepsilon}$ such that if a function $q(t)$ satisfies the inequality

$$|G'_{k,\varepsilon}| + p_0 \leqq |q'(t)|,$$

then $A(t)$ is a bounded operator from $\mathcal{D}^k(q; E) \to \mathcal{D}^k(q + G_{k,\varepsilon}; F)$ of norm less than ε.

2. Statement of the problem. Let us consider a real Hilbert space $\mathcal{U}$ and suppose that

(a) $\pi(u, v)$ is a symmetric bilinear continuous form on $\mathcal{U}$,

(b) $L(v)$ is a linear continuous form on $\mathcal{U}$,

(c) $\mathcal{U}_{ad}$ is a closed convex subset of $\mathcal{U}$ (the set of admissible controls), and

(d) $\mathcal{J}(v) = \pi(v, v) - 2L(v)$ is a quadratic functional on $\mathcal{U}$.

EXISTENCE AND UNIQUENESS THEOREM [3]. *If the form $\pi(u, v)$ is coercive on $\mathcal{U}$, then there exists a unique element $u \in \mathcal{U}_{ad}$, such that*

$$\mathcal{J}(u) =) \inf_{v \in \mathcal{U}_{ad}} \mathcal{J}(v)$$

(*u is the optimal control*).

Let V and H be two real Hilbert spaces, $V \subset H$, V dense in H, and let the injection $V \to H$ be continuous. We identify H to its dual so that if V' denotes the dual of V, we have $V \subset H \subset V'$.

We consider now the integro-differential operator

$$P(t, D) = \sum_{r=-n}^{1} \mathcal{B}_r(t) D^r, \qquad n \in \mathcal{N}, \quad \mathcal{B}_r(t) \in \mathcal{E}_t(\mathcal{L}(H, H)), \tag{4}$$

where $\mathcal{B}_r(t)$ is a Hermitian operator and satisfies the condition:

(I) there exists a function $b(t) \in \mathcal{E}_t$. $b(t) > 0$, such that for every $g \in H$,

$$(B_1(t)g, g)_H \geqq b(t)\|g\|_H^2 \quad \text{for all } t \in R. \tag{5}$$

Let $a(t; u, v)$ be a bilinear Hermitian continuous form on $V \times V$ such that for every $u, v \in V$,

(II) $|a(t; u, v)| \leqq k\|u\| \cdot \|v\|, \quad k > 0,$

(III) $a(t; u, u) \geqq \alpha(t)\|u\|_V^2, \quad \alpha(t) > 0 \quad \text{for all } t \in R.$

We suppose that the mapping $t \to a(t; u, v)$ is measurable.

Then there exists an operator $A(t) \in \mathcal{E}_t(\mathcal{L}(V, V))$ such that for all $u, v \in V$, we have

$$a(t; u, v) = (A(t)u, v)_V,$$

---

[2] $\mathcal{B}_t(\mathcal{L}(E, F))$ is a subspace of $\mathcal{E}_t(\mathcal{L}(E, F))$.($\mathcal{E}_t(\mathcal{L}(E, F))$ is the space of all bounded continuous operators from E into F which have continuous derivatives of every order with respect to t) and contains all functions $g(t)$ having the following property: for each integer r, $r \geqq 0$, there exists $\mathcal{B}_r < \infty$, such that

$$\|g^{(r)}(t)\|_{\mathcal{L}(E,F)} \leqq \mathcal{B}_r$$

for every real t.

and consequently:

(IV) there exists $\alpha(t)>0$ such that

$$(A(t)u, u)_V \geqq \alpha(t)\|u\|_V^2 \quad \text{for all } t \in R.$$

3. Results.

3.1. Consider a system governed by the integro-differential operator $P(t, D)$ defined by (4), such that the state y of the system is given by the solution of the equation

$$(A(t)y, u)_V + (P(t, D)y, u)_H = (g + \mathcal{B}v, u)_H, \qquad u \in V, \tag{6}$$

in the sense of scalar distributions. Suppose that $\mathcal{U} = V$. Then we have the following.

PROPOSITION 3.1. *If $p(t)$ is a real function that verifies the condition* (A) *and*

(a) $\mathcal{B} \in \mathcal{L}(V; \mathcal{D}^k(p; H))$,

(b) $g \in \mathcal{D}^k(p; H)$,

(c) $N \in \mathcal{L}(V, V)$ *is Hermitian and coercive,*

(d) $A(t)$ *satisfies the property* (IV),

(e) $P(t, D)$ *satisfies the property* (I),

(f) *the cost function is given by*

$$\mathcal{J}(v) = \|y(v) - z_d\|_*^2 + (Nv, v)_V, \tag{7}$$

where

$$\|y\|_*^2 = \|y\|_{V,p+G,k}^2 + \|y\|_{H;p,k}^2, \tag{8}$$

z_d is an observation of the state of the system and $y(v)$ the solution of (6), *then there exists an optimal control $v \in V_{ad}$ of the system (V_{ad} is a closed convex set in V).*

Proof. From Theorem 3.7 of [2, p. 119] it follows that there exists

(a′) a positive function $G(t)$, $G(t) \in C^1$, for every real t and

(b′) for every $k \in Z$ the positive functions $g_k(t) \in C^0$ and $G_k(t) \in C^1$ are such that if $p'(t) \geqq g_k$, for all $t \in R$, and $p + G$ verifies the condition (A), then for each $g \in \mathcal{D}^k(p; H)$ there exists a unique solution $y \in \mathcal{D}^k(p + G; V) \cap \mathcal{D}^k(p; H)$ of the equation

$$(A(t)y, u)_V + (P(t, D)y, u)_H = (g, u)_H \quad \text{for all } u \in V.$$

Let us consider $v \in V$. Then $\mathcal{B}v \in \mathcal{D}^k(p; H)$ and from (b) we obtain $g + \mathcal{B}v \in \mathcal{D}^k(p; H)$ for every $v \in V$.

Hence the equation

$$(A(t)y, u)_V + (P(t, D)y, u)_H = (g + \mathcal{B}v, u)_H, \qquad u \in V,$$

admits in $\mathcal{D}^k(p + G; V) \cap \mathcal{D}^k(p; H)$ a unique solution $y = y(t, v)$ for each control v which describes the state of the system at time t.

Let us consider the observation of the system as $z(u) = \mathcal{C}y(v)$, where $\mathcal{C}$ is the "observation" operator

$$\mathcal{C} : \mathcal{D}^k(p + G; V) \cap \mathcal{D}^k(p; H) \to \mathcal{H}.$$

$\mathcal{H}$ is a Hilbert space, and the cost function is defined by

$$\mathcal{J}(v) = \|\mathcal{C}y(v) - z_d\|_{\mathcal{H}}^2 + (Nv, v)_V.$$

If we assume that $\mathscr{C}$ is the canonical injection

$$i : \mathscr{D}^k(p+G;\, V)\cap\mathscr{D}^k(p;\, H)\to\mathscr{D}^k(p+G;\, V)\cap\mathscr{D}^k(p;\, H),$$

then the cost function is

$$\mathscr{T}(v)=\|y(v)-z_d\|_*^2+(Nv,\, v)_V,$$

where z_d is the given observation $z_d\in\mathscr{D}^k(p+G;\, V)\cap\mathscr{D}^k(p;\, H)$ and $\|\cdot\|_*$ the norm defined by (8).

Consequently we have

$$\Pi(u,\, v)=(y(u)-y(0),\, y(v)-y(0))_*+(Nu,\, v)_V,$$
$$L(v)=(z_d-y(0),\, y(v)-y(0))_*,$$

and then

$$\mathscr{T}(v)=\Pi(v,\, v)-2L(v)+\|z_d-y(0)\|_*^2.$$

The coerciveness of N implies the coerciveness of Π. Then, by the existence and uniqueness theorem it follows that there exists a $u\in V_{ad}$, V_{ad} a closed convex subset of V, which minimizes $\mathscr{T}(v)$ and hence realizes the optimal control.

We know that u is an optimal control if and only if

$$(y(u)-z_d,\, y(v)-y(u))_*+(Nu,\, v-u)_V\geqq 0 \quad \text{for all } v\in V_{ad} \tag{9}$$

which may be written

$$\int_R (e^{-(p+G)}D^k(y(u)-z_d),\, e^{-(p+G)}D^k(y(v)-y(u)))_V\, dt$$
$$+\int_R (e^{-p}D^k(y(u)-z_d),\, e^{-p}D^k(y(v)-y(u)))_H\, dt+(Nu,\, v-u)_V\geqq 0$$

or

$$\int_R (D^k(e^{-2(p+G)}D^k(y(u)-z_d)),\, y(v)-y(u))_V$$
$$+\int_R (D^k(e^{-2p}D^k(y(u)-z_d)),\, y(v)-y(u))_H\, dt+(Nu,\, v-u)_V\geqq 0. \tag{10}$$

We now transform this expression by means of the adjoint state. For each control $v\in V_{ad}$ we define the adjoint states $c_1(v)$, $c_2(v)\in\mathscr{D}^k(p+G;\, V)\cap\mathscr{D}^k(p;\, H)$ as being the solutions of the equations

$$[P(t,\, D)+A(t)]^*c_1(v)=D^k(e^{-2(p+G)}D^k(y(v)-z_d)) \tag{11}$$

and

$$[P(t, D)+A(t)]^* c_2(v) = D^k(e^{-2p} D^k(y(v)-z_d)). \tag{12}$$

Consequently,

$$\begin{aligned}(D^k(e^{-2(p+G)}D^k(y(u)-z_d)),\ y(v)-y(u))_V \\ = ([P(t, D) + A(t)]^* c_1(u),\ y(v)-y(u))_V \\ = (c_1(u),\ [P(t, D)+A(t)](y(v)-y(u))) = (c_1(u),\ B(v-u)) \\ = (\mathcal{B}^* c_1(u),\ v-u)_V,\end{aligned}$$

where $\mathcal{B}^*$ is the adjoint of $\mathcal{B}$ and

$$\begin{aligned}(D^k(e^{-p}D^k(y(u)-z_d)),\ y(v)-y(u))_H = ([P(t, D) + A(t)]^* c_2(u),\ y(v)-y(u))_H \\ = (c_2(u), [P(t, D)+A(t)](y(v)-y(u)))_H = (c_2(u),\ \mathcal{B}(v-u)) \\ = (\mathcal{B}^* c_1(u),\ v-u)_H.\end{aligned}$$

Then we obtain from (10), (11) and (12),

$$\int_R (\mathcal{B}^* c_1(u),\ v-u)_V\, dt + \int_R (\mathcal{B}^* c_2(u),\ v-u)_H\, dt + (Nu,\ v-u)_V \geqq 0 \tag{13}$$

for every $v \in V_{ad}$ and $\mathcal{B}^* \in \mathcal{L}(\mathcal{D}'^k(p; H), V') = \mathcal{L}(\mathcal{D}'^k(-p; H), V')$. Hence the optimal control u is characterized by the inequality (13), where $c_1(u)$ and $c_2(u)$ are given by (11) and (12).

3.2. Let us consider the integro-differential operator

$$P(t, D) = \sum_{r=-n}^{2} \mathcal{B}_r(t) D^r, \tag{14}$$

where $n \in N$, $\mathcal{B}_r(t) \in \mathcal{E}_t(\mathcal{L}(H, H))$.

Suppose that $\mathcal{B}_2(t)$ is Hermitian and satisfies condition (I), in other words, that there exists a function $b(t) \in \mathcal{E}$ such that for every $g \in H$ and t real,

$$(\mathcal{B}_2(t)g,\ g)_H \geqq b(t)\|g\|_H. \tag{15}$$

Let V be the space of controls and let us consider a system governed by the operator (14) such that the state y of the system is given by the solution of the equation

$$(A(t)y,\ u)_V + (P(t, D)y,\ u)_H = (g + \mathcal{B}v,\ u)_H, \tag{16}$$

the sense of scalar distributions.

We have the following result.

PROPOSITION 3.2. *If: p is a real function that verifies the condition* (A) *and*

(a_1) $\mathcal{B} \in \mathcal{L}(V; \mathcal{D}^k(p; H))$,

(b_1) $g \in \mathcal{D}^k(p; H)$,

(c_1) $N \in \mathcal{L}(V, V)$ *is Hermitian and coercive,*

(d_1) $A(t)$ *satisfies the condition* (IV),

(e_1) $\mathcal{B}_2(t)$ *satisfies the inequality* (15), *then we can find a unique element* $u \in V_{ad} \subset V$ (V_{ad}—*a closed convex set of* V) *which minimizes the cost function*

$$\mathcal{J}(v) = \|y(v)-z_d\|_*^2 + (Nv,\ v)_V, \tag{17}$$

where y is the solution of (16), z_d *is an observation of the state of the system and*

$$\|y\|_*^2 = \|y\|_{v;p+G,k}^2 + \|y\|_{H;p,k+1}^2.$$

Proof. By virtue of a theorem of Trèves ([2, Thm. 3.7]) there exists:

(a″) a function $G(t) \geqq 0$ of class C^1 and

(b″) for every $k \in Z$, positive functions $g_k(t) \in C^0$ and $G_k(t) \in C^1$ such that if $p' \geqq g_k$ and $p+G$ verifies the condition (A), then for each $g \in \mathcal{D}^k(p; H)$ we can find a unique element $y \in \mathcal{D}^k(p+G; V) \cap \mathcal{D}^{k+1}(p; H)$ such that

$$(A(t)y, u)_V + (P(t, D)y, u)_H = (g, u)_H \quad \text{for all } u \in V.$$

From the hypothesis we deduce that $\mathcal{B} \in \mathcal{L}(V; \mathcal{D}^k(p; H))$ and consequently, $g + \mathcal{B}v \in \mathcal{D}^k(p; H)$. Hence we have a unique solution $y = y(t, v)$ of the equation

$$(A(t)y, u)_V + (P(t, D)y, u)_H = (g + \mathcal{B}v, u)_H$$

that represents the state of the system at time t for each control.

Let us suppose that the "observation" operator is the canonical injection

$$i : \mathcal{D}^k(p+G; V) \cap \mathcal{D}^{k+1}(p; H) \to \mathcal{D}^k(p+G; V) \cap \mathcal{D}^{k+1}(p; H)$$

and the cost function is

$$\mathcal{J}(v) = \|y(v) - z_d\|_*^2 + (Nv, v). \tag{18}$$

Then it follows from the same theorem that there exists a unique element $u \in V_{ad} \subset V$ which minimizes $\mathcal{J}(v)$.

The optimal control is characterized by the inequality

$$(y(u) - z_d, y(v) - y(u))_* + (Nu, v-u)_V \geqq 0 \quad \text{for all } v \in V_{ad} \tag{19}$$

or by

$$\int_R \bar{e}^{(p+G)} D^k(y(u) - z_d), e^{-(p+G)} D^k(y(v) - y(u))_V \, dt$$

$$+ \int_R (e^{-p} D^{k+1}(y(u) - z_d), e^{-p} D^{k+1}(y(v) - y(u)))_H \, dt + (Nu, v-u)_V \geqq 0$$

or

$$\int_R (D^k(e^{-2(p+G)} D^k(y(u) - z_d)), y(v) - y(u))_V \, dt$$

$$+ \int_R (D^{k+1}(e^{-2p} D^{k+1}(y(u) - z_d)), y(v) - y(u))_H \, dt + (Nu, v-u)_V \geqq 0.$$

Let us introduce the adjoint states

$$d_1 : V_{ad} \to \mathcal{D}^k(p+G; V) \cap \mathcal{D}^k(p; H),$$

$$d_2 : V_{ad} \to \mathcal{D}^k(p+G; V) \cap \mathcal{D}^{k+1}(p; H)$$

as solutions of the equations

$$[P(t, D)+A(t)]^* d_1(v) = D^k(e^{-2(p+G)}D^k(y(v)-z_d)),$$
$$[P(t, D)+A(t)]^* d_2(v) = D^{k+1}(e^{-2p}D^{k+1}(y(v)-z_d)).$$

Then we obtain

$$\int_R (d_1(u), \mathcal{B}(v-u))\, dt + \int_R (d_2(u), \mathcal{B}(v-u))\, dt + (Nu, v-u)_V \geqq 0$$

and

$$\int_R (\mathcal{B}^* d_1(u), v-u)\, dt + \int_R (\mathcal{B}^* d_2(u), v-u)_H\, dt + (Nu, v-u)_V \geqq 0,$$

where

$$\mathcal{B}^* \in \mathcal{L}(\mathcal{D}'^k(p; H); V') = \mathcal{L}(\mathcal{D}^{-k}(-p; H), V').$$

Acknowledgments. I would like to thank Prof. A. V. Balakrishnan for many useful remarks and for his constant encouragement. I would like to thank Prof. H. Fattorini for having the amiability to read the manuscript.

REFERENCES

[1] F. TRÈVES, *Domination et problèmes aux limites de type mixte*, C. R. Acad. Sci. Paris, 245 (1957), pp. 2454–2457.

[2] ———, *Relation de domination entre opérateurs différentiels*, Acta Math., 101 (1959), pp. 1–139.

[3] J. L. LIONS, *Contrôle optimale des systèmes gòuvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

ON SOME PROPERTIES OF MIN-MAX FUNCTIONS*

T. MATSUMOTO†

Abstract. Some of the properties of functions resulting from min-max operations are discussed. First, an implicit function theorem involving min-max functions is proved. Then a formula for the directional derivatives of the implicit function is given. It is shown that these results can be successfully applied to some of the problems in differential games.

1. Introduction. During the course of a study in differential games, the author was led to functions of the form

$$H(z, \tau) = \min_{y \in Y} \max_{x \in X} F(x, y, z, \tau).$$

We will discuss some of the properties of such functions. We will first prove an implicit function theorem for the equation

$$H(z, \tau) = 0. \tag{1.1}$$

Namely, we will give conditions under which (1.1) uniquely (locally) defines a function $\tau(z)$. We will then give a formula for calculating the directional derivative of $\tau(z)$ in a direction g:

$$\lim_{\substack{\alpha \to 0 \\ \alpha > 0}} \frac{\tau(z + \alpha g) - \tau(z)}{\alpha} = D_g \tau(z).$$

Finally, we will show how these results can be applied to some of the problems in differential games.

2. Results. We first state a known result on the directional differentiability of min-max functions.

THEOREM 1. *Let (X, d_X) and (Y, d_Y) be compact metric spaces, and let Z be a normed linear space. Let $F(x, y, z)$ be a real-valued function continuous on $X \times Y \times Z$, and let*

$$G(y, z) = \max_{x \in X} F(x, y, z), \qquad M(y, z) = \{x \in X | F(x, y, z) = G(y, z)\},$$

$$H(z) = \min_{y \in Y} G(y, z), \qquad N(z) = \{y \in Y | G(y, z) = H(z)\}.$$

Suppose that the following hold:

(i) *At a certain point $z_0 \in Z$ and for a certain $g \in Z$, the derivative*

$$\frac{\partial}{\partial \nu} F(x, y, z_0 + \nu g)$$

exists and is continuous in (x, y, ν) on $X \times Y \times [0, \alpha]$, $\alpha > 0$.

* Received by the editors January 15, 1974, and in revised form July 4, 1974.
† Department of Electrical Engineering, Waseda University, Tokyo 160, Japan.

(ii) *$M(y, z_0)$ is lower semicontinuous in y with respect to inclusion, i.e., at each y_0 and for every $\varepsilon > 0$ there is a $\delta > 0$ such that $d_Y(y_0, y) < \delta$ implies*

$$M(y_0, z_0) \subset U(M(y, z_0); \varepsilon) = \left\{ x \in X \mid \inf_{x_0 \in M(y, z_0)} d_X(x_0, x) < \varepsilon \right\}.$$

Then at z_0, the function $H(z)$ is directionally differentiable in the direction g, and

$$D_g H(z_0) \equiv \lim_{\substack{\nu \to 0 \\ \nu > 0}} \frac{H(z_0 + \nu g) - H(z_0)}{\nu} \tag{2.1}$$

$$= \min_{y \in N(z_0)} \max_{x \in M(y, z_0)} \frac{\partial}{\partial g} F(x, y, z_0),$$

where

$$\frac{\partial}{\partial g} F(x, y, z_0) = \left[\frac{\partial}{\partial \nu} F(x, y, z_0 + \nu g) \right]_{\nu = 0}.$$

Proof of this fact can be found in [1]. Related results on min-max functions are found in [2]–[6].

Remark 1. Condition (ii) cannot, in general, be relaxed, as the following example shows. Let

$$F(x, y, z) = x(y + z), \qquad X = Y = [-1, 1].$$

Then

$$G(y, z) = |y + z|,$$

with

$$M(y, z) = \begin{cases} \{\operatorname{sgn}(y + z)\} & \text{if } y + z \neq 0, \\ [-1, 1] & \text{if } y + z = 0. \end{cases}$$

Hence $M(y, z)$ is only upper semicontinuous in y at $y = 0$ with $z = 0$. It is clear that for $|z| \leqq 1$,

$$H(z) = 0, \quad \text{with } N(z) = \{-z\}.$$

Clearly, then,

$$[D_g H(z_0)]_{z_0 = 0, g = 1} = 0.$$

On the other hand,

$$\left[\frac{\partial}{\partial g} F(x, y, z_0) \right]_{g = 1} = x,$$

so that

$$\left[\min_{y\in N(z_0)}\max_{x\in M(y,z_0)}\frac{\partial}{\partial g}F(x,y,z_0)\right]_{z_0=0,\,g=1}=\max_{x\in M(0,0)}x=\max_{x\in[-1,1]}x=1.$$

Remark 2. Even if condition (ii) does not hold, the following estimates are valid:

$$\liminf_{\substack{\nu\to 0\\ \nu>0}}\frac{H(z_0+\nu g)-H(z_0)}{\nu}\geqq\min_{y\in N(z_0)}\min_{x\in M(y,z_0)}\frac{\partial}{\partial g}F(x,y,z_0), \tag{2.2}$$

$$\limsup_{\substack{\nu\to 0\\ \nu>0}}\frac{H(z_0+\nu g)-H(z_0)}{\nu}\leqq\max_{y\in N(z_0)}\max_{x\in M(y,z_0)}\frac{\partial}{\partial g}F(x,y,z_0). \tag{2.3}$$

Remark 3. $M(y, z_0)$ is automatically upper semicontinuous in y with respect to inclusion, i.e., the inclusion

$$M(y,z_0)\subset U(M(y_0,z_0);\varepsilon)$$

holds. Hence, if condition (ii) holds, then $M(y, z_0)$ is continuous with respect to the Hausdorff metric in y. We will later show that if $M(y, z_0)$ is Hausdorff continuous in y, then the function

$$\max_{x\in M(y,z_0)}\frac{\partial}{\partial g}F(x,y,z_0) \tag{2.4}$$

is continuous in y (as a real-valued function), so that the formula (2.1) makes sense. It is clear that the function defined by (2.4) is upper semicontinuous in y (as a real-valued function), so that (2.3) makes sense. A similar statement is valid for (2.2).

Remark 4. Since $M(y, z_0)$ is always contained in a compact space X, upper semicontinuity and lower semicontinuity with respect to inclusion are equivalent to sequential upper semicontinuity and sequential lower semicontinuity, respectively. More precisely, $M(y, z_0)$ is sequentially upper semicontinuous in y at y_0 if

$$y_k\to y_0,\quad x_k\to x_0\quad\text{and}\quad x_k\in M(y_k,z_0)$$

imply

$$x_0\in M(y_0,z_0).$$

$M(y, z_0)$ is sequentially lower semicontinuous in y at y_0 if, for every $x_0\in M(y_0, z_0)$ and for every sequence $\{y_k\}$ with $y_k\to y_0$, there is a sequence $\{x_k\}$ with

$$x_k\in M(y_k,z_0)\quad\text{and}\quad x_k\to x_0.$$

COROLLARY. *Let condition* (i) *hold. If* $M(y, z_0)$ *is a singleton set* $\{m(y, z_0)\}$ *for all* y *in a neighborhood of* y_0, *then*

$$D_gH(z_0)=\min_{y\in N(z_0)}\frac{\partial}{\partial g}F(m(y,z_0),y,z_0). \tag{2.5}$$

Proof. If $M(y, z_0)$ is a singleton for all y in a neighborhood of y_0, then it is continuous in y. Hence condition (ii) is satisfied.

Remark. The fact that $N(z_0)$ is a singleton does not ensure that the formula is valid. For instance, in the example of Remark 1 after Theorem 1, the set $N(z_0)$ at $z_0 = 0$ is the singleton set $\{0\}$. But the formula (2.1) does not hold.

Now, consider a real-valued function $F(x, y, z, \tau)$ continuous on $X \times X \times V(z_0) \times W(\tau_0)$, where X and Y are compact metric spaces, $V(z_0)$ is a neighborhood of a point z_0 in a normed linear space Z, and $W(\tau_0)$ is a neighborhood of a point τ_0 of the real line. Let

$$G(y, z, \tau) = \max_{x \in X} F(x, y, z, \tau),$$

$$M(y, z, \tau) = \{x \in X | F(x, y, z, \tau) = G(y, z, \tau)\},$$

$$H(z, \tau) = \min_{y \in Y} G(y, z, \tau),$$

$$N(z, \tau) = \{y \in Y | G(y, z, \tau) = H(z, \tau)\}.$$

We will prove an implicit function theorem for

$$H(z, \tau) = 0.$$

THEOREM 2. *Suppose that*

$$H(z_0, \tau_0) = 0,$$

and that the following hold:

(i) *The derivative*

$$\frac{\partial}{\partial \tau} F(x, y, z, \tau)$$

exists and is continuous on $X \times Y \times V(z_0) \times W(\tau_0)$.

(ii) $M(y, z, \tau)$ *is lower semicontinuous in* y *with respect to inclusion for each* $(z, \tau) \in V(z_0) \times W(\tau_0)$.

(iii) $$\min_{y \in N(z_0, \tau_0)} \min_{x \in M(y, z_0, \tau_0)} \frac{\partial}{\partial \tau} F(x, y, z_0, \tau_0) > 0.$$

Then there is a neighborhood $V'(z_0) \subset V(z_0)$ *of* z_0 *and a continuous real-valued function* $\tau(z)$ *on* $V'(z_0)$ *satisfying*

$$H(z, \tau(z)) = 0.$$

Proof. It follows from Theorem 1 and conditions (ii) and (iii) that

$$D_r H(z_0, \tau_0) \equiv \lim_{\substack{\alpha \to 0 \\ \alpha > 0}} \frac{H(z_0, \tau_0 + \alpha) - H(z_0, \tau_0)}{\alpha}$$

$$= \min_{y \in N(z_0, \tau_0)} \max_{x \in M(y, z_0, \tau_0)} \frac{\partial}{\partial \tau} F(x, y, z_0, \tau_0) > 0,$$

$$D_l H(z_0, \tau_0) \equiv \lim_{\substack{\alpha \to 0 \\ \alpha < 0}} \frac{H(z_0+\alpha) - H(z_0, \tau_0)}{\alpha}$$

$$= (-1) \min_{y \in N(z_0, \tau_0)} \max_{x \in M(y, z_0, \tau_0)} (-1) \frac{\partial}{\partial \tau} F(x, y, z_0, \tau_0)$$

$$= \max_{y \in N(z_0, \tau_0)} \min_{x \in M(y, z_0, \tau_0)} \frac{\partial}{\partial \tau} F(x, y, z_0, \tau_0) > 0.$$

It follows from condition (i) that the function

$$\min_{y \in N(z, \tau)} \min_{x \in M(y, z, \tau)} \frac{\partial}{\partial \tau} F(x, y, z, \tau) \tag{2.6}$$

is lower semicontinuous in (z, τ). Hence there are neighborhoods $V'(z_0) \subset V(z_0)$ and $W'(\tau_0) \subset W(\tau_0)$ such that the function defined by (2.6) is positive on $V'(z_0) \times W'(\tau_0)$. Clearly, then, $D_r H(z, \tau)$ and $D_l H(z, \tau)$ are also positive on $V'(z_0) \times W'(\tau_0)$. Hence, for each z in $V'(z_0)$, $H(z, \tau)$ is monotonically increasing with respect to τ on $W'(\tau_0)$, so that there is a unique $\tau(z)$ satisfying

$$H(z, \tau(z)) = 0.$$

In order to prove continuity of $\tau(z)$, let $\{z_k\} \subset V'(z_0)$ be convergent to z, and let $\{\tau(z_k)\}$ be the corresponding sequence:

$$H(z_k, \tau(z_k)) = 0, \qquad k = 1, 2, \cdots.$$

Since $\{\tau(z_k)\}$ is bounded, there is a subsequence $\{\tau(z_{k_j})\}$ convergent to a τ^*. Since $H(z, \tau)$ is jointly continuous,

$$0 = \lim_{j \to \infty} H(z_{k_j}, \tau(z_{k_j})) = H(z, \tau^*).$$

But, since $\tau(z)$ is unique, we have

$$\tau^* = \tau(z).$$

Thus $\{\tau(z_k)\}$ has a unique accumulation point $\tau(z)$, and hence it is, in fact, the limit point of $\{\tau(z_k)\}$.

We next give a formula for the directional derivatives of the function $\tau(z)$.

THEOREM 3. *Let conditions* (i)–(iii) *of Theorem* 2 *hold. If, in addition*:

(iv) $\partial / \partial \nu \, F(x, y, z_0 + \nu g, \tau)$ *exists and is continuous on* $X \times Y \times [0, \alpha] \times W(\tau_0)$, $\alpha > 0$,

then $\tau(z)$ *is directionally differentiable at* z_0 *in the direction* g, *and*

$$D_g \tau(z_0) = \max_{y \in N(z_0, \tau(z_0))} \min_{x \in M(y, z_0, \tau(z_0))} - \frac{\partial F(x, y, z_0, \tau(z_0))/\partial g}{\partial F(x, y, z_0, \tau(z_0))/\partial \tau}. \tag{2.7}$$

Proof. By Theorem 2, $\tau(z_0+\nu g)$ is uniquely defined and continuous for sufficiently small ν. It follows from the mean-value theorem that

$$F(x, y, z_0+\nu g, \tau(z_0+\nu g)) = F(x, y, z_0, \tau(z_0)) + \nu\frac{\partial}{\partial \nu}F(x, y, z', \tau') + (\tau(z_0+\nu g)-\tau(z_0))\frac{\partial}{\partial \tau}F(x, y, z', \tau'), \tag{2.8}$$

where $z' = z_0+\theta\nu g$, $\tau' = \tau(z_0)+\theta^*(\tau(z_0+\nu g)-\tau(z_0))$, with $0<\theta, \theta^*<1$. For every $y \in Y$ and $x \in M(y, z_0, \tau(z_0))$,

$$G(y, z_0+\nu g, \tau(z_0+\nu g)) \geqq G(y, z_0, \tau(z_0)) + \nu\frac{\partial}{\partial \nu}F(x, y, z', \tau') + (\tau(z_0+\nu g)-\tau(z_0))\frac{\partial}{\partial \tau}F(x, y, z', \tau'). \tag{2.9}$$

In particular, for $y \in N(z_0+\nu g, \tau(z_0+\nu g))$ and $x \in M(y, z_0, \tau(z_0))$,

$$\begin{aligned} 0 = H(z_0+\nu g, \tau(z_0+\nu g)) &= G(y, z_0+\nu g, \tau(z_0+\nu g)) \\ &\geqq G(y, z_0, \tau(z_0)) + \nu\frac{\partial}{\partial \nu}F(x, y, z', \tau') \\ &\quad + (\tau(z_0+\nu g)-\tau(z_0))\frac{\partial}{\partial \tau}F(x, y, z', \tau') \\ &\geqq \nu\frac{\partial}{\partial \nu}F(x, y, z', \tau') \\ &\quad + (\tau(z_0+\nu g)-\tau(z_0))\frac{\partial}{\partial \tau}F(x, y, z', \tau'), \end{aligned} \tag{2.10}$$

where we have used the fact that

$$G(y, z_0, \tau(z_0)) \geqq \min_{y \in Y} G(y, z_0, \tau(z_0)) = 0.$$

If condition (iii) is satisfied, for every $y \in N(z_0, \tau(z_0))$ and every $x \in M(y, z_0, \tau(z_0))$,

$$\frac{\partial}{\partial \tau}F(x, y, z_0, \tau(z_0)) > 0.$$

Since $N(z, \tau)$ is upper semicontinuous in (z, τ) with respect to inclusion, $M(y, z, \tau)$ is upper semicontinuous in (y, z, τ) with respect to inclusion, and since $\tau(z)$ is continuous, we have that for all sufficiently small $\nu > 0$,

$$\frac{\partial}{\partial \tau}F(x, y, z', \tau') > 0, \tag{2.11}$$

where $y \in N(z_0+\nu g, \tau(z_0+\nu g))$ and $x \in M(y, z_0, \tau(z_0))$. Now, by a proposition

which will be given shortly, the function

$$\min_{x\in M(y,z_0,\tau(z_0))} \frac{\partial}{\partial\tau}F(x, y, z_0, \tau(z_0))$$

is continuous in y. It then follows from (2.10) and (2.11) that for $\nu>0$ sufficiently small,

$$\frac{\tau(z_0+\nu g)-\tau(z_0)}{\nu}\leqq \max_{y\in N(z_0+\nu g,\tau(z_0+\nu g))}\ \min_{x\in M(y,z_0,\tau(z_0))} -\frac{\partial F(x, y, z', \tau')/\partial\nu}{\partial F(x, y, z', \tau')/\partial\tau}.$$

It follows from condition (ii) and a proposition below that the right-hand side of the above inequality is upper semicontinuous in ν at $\nu=0$, so that

$$\limsup_{\substack{\nu\to 0\\ \nu>0}} \frac{\tau(z_0+\nu g)-\tau(z_0)}{\nu}\leqq \max_{y\in N(z_0,\tau(z_0))}\ \min_{x\in M(y,z_0,\tau(z_0))} -\frac{\partial F(x, y, z_0, \tau(z_0))/\partial g}{\partial F(x, y, z_0, \tau(z_0))/\partial\tau}. \tag{2.12}$$

On the other hand, it follows from (2.8) that for every $y\in Y$ and every $x\in M(y, z_0+\nu g, \tau(z_0+\nu g))$,

$$\begin{aligned} G(y, z_0+\nu g, \tau(z_0+\nu g)) &= F(x, y, z_0, \tau(z_0))+\nu\frac{\partial}{\partial\nu}F(x, y, z', \tau')\\ &\quad +(\tau(z_0+\nu g)-\tau(z_0))\frac{\partial}{\partial\tau}F(x, y, z', \tau')\\ &\leqq G(y, z_0, \tau(z_0))+\nu\frac{\partial}{\partial\nu}F(x, y, z', \tau')\\ &\quad +(\tau(z_0+\nu g)-\tau(z_0))\frac{\partial}{\partial\tau}F(x, y, z', \tau'). \end{aligned}$$

In particular, for $y\in N(z_0, \tau(z_0))$ and $x\in M(y, z_0+\nu g, \tau(z_0+\nu g))$,

$$\begin{aligned} H(z_0+\nu g, \tau(z_0+\nu g))\leqq H(z_0, \tau(z_0))&+\nu\frac{\partial}{\partial\nu}F(x, y, z', \tau')\\ &+(\tau(z_0+\nu g)-\tau(z_0))\frac{\partial}{\partial\tau}F(x, y, z', \tau'). \end{aligned}$$

Since, for ν sufficiently small,

$$H(z_0+\nu g, \tau(z_0+\nu g)) = H(z_0, \tau(z_0)) = 0,$$

we have, for $\nu>0$ sufficiently small,

$$\frac{\tau(z_0+\nu g)-\tau(z_0)}{\nu}\geqq \min_{x\in M(y,z_0+\nu g,\tau(z_0+\nu g))} -\frac{\partial F(x, y, z', \tau')/\partial\nu}{\partial F(x, y, z', \tau')/\partial\tau} \tag{2.13}$$

for all $y\in N(z_0, \tau(z_0))$. Since $M(y, z, \tau)$ is upper semicontinuous in (z, τ), with respect to inclusion, (2.13) implies that

$$\liminf_{\substack{\nu\to 0\\ \nu>0}} \frac{\tau(z_0+\nu g)-\tau(z_0)}{\nu}\geqq \min_{x\in M(y,z_0,\tau(z_0))} -\frac{\partial F(x, y, z_0, \tau(z_0))/\partial g}{\partial F(x, y, z_0, \tau(z_0))/\partial\tau}$$

for every $y \in N(z_0, \tau(z_0))$. This, together with (2.12), implies the desired formula (2.7).

The following fact is proved in [1]. Since it plays an important role throughout the paper, we will give a proof which is different from that of [1].

PROPOSITION. *Let $f(x, y)$ be a real-valued function continuous on $X \times Y$, where X is a compact metric space and Y is a metric space. Let $M(y)$ be a compact set-valued function which is always contained in X. If $M(y)$ is continuous with respect to the Hausdorff metric, then the function*

$$\max_{x \in M(y)} f(x, y) \tag{2.14}$$

is continuous. Similarly,

$$\min_{x \in M(y)} f(x, y)$$

is continuous.

Proof. We will prove only the first statement. The proof of the second is similar. It is clear that the function defined by (2.14) is upper semicontinuous. In order to prove the lower semicontinuity, let $\{y_k\}$ be an arbitrary sequence with $y_k \to y_0$. Let x_0 be an element of $M(y_0)$ such that

$$f(x_0, y_0) = \max_{x \in M(y_0)} f(x, y_0)$$

holds. Since $M(y)$ is continuous and hence lower semicontinuous, there is a sequence $\{x_k\}$ such that $x_k \in M(y_k)$ and $x_k \to x_0$ (see Remark 3 after Theorem 1). It is clear that

$$f(x_k, y_k) \to f(x_0, y_0),$$

so that for a given $\varepsilon > 0$, there is an integer K such that $k \geqq K$ implies that

$$|f(x_0, y_0) - f(x_k, y_k)| < \frac{\varepsilon}{2}.$$

Hence, for $k \geqq K$,

$$\begin{aligned} \max_{x \in M(y_0)} f(x, y_0) = f(x_0, y_0) &< f(x_k, y_k) + \varepsilon \\ &\leqq \max_{x \in M(y_k)} f(x, y_k) + \varepsilon. \end{aligned}$$

Since $\{y_k\}$ and $\varepsilon > 0$ are arbitrary, we conclude that

$$\max_{x \in M(y_0)} f(x, y_0) \leqq \liminf_{y \to y_0} \max_{x \in M(y)} f(x, y).$$

3. Applications. In this section, we will show how the results of the previous section can be applied to some of the problems in differential games. Consider the linear system

$$\dot{z} = Az + Bu + Cv, \tag{3.1}$$

where z in R^n is the state, u in $U \subset R^r$ is the control of controller I, and v in $V \subset R^s$ is the control of controller II. A, B and C are constant matrices with appropriate dimensions. Let π be the operator of orthogonal projection of R^n onto its linear subspace R^x, and let

$$T = \{z \in R^n \mid \|\pi z\| \leqq \gamma\},$$

where γ is a nonnegative real number. The objective of controller I is to steer an initial state z_0 of (3.1) to a point of T, whereas the objective of controller II is to prevent that from occurring. This problem is called the pursuit-evasion problem. T is called the target set.

We will first look at the problem from controller II's point of view. Let

$$B_x = \{\psi \in R^x \mid \|\psi\| = 1\},$$

and for ψ in B_x, set

$$\varphi(\psi, s) = \max_{v \in V} \min_{u \in U} \langle \psi, \pi e^{As}(Bu + Cv) \rangle,$$

where e^{As} is the fundamental matrix of (3.1). Let

$$F(\psi, \tau, z) = \langle \psi, \pi e^{A\tau} z \rangle + \int_0^\tau \varphi(\psi, s)\, ds, \tag{3.2}$$

and assume that the following hold:

(i) $\varphi(\psi, s)$ is independent of ψ (written as $\varphi(s)$);

(ii) there is a positive number θ such that

$$\int_0^\theta \varphi(s)\, ds > \gamma.$$

Condition (i) is very strong, but it is satisfied for several nontrivial examples [3], [7]–[11]. Condition (ii) is a natural one for controller II to accomplish his objective. Let

$$H(z) = \min_{\tau \in [0, \theta]} \max_{\psi \in B_x} F(\psi, \tau, z),$$

which is still continuous. Controller II can prevent $z \in T$ from occurring, as long as $H(z) > \gamma$. Hence, if controller II can choose a control in such a way that

$$D_+ H(z(t)) \equiv \lim_{\substack{\alpha \to 0 \\ \alpha > 0}} \frac{H(z(t+\alpha)) - H(z(t))}{\alpha} \geqq 0, \tag{3.3}$$

$$H(z(0)) > \gamma$$

hold, then he can prevent $z \in T$ from occurring. Theorem 1 can be used to calculate the right derivative (3.3). Let

$$G(\tau, z) = \max_{\psi \in B_x} F(\psi, \tau, z),$$

$$M(\tau, z) = \{\psi \in B_x \mid F(\psi, \tau, z) = G(\tau, z)\},$$

$$N(z) = \{\tau \in [0, \theta] \mid G(\tau, z) = H(z)\}.$$

It follows from condition (i) that

$$M(\tau, z) = \left\{ \frac{\pi e^{A\tau} z}{\|\pi e^{A\tau} z\|} \right\} \equiv \{\psi(\tau, z)\},$$

as long as $\pi e^{A\tau} z \neq 0$. Hence $M(\tau, z)$ is continuous in τ if $\pi e^{A\tau} z \neq 0$ for $\tau \in [0, \theta]$. If the admissible controls are right continuous, then

$$\lim_{\substack{\alpha \to 0 \\ \alpha > 0}} \frac{z(t+\alpha) - z(t)}{\alpha} = Az(t) + Bu(t) + Cv(t),$$

so that Theorem 1 is applicable. The right derivative is given by

$$D_r H(z(t)) = \min_{\tau \in N(z(t))} \langle \psi(\tau, z), \pi e^{A\tau}(Az(t) + Bu(t) + Cv(t)) \rangle.$$

Details can be found in [11].

We will next consider the problem from controller I's point of view. Let

$$\dot{z} = A(t)z + B(t)u + C(t)v \tag{3.4}$$

be the dynamics, where $A(t)$, $B(t)$ and $C(t)$ are matrices with the same dimensions as before with their components continuous. In order to analyze the problem, we distinguish two information patterns, i.e., the information available to each controller. The information pattern is called open-loop if controller II first chooses $v(\cdot)$ on $[t_0, \infty)$, and controller I then chooses $u(\cdot)$ on $[t_0, \infty)$ knowing the $v(\cdot)$. The information pattern is called closed-loop if, at time t, controller II chooses $v(t)$ knowing $(z(t), t)$, and controller I chooses $u(t)$ knowing $(z(t), t, v(t))$. It is, in general, extremely difficult to solve the closed problem directly so that one sometimes makes use of the properties of the open-loop problem. The open-loop problem is easier to analyze, since it can be characterized by support functions of various sets.

Let $\mathcal{U}(\mathcal{V})$ be the class of all open-loop admissible controls of controller I (II), and let

$$F(\psi, u(\cdot), v(\cdot), t, \tau) = \langle \psi, \pi\Phi(t+\tau, t)z(t) \rangle + \int_t^{t+\tau} \langle \psi, \pi\Phi(t+\tau)(B(s)u(s) + C(s)v(s)) \rangle \, ds,$$

where $u(\cdot) \in \mathcal{U}$, $v(\cdot) \in \mathcal{V}$, and $\Phi(\cdot, \cdot)$ is the fundamental matrix of (3.4). If $\mathcal{U}$ ($\mathcal{V}$) is the class of all measurable functions $u(\cdot)$ ($v(\cdot)$) such that $u(t) \in U$ ($v(t) \in V$), and if U (V) is compact and convex, then $\mathcal{U}$ ($\mathcal{V}$) is weakly compact. Since $F(\psi, u(\cdot), v(\cdot), t, \tau)$ is linear in $(u(\cdot), v(\cdot))$, it is weakly continuous. Hence

$$H(t, \tau) = \max_{\psi \in B_X} \max_{v(\cdot) \in \mathcal{V}} \min_{u(\cdot) \in \mathcal{U}} F(\psi, u(\cdot), v(\cdot), t, \tau)$$

makes sense and is continuous. It is clear that if $H(t, \tau) \leqq \gamma$, then $z(t)$ can be

steered to T at time $t+\tau$ no matter what admissible (open-loop) control is chosen by controller II. Let $\tau(t)$ be the smallest value of τ satisfying

$$H(t, \tau) = \gamma.$$

If controller I can choose an admissible closed-loop control in such a way that $\tau(t)$ is monotonically decreasing, then he can steer the initial state to T under the closed-loop information pattern, since $\tau(t) = 0$ if and only if $z(t) \in T$. In order to do this, the right derivative

$$D_r\tau(t)$$

is again important. This can be calculated by Theorem 3 under certain conditions.

Let

$$G(\psi, v(\cdot), t, \tau) = \min_{u(\cdot)\in\mathcal{U}} F(\psi, u(\cdot), v(\cdot), t. \tau),$$

$$G'(\psi, t, \tau) = \max_{v(\cdot)\in\mathcal{V}} G(\psi, v(\cdot), t, \tau),$$

$$M(\psi, v(\cdot), t, \tau) = \{u(\cdot) \in \mathcal{U} | F(\psi, u(\cdot), v(\cdot), t, \tau) = G(\psi, v(\cdot), t, \tau)\},$$

$$N(\psi, t, \tau) = \{v(\cdot) \in \mathcal{V} | G(\psi, v(\cdot), t, \tau) = G'(\psi, t, \tau)\}.$$

It is clear that M is independent of $v(\cdot)$, so that it can be written as $M(\psi, t, \tau)$. If upper semicontinuity and lower semicontinuity of $M(\psi, t, \tau)$ and $N(\psi, t, \tau)$ are understood in the sense of weak convergence, everything remains valid. Now an important assumption of Theorem 2 and Theorem 3 was the continuity of $M(\psi, t, \tau)$ and $N(\psi, t, \tau)$ with respect to ψ. This condition is satisfied if one assumes the following:

(iii) For each ψ in B_x and for each interval $[t, t+\tau]$, the formula

$$\min_{u(s)\in U} \langle \psi, \pi\Phi(t+\tau, s)B(s)u(s)\rangle = \langle \psi, \pi\Phi(t+\tau, s)B(s)u(s; \psi, t, \tau)\rangle$$

uniquely defines $u(s; \psi, t, \tau)$, except for at most a finite number of points of $[t, t+\tau]$.

A similar condition is assumed for controller II. Under condition (iii), $u(s; \psi, t, \tau)$ is piecewise continuous in s on $[t, t+\tau]$, so that $u(\cdot; \psi, t, \tau)$ belongs to $\mathcal{U}$. More important is the fact that $M(\psi, t, \tau)$ is a singleton:

$$M(\psi, t, \tau) = \{u(\cdot; \psi, t, \tau)\}.$$

Hence the continuity condition is satisfied. A similar statement is valid for controller II. Condition (iii) is satisfied for many nontrivial examples [3], [7]–[11].

Another assumption of Theorem 3 was the differentiability of $F(\psi, u(\cdot), v(\cdot), t, \tau)$ with respect to τ. By a slight modification of the proof of Theorem 3, one can see that the differentiability of

$$F(\psi, u(\cdot; \psi, t', \tau'), v(\cdot; \psi, t'', \tau''), t, \tau) \tag{3.5}$$

with respect to τ for each (t', τ') and (t'', τ'') in a neighborhood of (t, τ) will suffice. Under condition (iii) the function defined by (3.5) is differentiable in τ. If the closed-loop admissible controls are right continuous, then $D_\tau\tau(t)$ can be explicitly calculated (see [10] for details).

Acknowledgments. The author is indebted to Prof. Y. Ishizuka for his help during the course of this work. He is also grateful to the reviewer for his valuable comments.

REFERENCES

[1] V. MALOZJOMOV, *On an extremal problem*, Cybernetics, 4 (1970), pp. 111–114.
[2] J. DANSKIN, *The theory of max–min, with applications*, SIAM J. Appl. Math., 14 (1966), pp. 641–664.
[3] A. FRIEDMAN, *Differential Games*, John Wiley, New York, 1971.
[4] V. DEMJANOV AND A. RUBINOV, *Approximate Methods in Extremal Problems*, American Elsevier, New York, 1970.
[5] V. DEMJANOV, *Differentiability of min–max functions*, Soviet Comp. Math. and Math. Phys., 8 (1968), pp. 1186–1195.
[6] B. PSHENICHNYI, *Necessary Conditions for an Extremum*, Dekker, 1972.
[7] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.
[8] L. PONTRYAGIN, *On the theory of differential games*, Russian Math. Surveys, 21 (1966), no. 4.
[9] B. PSHENICHNYI, *Linear differential games*, Automat. Remote Control (1968), pp. 55–67.
[10] T. MATSUMOTO AND E. SHIMEMURA, *On a class of linear pursuit games*, J. Differential Equations, 12 (1972), pp. 266–283.
[11] T. MATSUMOTO, *A class of linear evasion problems*, J. Optimization Theory and Applications, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

THE CLOSED-LOOP TIME OPTIMAL CONTROL. II: STABILITY*

PAVOL BRUNOVSKÝ†

Abstract. The problem, to which extent does the closed-loop time-optimal control of a linear system fulfill its task if the system is subject to small perturbations, is studied.

In part I of this paper (ref. [1]) we have studied the problem of optimality of the (Filippov) solutions of the linear system in R^n under the action of the closed-loop control v:

$$\dot{x} = Ax + v(x) \tag{1}$$

(i.e., of the solutions of the associated multivalued differential equation

$$\dot{x} \in Ax + V(x), \qquad V(x) = \bigcap_{\delta>0} \bigcap_{\mu(N)=0} \operatorname{co}\operatorname{cl} v((x+\delta D)\setminus N), \tag{2}$$

D the unit disc in R^n), which is obtained by synthesizing the open-loop time optimal control of the linear system

$$\dot{x} = Ax + u \tag{3}$$

with control constraint $u \in U$, under the assumptions that U is a convex polytope containing the origin in its relative interior and that the system is normal.

We now turn to the problem (mentioned in [1, § 1]) of stability. Namely, we ask how far the closed-loop control v, which was constructed for the unperturbed system (3), fulfills its task if the system is subject to small but permanently acting perturbations, i.e., whether, and in what time, the trajectories of the perturbed system under the action of the control v will approach the origin. Of course, one can expect positive results only if the closed-loop control v yields optimal solutions if no perturbations are present.

Let us note that the consideration of perturbations exhibits clearly the inadequacy of the classical definition of the solution. An excellent explanation can be found in [3] (the author was unaware of this paper while writing [1]).

The perturbations shall be modeled along the lines of [2] as measurable functions added to the right-hand side of (2), the magnitude of which does not exceed a given positive constant ε (note that the "measurement perturbations" of [3] are included). We shall be interested in estimates and results concerning the behavior of the system under any perturbation of this kind, i.e., results which take into account only the information about ε. Using the definition of solutions of multivalued differential equations, one sees that the Filippov solutions of the equation

$$\dot{x} = Ax + v(x) + p(t),$$

where $|p(t)| \leqq \varepsilon$ is measurable, are solutions of the multivalued differential

* Received by the editors July 20, 1973, and in revised form March 7, 1974.

† Mathematical Institute, Slovak Academy of Sciences, 886 25 Bratislava, Czechoslovakia. Now at Institute of Applied Mathematics, Comenius University, 816 31 Bratislava, Czechoslovakia.

equation

$$(4_\varepsilon) \qquad \dot{x} \in Ax + V(x) + \varepsilon D$$

and vice versa (cf. [2]). Therefore, we are led to study the solutions of (4_ε) and we shall formulate the results, summarized in the following two theorems, in terms of these solutions. Note that all solutions of (4_ε) are solutions of (4_η) for $\varepsilon \leqq \eta$.

Henceforth we shall assume that the system (3) with control domain U is normal and that U contains 0 in its relative interior without further notice. As in [1], we denote by $T(x)$ the minimal time in which x can be steered to 0, $\mathcal{R}(\tau) = \{x | T(x) \leqq \tau\}$ and $\mathcal{R} = \cup_{\tau \geqq 0} \mathcal{R}(\tau)$. We recall that under these assumptions and notations, the sets $\mathcal{R}(\tau)$ have Properties 1–3 of [1, § 2].

While the first one of the following two theorems does not require any knowledge of the structure of the synthesis and is valid for systems of any dimension, the proof of the second one is heavily based upon the study of the structure of the closed-loop time optimal control in two-dimensional systems from [1].

THEOREM 1. *Assume that all the solutions of* (2) *are optimal trajectories of* (3). *Then given* $\tau \in [0, \infty)$ *and* $\rho > 0$, *there exists an* $\varepsilon > 0$ *such that any solution* φ *of* (4_ε) *with* $\rho(0) \in \mathcal{R}(\tau)$ *satisfies* $|\varphi(t)| \leqq \rho$ *for* $t \geqq T(x)$.

THEOREM 2. *Let the assumption of Theorem* 1 *be satisfied and let* $\dim x = \dim U = 2$. *Then given* $\tau \in [0, \infty)$, *for sufficiently small* $\varepsilon > 0$, *there exists a function* $T_\varepsilon : \mathcal{R}(\tau) \to R$ *such that any solution* φ *of* (4_ε) *with* $\varphi(0) = x \in \mathcal{R}(\tau)$ *satisfies* $\varphi(t) = 0$ *for* $t \geqq T_\varepsilon(x)$ *and* $T_\varepsilon(x) \to T(x)$ *as* $\varepsilon \to 0$ *uniformly over* $\mathcal{R}(\tau)$.

Let us note that the Theorem of [1] gives necessary and sufficient conditions for two-dimensional systems to satisfy the assumption of Theorem 1. Further, let us note that Theorem 2 cannot be extended to the case $\dim U = 1$. This follows from [2], where the smallest neighborhood of the origin in which the system can be kept is constructed and shown not to be the origin in general.

For the proof of Theorem 1 we need two lemmas.

LEMMA 1. *Let* $\{\varepsilon_k\}$ *be a sequence of positive constants tending to* 0 *and let* $\{\varphi_k\}$ *be a sequence of trajectories of* (4_{ε_k}) *on* $[0, T]$ *such that* $\lim_{k \to \infty} \varphi_k(t) = \varphi(t)$ *uniformly on* $[0, T]$. *Then* φ *is a solution of* (2) *on* $[0, T]$.

Proof. φ is a quasi-trajectory of (2) as defined in [4, Def. 5]. Since $Ax + V(x)$ is convex for every x, the lemma follows from [4, Thm. 5].

LEMMA 2. *Let the assumptions of Theorem* 1 *be satisfied. Then given* $T > 0$, $\tau \geqq 0$, *for every* $\eta > 0$, *there exists an* $\varepsilon > 0$ *such that any solution* φ *of* (4_ε) *on* $[0, T]$ *with* $\varphi(0) \in \mathcal{R}(\tau)$ *satisfies* $\|\varphi - \psi\| < \eta$, *where* ψ *is the solution of* (2) *with* $\varphi(0) = \psi(0)$ *and* $\|\cdot\|$ *is the supremum norm in* $C(0, T)$.

Proof. Let us note first that because of the linear growth of the right-hand side of (4_ε), all its solutions can be extended to the entire right half-line.

Assume that the assertion of the lemma does not hold. Then there exists an $\eta > 0$ and sequences $\{\varepsilon_k\}$ of positive constants tending to 0 and $\{\psi_k\}$ of solutions of (4_{ε_k}) such that

$$(5) \qquad \|\psi_k - \varphi_k\| \geqq \eta,$$

where ψ_k is the solution of (2) with $\varphi_k(0) = \psi_k(0)$. Since both the sequences $\{\varphi_k\}$

and $\{\psi_k\}$ are uniformly bounded and equicontinuous, we can assume that they are uniformly convergent. Denote by φ, ψ, respectively, their limits. From (5) it follows that $\|\varphi - \psi\| \geqq \eta$. On the other hand, $\varphi(0) = \psi(0)$ and, by Lemma 1, both φ and ψ are solutions of (2). Since under the assumptions of Theorem 1 the solution of (2) starting at a given point is unique, then $\varphi = \psi$, which is impossible.

Proof of Theorem 1. Let $\omega > 0$ be chosen in such a way that $\omega D \subset \mathcal{R}(\vartheta)$ and $\mathcal{R}(\vartheta) + \omega D \subset \rho D$ for some suitably chosen $\vartheta > 0$. Due to [1, § 1, Property 1], such an ω exists. By Lemma 2, there exists an $\varepsilon > 0$ such that for every solution φ of (4_ε) with $\varphi(0) = x \in \mathcal{R}(\tau)$, we have $|\psi(t) - \varphi(t)| < \omega$ for $t \in [0, \tau]$, where ψ is the solution of (2) with $\varphi(0) = \psi(0)$. Since $\psi(T(x)) = 0$, we have $\varphi(T(x)) \in \omega D$. We prove $\varphi(t) \in \rho D$ for $t \geqq T(x)$ by induction.

Assume $\varphi(T(x) + k\tau) \in \omega D$ for some positive integer k. Since (4_ε) and (2) are autonomous, then

$$|\varphi(t) - \psi_k(t)| \leqq \omega \tag{6}$$

for $t \in [T(x) + k\tau,\ T(x) + (k+1)\tau]$, where ψ_k is the solution of (2) with $\psi_k(T(x) + k\tau) = \varphi(T(x) + k\tau)$. Since ψ_k is a solution of (2), we have $T(\psi_k(t)) \leqq T(\psi_k(T(x)) + k\tau) \leqq \vartheta$ and $\psi_k(T(x) + (k+1\tau) = 0$. From this and (6) we obtain $\varphi(t) \in \rho D$ for $t \in [T(x) + k\tau,\ T(x) + (k+1)\tau]$ and $\varphi(T(x) + (k+1)\tau) \in \omega D$. This completes the proof.

For the proof of Theorem 2 we have to recall some of the notations and results of [1]. In the rest of the paper we shall always assume $n = 2$.

For $u \in U$, denote $H(u, U) = \{\psi | \langle \psi, u \rangle = \max_{v \in U} \langle \psi, v \rangle\}$, where $\langle \cdot, \cdot \rangle$ stands for scalar product. The set $H(u, U)$ is a closed convex cone with vertex at 0 for every $u \in U$. If w is a vertex of U, due to normality and the linearity of the adjoint equation $\dot{\psi} = -A'\psi$, all its solutions which meet a boundary half-line of $H(w, U)$ have to cross it transversally in one direction. We call w

—*attracting*, if both boundary half-lines of $H(w, U)$ are crossed inwards,

—*neutral*, if one of the half-lines is crossed inwards, the other outwards,

—*repulsing*, if both boundary half-lines are cross outwards.

A vertex w is repulsing, neutral, or attracting according to whether Case 1, 2, or 3 of the proof of Theorem 1 of [1] takes place. It is shown in this proof that all the solutions of (2) are optimal precisely if no vertex of U is attracting and that w is attracting if and only if it satisfies the condition of [1, Thm.], i.e., if $H(w, U)$ contains the eigenvector of the larger eigenvalue of $-A'$ but not the other eigenvector of $-A'$. We now prove the following.

LEMMA 3. *If no vertex of U is attracting, then all vertices of U are neutral.*

Proof. Assume there exists a repulsing vector w_0 of U. This means that the solutions of the equation $\dot{\psi} = A'\psi$ (which are the solutions of the adjoint equation with time reversed) cross both boundary half-lines of $H(w_0, U)$ inwards. However, this is possible only if A' has two distinct eigenvalues and $H(w_0, U)$ contains the eigenvector of the larger eigenvalue of A' but not the other eigenvector of A'. This implies that there exists another vertex w which contains the eigenvector of the smaller eigenvalue of A', which is the larger eigenvalue of $-A'$. Consequently, w is attracting, which contradicts the assumption of the lemma.

For a vertex w of U define

$$\Gamma(w)=\left\{-\int_0^t e^{-sA}w\,ds \mid 0\leqq t\leqq \tau(w)\right\},$$

where $\tau(w)=\max\{t\mid \cap_{0\leqq s\leqq t} e^{sA'}H(w,U)\neq 0\}$ (cf. [1, Corollary 2]).

LEMMA 4. *For every vertex w of U and any $\tau<\tau(w)$, there exists a neighborhood B of $\Gamma(w)\cap\mathcal{R}(\tau)$ which is divided by $\Gamma(w)$ into one-sided neighborhoods B^+, B^- such that*

(i) *There exists a C^1 function s on B such that*

$$\Gamma(w)\cap\mathcal{R}(\tau)=\{x\in\mathcal{R}(\tau)\mid s(x)=0\},\qquad y(x)=\partial s/\partial x(x)\neq 0$$

for $x\in\bar{B}$, $s(x)>0$ for $x\in B^+$ and $s(x)<0$ for $x\in B^-$.

(ii) *If w is neutral and w_1 is the vertex adjacent to w such that the solutions of the adjoint equation leave $H(w_1,U)$ for $H(w,U)$, then*

$$\langle y(x), Ax+w_1\rangle<-\varkappa \tag{7}$$

for $x\in B$ and some $\varkappa>0$, and

$$V(x)=\begin{cases}\{w\} & \text{for } x\in B^-,\\ \{w_1\} & \text{for } x\in B^+,\\ \operatorname{co}\{w,w_1\} & \text{for } x\in\Gamma(w)\cap B.\end{cases}$$

Proof. The neighborhood B is obtained by patching together the neighborhoods B of [1, Lemma 9]. The validity of (ii) follows from the analysis of Case 2 in the proof of [1, Thm.] (to obtain (7) we use [1, (13)] and the compactness of $\bar{B}$).

LEMMA 5. *Let* $\dim U=2$. *Then for every $\tau>0$, there exists a $\mu_\tau>0$ such that for every $x\in\mathcal{R}(\tau)$, $\psi\in E_0(x)$,*

$$\max_{u\in U}\langle\psi, Ax+u\rangle\ (=\langle\psi, Ax+v(x)\rangle)>\mu_\tau.$$

Proof. By [1, § 2, Property 3], $\max_{u\in U}\langle\psi, Ax+u\rangle=\max_{u\in U}\langle e^{-T(x)A'},u\rangle$ for every $\psi\in E_0(x)$. Since $\dim U=2$, there exists an $\eta>0$ such that $\eta D\subset U$. Further, since $S_\tau=\cup_{0\leqq t\leqq\tau}\{e^{-tA'}\psi\mid|\psi|=1\}$ does not contain the origin and is compact, we have $\min_{\chi\in S_\tau}|\chi|>0$. Thus we have

$$\begin{aligned}&\min_{0\leqq T(x)\leqq\tau}\ \min_{\psi\in E_0(x)}\ \max_{u\in U}\langle\psi, Ax+u\rangle\\ &=\min_{0\leqq T(x)\leqq\tau}\ \min_{\psi\in E_0(x)}\ \max_{u\in U}\langle e^{-T(x)A'}\psi, u\rangle\\ &\geqq\min_{0\leqq T(x)\leqq\tau}\ \min_{|\psi|=1}\ \max_{u\in U}\langle e^{-T(x)A'}\psi, u\rangle\geqq\min_{\chi\in S_\tau}\max_{u\in U}\langle\chi, u\rangle\\ &=\eta\min_{\chi\in S_\tau}|\chi|>o.\end{aligned}$$

The following lemma is a transcription of [5, Thm. 3.2 and Remark 2]. We use the notation dT for the differential of T and $\partial T/\partial e(x)=\lim_{h\to 0}h^{-1}(T(x+he)-T(x))$ for the directional derivative of T in the direction e.

LEMMA 6. *If* $\min_{\psi \in E_0(x)} \max_{u \in U} \langle \psi, Ax+u \rangle > 0$, *then* $(\partial T/\partial e)(x)$ *exists for any* $e \in R^n$, *and*

$$(\partial T/\partial e)(x) = -\max_{\psi \in E_0(x)} [\langle \psi, Ax+v(x) \rangle^{-1} \langle \psi, e \rangle].$$

Moreover, if $\lim_{t \to 0} e(t) = e$, *then*

$$\lim_{t \to 0} t^{-1}[T(x+te(t)) - T(x)] = (\partial T/\partial e)(x).$$

Note that Lemma 7 of [1] is a consequence of Lemma 6.

COROLLARY 1. *If* $\varphi : [-\sigma, \sigma] \to R^n$ *and* $\dot{\varphi}(0))$ *exists, then* $(d(T \circ \varphi)/dt)(0) = (\partial T/\partial \dot{\varphi}(0))(\varphi(0))$. *In particular,* $(\partial T/\partial(Ax+v(x)))(x) = (\partial(T \circ \xi_x(t))/dt)(0) = -1$ (ξ_x *being the open-loop optimal trajectory from* x).

LEMMA 7. *Let* $\dim U = 2$. *Then given* $\tau > 0$, *there exists a* K_τ *such that* $|(\partial T/\partial e)(x)| \leqq K_\tau$ *for every* $x \in \mathcal{R}(\tau)$ *and every* $|e| \leqq 1$.

Proof. The proof follows immediately from Lemmas 5 and 6.

COROLLARY 2. $|dT(x)| \leqq K_\tau$ *as soon as it exists and* $K_\tau \geqq |Ax+v(x)|^{-1}$ *for all* $x \in \mathcal{R}(\tau)$.

Proof of Theorem 2. By virtue of Theorem 1, it suffices to prove Theorem 2 for $\tau > o$ sufficiently small. Namely, assume that we have proved Theorem 2 for $0 \leqq \tau \leqq \tau_0$. Then given $\eta > 0$, we can choose $\rho > 0$ such that $\rho D \subset \mathcal{R}(\eta/2)$ and $\varepsilon > 0$ such that every solution starting in ρD reaches the origin in time $t \leqq \eta/2 + \eta/2 = \eta$. Let us now choose $\tau \geqq 0$ arbitrarily. By Theorem 1, if ε is suitably restricted, the solutions of (4_ε), starting at points $x \in \mathcal{R}(\tau)$, reach ρD in time $T(x)$ and, thus, reach the origin in time $T(x)+\eta$.

Let us therefore assume that $\tau < \min\{\tau(w) | w = \text{vertex of } U\}$. We prove that given $\varepsilon > 0$ sufficiently small, for any $x \in \mathcal{R}(\tau)$, $x \neq 0$, there exists a $\sigma > 0$ such that

$$T(\varphi(t)) - T(x) \leqq (-1 + K_\tau \varepsilon)t \tag{8}$$

for any $t \in [0, \sigma]$ and any solution of (4_ε) with $\varphi(0) = x$.

Let us assume first that $x \in \mathcal{R}(\tau) - \Gamma$, where $\Gamma = \cup\{\Gamma(w) | = \text{vertex of } U\}$. Since for any $\varepsilon > 0$ the right-hand side of (4_ε) is uniformly bounded in $\mathcal{R}(\tau+1)$ and $\operatorname{int} \mathcal{R}(\tau+1) \setminus \bar{\Gamma}$ is open, given $\varepsilon > 0$, for any $x \in \mathcal{R}(\tau) \setminus \Gamma$, there exists a $\sigma > 0$ such that for $t \in [0, \sigma]$, $\varphi(t) \in \operatorname{int} \mathcal{R}(\tau+1) \setminus \bar{\Gamma}$. Since T is differentiable in $\mathcal{R}(\tau+1) \setminus \bar{\Gamma}$, we have for almost all $t \in [0, \sigma]$,

$$dT(\varphi(t))/dt = dT(\varphi(t))(A\varphi(t) + w(t) + \varepsilon\delta(t)),$$

$w(t)$ and $\delta(t)$ being measurable and $w(t) \in W(\varphi(t))$, $|\delta(t)| \leqq 1$, where $W(x) = \operatorname{co} \cup_{\psi \in E_0(x)} \{u \in U | \psi \in H(u, U)\}$[1] (the existence of measurable w and δ follows from the upper semicontinuity of W ([1, Lemma 6]) and the Filippov implicit function lemma [6]).

[1] In the definition of W in [1, after Lemma 5] there is an obvious error: $\cap$ should be replaced by $\cup$.

For $x \in \mathcal{R}(\tau) \setminus \Gamma$, $E_0(x)$ consists of a unique point which we denote by $\Psi(x)$ (cf. [1, Cor. 2]). We have

$$
\begin{aligned}
&dT(\varphi(t))[A\varphi(t) + w(t) + \varepsilon\delta(t)] \\
&\quad = dT(\varphi(t))[A\varphi(t) + w(t)] + \varepsilon\, dT(\varphi(t))\delta(t) \\
&\quad \leqq -\max_{u \in U} \langle \Psi(\varphi(t)), A\varphi(t) + u\rangle^{-1} \langle \Psi(\varphi(t)), A\varphi(t) + w(t)\rangle \\
&\qquad + \varepsilon K_{\tau+1} = -1 + \varepsilon K_{\tau+1}.
\end{aligned} \tag{9}
$$

Choosing $\varepsilon < K_{\tau+1}^{-1}$, we see that T is decreasing and therefore $T(\varphi(t)) \leqq \tau$ for all $t \in [0, \sigma]$, whence $K_{\tau+1}$ in (9) can be replaced by K_τ. Integrating (9), we obtain (8).

Now let $x \in \Gamma(w) \cap \mathcal{R}(\tau)$. We define B, B^+ etc. as in Lemma 4. Again, ε, σ can be chosen so small that for $t \in [0, \sigma]$, φ does not leave B. We prove that for $\varepsilon > 0$ sufficiently small, φ does not enter B^+ for $t \in [0, \sigma]$.

Assume the contrary. Then there exists an interval $[\sigma_1, \sigma_2]$ such that $\varphi(\sigma_1) \in \Gamma(w)$ and $\varphi(t) \in B^+ \cap \mathcal{R}(\tau(w))$ for $t \in (\sigma_1, \sigma_2]$, which is possible only if $ds(\varphi(t))/dt = \langle y(\varphi(t)), \dot{\varphi}(t)\rangle > 0$ on some subset of nonzero measure of $[\sigma_1, \sigma_2]$. However, if $0 < \varepsilon < \max_{x \in \bar{B}^+} |y(x)| \varkappa^{-1}$, we have

$$
\begin{aligned}
\langle y(\varphi(t)), \dot{\varphi}(t)\rangle &= \langle y(\varphi(t)), A\varphi(t) + w_1 + \varepsilon\delta(t)\rangle \\
&\leqq \langle y(\varphi(t)), A\varphi(t) + w_1\rangle + \varepsilon |y(\varphi(t))| < -\varkappa + \varkappa = 0
\end{aligned}
$$

(where $|\delta(t)| \leqq 1$ and w_1 is as in Lemma 4) as soon as $\varphi(t) \in B^+$ and $\dot{\varphi}(t)$ exists, which is impossible.

Since φ cannot enter B^+, we have $\langle y(\varphi(t)), \dot{\varphi}(t)\rangle \leqq 0$ as soon as $\varphi(t) \in \Gamma(w)$ and $\dot{\varphi}(t)$ exists. Given $s \in [0, \sigma]$, denote by I the set of those points $t \in [0, s]$ for which $\dot{\varphi}(t)$ exists, by I_1 the set of those $t \in I$ for which $\varphi(t) \in \Gamma(w)$ and $\dot{\varphi}(t)$ is not tangent to $\Gamma(w)$, by I_2 the set of those $t \in I$ for which $\varphi(t) \in \Gamma(w)$ and $\dot{\varphi}(t)$ is tangent to $\Gamma(w)$ and by I_3 the set of those $t \in I$ for which $\varphi(t) \notin \Gamma(w)$. Obviously I, as well as I_1, I_2, I_3, are measurable and $\mu(I) = s$. Furthermore, I_1 obviously consists of isolated points and therefore has zero measure. Since $Ax + w$ is tangent to $\Gamma(w)$ at x, for $t \in I_2$ we have by Lemma 4 and Corollary 2,

$$
\begin{aligned}
\dot{\varphi}(t) &\in \{A\varphi(t) + \operatorname{co}\{w, w_1\} + \varepsilon D\} \cap \{\lambda(A\varphi(t) + w) | \lambda \geqq 0\} \\
&= \{(A\varphi(t) + w)(1 + |A\varphi(t) + w|^{-1})\vartheta(t)\varepsilon \,|\, |\vartheta(t)| \leqq 1\} \\
&\subset \{(A\varphi(t) + w)(1 + K_{\tau(w)}\vartheta(t))\varepsilon \,|\, |\vartheta(t)| \leqq 1\},
\end{aligned}
$$

i.e., $\dot{\varphi}(t)$ is a multiple of $A\varphi(t) + w$ by a coefficient $\geqq (1 - K_{\tau(w)}\varepsilon)$. Since $w = v(\varphi(t))$, we have by Corollaries 1 and 2,

$$
\begin{aligned}
&(\partial T/\partial(\dot{\varphi}(t)))(\varphi(t)) \\
&\quad = |\dot{\varphi}(t)| |A\varphi(t) + w|^{-1} (\partial T/\partial(A\varphi(t) + w))(\varphi(t)) \\
&\quad = (\dot{\varphi}(t))(A\varphi(t) + w)^{-1}(-1) \leqq -1 + K_{\tau(w)}\varepsilon.
\end{aligned}
$$

From this we obtain for some $|\delta(t)| \leqq 1$,

$$
\begin{aligned}
T(\varphi(s)) - T(x) &= \int_0^s (dT(\varphi(t))/dt)\, dt \\
&= \int_{I_2} (dT(\varphi(t))/dt)\, dt + \int_{I_3} (dT(\varphi(t))/dt) dt \\
&= \int_{I_2} (\partial T/\partial \dot{\varphi}(t))(\varphi(t))\, dt + \int_{I_3} dT(\varphi(t))\dot{\varphi}(t)\, dt \\
&= [-1 + K_{\tau(w)}\varepsilon]\mu(I_2) + \int_{I_3} \{\langle \Psi(\varphi(t), A\varphi(t) + w)\rangle^{-1} \\
&\qquad \cdot \langle \Psi(\varphi(t), A\varphi(t) + w)\rangle + dT(\varphi(t))\varepsilon\delta(t)\}\, dt \\
&\leqq [-1 + K_{\tau(w)}\varepsilon]\mu(I_2) + [-1 + K_{\tau(w)}\varepsilon]\mu(I_3) \\
&\qquad \cdot [-1 + K_{\tau(w)}\varepsilon]s.
\end{aligned}
$$

Again, if ε is chosen sufficiently small, $K_{\tau(w)}$ can be replaced by K_τ. Thus (8) is established for all $x \in \mathcal{R}(\tau)$.

We now prove that if $\varepsilon > 0$ is sufficiently small and φ is a solution of (4_ε) starting at some point $x \in \mathcal{R}(\tau)$, then

$$
T(\varphi(t)) - T(x) \leqq \max\{-T(x), (-1 + K_\tau\varepsilon)t\}. \tag{10}
$$

Assume the contrary, and denote by t_0 the supremum of those t for which (10) is valid. Since (8) is valid for all $x \neq 0$, $T(\varphi(t))$ is nonincreasing. Therefore, $T(\varphi(t_0)) > 0$ and there exists a sequence of points $\{t_k\}$, $t_k \searrow t_0$ such that $T(\varphi(t_k)) - T(\varphi(t_0)) > (-1 + K_\tau\varepsilon)(t_k - t_0)$, which violates (8).

Now, given $x \in \mathcal{R}(\tau)$, it follows from (10) that $\varphi(t) = 0$ as soon as $t \geqq (1 - K_\tau\varepsilon)^{-1} T(x)$, which completes the proof.

REFERENCES

[1] P. BRUNOVSKÝ, *The closed-loop time optimal control. I: Optimality*, this Journal, 12 (1974), pp. 624–634.

[2] ———, *On the best stabilizing control under a given class of perturbations*, Czech. Math. J., 15 (1965), pp. 329–369.

[3] H. HERMES, *Discontinuous vector fields and feedback control*, Differential Equations and Dynamical Systems, J. Hale and J. LaSalle, eds., Academic Press, New York, 1967, pp. 155–166.

[4] T. WAŻEWSKI, *On an optimal control problem*, Differential Equations and their Applications, Publ. House of the Czechoslovak Academy of Sciences, Prague, 1963, pp. 229–242.

[5] B. N. PŠENIČNYJ, *Line jny je differencial' ny je igry*, Avtomat. i Telemeh., (1968), no. 1, pp. 65–78.

[6] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, Vestnik Moskov. Univ. Ser. I Mat. Meh., (1959), no. 2, pp. 25–32; English transl., this Journal, 1 (1962), pp. 76–84.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

CONTROL OF LINEAR SYSTEMS THROUGH SPECIFIED INPUT CHANNELS*

J. P. CORFMAT AND A. S. MORSE†

Abstract. It is shown for the controllable linear system $\dot{x} = Ax + Bu + Dv$, $y = Cx$ that there exists a feedback map F for which $\dot{x} = (A + DFC)x + Bu$ is controllable if and only if the number of transmission polynomials of (C, A, B) is no greater than the rank of the (nonzero) transfer matrix of (C, A, B). If this condition fails to hold, then for all F, the spectrum of $A + DFC$ contains a uniquely determined subset of transmission zeros, and this subset coincides with the spectrum of $A + DFC$ modulo the controllable space of $(A + DFC, B)$ whenever F is selected so that the dimension of the controllable space is as large as possible. Under mild assumptions, the transmission polynomials are identified as the numerator polynomials of the rational functions which appear in the Smith–McMillan form of the transfer matrix of (C, A, B).

Introduction. In this paper, we consider the problem of selecting for the two-input channel, controllable, linear system

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

an output feedback law $v = Fy$ so that the resulting closed-loop system

$$\dot{x}(t) = (A + DFC)x(t) + Bu(t) \tag{2}$$

is controllable with $u(\cdot)$. The solution to the problem extends earlier results ([1]–[4]) and is central to the construction of decentralized feedback laws for assigning the closed-loop spectrum of a linear system (cf. [5]). Applications to decentralized control will be discussed in a future paper.

The present study differs from previous work [2] in that here (1) is not required to be an observable model. Unobservable models appear more the rule than the exception when one begins to examine the structural properties of various interconnections of linear systems. For example, if (1) is the composite model of two noninteracting subsystems, one (Σ_1) with input u and output y, the other (Σ_2) with input v and output

$$z(t) = Lx(t), \tag{3}$$

then (2) is the state equation which results if Σ_1 is connected in cascade with Σ_2 by means of the control $v = Fy$. Note that the subsystem $\Sigma_1 = (C, A, B)$ can be neither controllable nor observable, even if (1) together with (3) is. By studying the conditions under which F can be chosen to make (2) controllable with u (and observable with z), one can determine what is required to stabilize or otherwise control dynamic response with feedback from z to u.

* Received by the editors July 9, 1974, and in revised form December 23, 1974. This work was supported by the U.S. Air Force Office of Scientific Research under Grant 72-2211.

† Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520.

In the sequel (§ 1), it will be shown that (2) can be made controllable with F just in case the number of transmission polynomials of (C, A, B) (cf. [6]) is no greater than the rank of the (nonzero) transfer matrix $C(\lambda I - A)^{-1}B$ over the field of rational functions in λ. If (2) cannot be made controllable, then for each choice of F, (2) has an uncontrollable spectrum Λ_F. It is shown that for arbitrary F, Λ_F contains a fixed subset Λ_0 consisting of certain uniquely determined transmission zeros of (C, A, B); it is further shown that $\Lambda_F = \Lambda_0$ whenever F is selected so that the dimension of the controllable space of (2) is as large as possible. Finally, in § 2, we relate the transmission polynomials of (C, A, B) (a state space concept) to a familiar invariant of transfer matrices by showing that under mild assumptions, the transmission polynomials coincide with the numerator polynomials of the rational functions which appear in the Smith–McMillan form of the transfer matrix of (C, A, B).

Notation. In the sequel, $\mathbb{R}[\lambda]$ is the ring of polynomials with coefficients in the field of reals $\mathbb{R}$. Script letters $\mathscr{X}, \mathscr{Y}, \cdots$, denote $\mathbb{R}$-vector spaces with elements $x, y, \cdots$, and $d(\mathscr{X})$ is the dimension of $\mathscr{X}$. Both linear maps and matrices are denoted by capital letters $A, B, \cdots$, while im A and ker A abbreviate image A and kernel A, respectively. The restriction or definition of $M : \mathscr{R} \to \mathscr{S}$ on $\mathscr{T} \subset \mathscr{R}$ is written as $M|\mathscr{T}$.

The spectrum of a map $A : \mathscr{X} \to \mathscr{X}$, written $\sigma(A)$, is the symmetric set of complex roots of the characteristic polynomial (c.p.) of A repeated according to multiplicity. The set of invariant factors (i.f.) of A is written as a list $\{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ ordered so that $\alpha_i | \alpha_{i-1}$ (divides), $i = 2, 3, \cdots, k$. If $\mathscr{W}$ is A-invariant, $A|\mathscr{W}$ denotes the restriction of A to $\mathscr{W}$ and $A\|(\mathscr{X}/\mathscr{W})$ is the map induced by A in $\mathscr{X}/\mathscr{W}$.

If k is a positive integer, $\mathbf{k} \equiv \{1, 2, \cdots, k\}$, and $\{\alpha_i, i \in \mathbf{k}\}$ abbreviates the list $\{\alpha_1, \alpha_2, \cdots, \alpha_k\}$.

The maps $A : \mathscr{X} \to \mathscr{X}$, $B : \mathscr{U} \to \mathscr{X}$, $C : \mathscr{X} \to \mathscr{Y}$, and $D : \mathscr{V} \to \mathscr{X}$ ($d(\mathscr{X}) = n$, $d(\mathscr{U}) = m$, $d(\mathscr{Y}) = p$, $d(\mathscr{V}) = q$) are fixed and are associated with the linear system (1), (2). We write $\mathscr{B}$ for im B, $\langle A|\mathscr{B}\rangle = \mathscr{B} + A\mathscr{B} + \cdots + A^{n-1}\mathscr{B}$ for the controllable space of (A, B), and $[C|A] = \cap_{i=1}^{n} \ker CA^{i-1}$ for the unobservable space of (C, A). If $A\mathscr{W} \subset \mathscr{W} \subset \ker C$, the system induced by (C, A, B) in $\mathscr{X}/\mathscr{W}$ is the triple $(\bar{C}, \bar{A}, \bar{B})$ where $\bar{A} = A\|(\mathscr{X}/\mathscr{W})$, $\bar{B} = PB$, $P : \mathscr{X} \to \mathscr{X}/\mathscr{W}$ is the canonical projection, and $\bar{C}$ is the unique solution to $C = \bar{C}P$.

1. Main problem. The problem to be analyzed is as follows.

Main Problem. Let A, B, C, D be fixed with $\langle A|\mathscr{B} + \mathscr{D}\rangle = \mathscr{X}$. Find conditions for the existence of a map $F : \mathscr{Y} \to \mathscr{V}$ such that

$$\langle A + DFC|\mathscr{B}\rangle = \mathscr{X}.$$

Remark 1. In the sequel, it will be assumed without loss of generality that

$$C(\lambda I - A)^{-1}B \neq 0. \tag{4}$$

For if (4) were false, $\langle A + DFC|\mathscr{B}\rangle$ would be independent of F.

First, let $\mathbf{F}^*$ denote the class of all maps F for which the dimension of $\langle A+DFC|\mathcal{B}\rangle$ is as large as possible. It is reasonably clear (and can easily be shown) that if F is chosen arbitrarily, it is almost certain to be in $\mathbf{F}^*$; in other words, either $\mathbf{F}^*$ coincides with the space of all maps $F: \mathcal{Y} \to \mathcal{V}$, or, at worst, $\mathbf{F}^*$ can be viewed as the complement of a suitably defined proper variety in the space of all maps from $\mathcal{Y}$ to $\mathcal{V}$.[1]. Thus the problem of selecting F so that $d(\langle A+DFC|\mathcal{B}\rangle)$ is as large as possible is, in principle, a simple computational matter.

Our interest is not so much in computing $F \in \mathbf{F}^*$ as in expressing $d(\langle A+DFC|\mathcal{B}\rangle)$ for $F \in \mathbf{F}^*$ in terms of the problem data A, B, C.[2] To this end, let $\{\alpha_i, i \in \mathbf{t}\}$ denote the list of transmission polynomials (t.p.) of (C, A, B) (cf. [6]), ordered so that $\alpha_i | \alpha_{i-1}$, $i = 2, \cdots, t$, write r for the rank of the nonzero transfer matrix $C(\lambda I - A)^{-1}B$ over the field of rational functions in λ, and define

$$n^* = \begin{cases} n & \text{if } r \geqq t, \\ n\text{-deg}\left(\prod_{i=r+1}^{t} \alpha_i\right) & \text{if } r < t. \end{cases}$$

Our first result is as follows.

THEOREM 1. *Let A, B, C, D be fixed with $\langle A|\mathcal{B}+\mathcal{D}\rangle = \mathcal{X}$ and $C(\lambda I - A)^{-1}B \neq 0$. Then*

$$\max_F d(\langle A+DFC|\mathcal{B}\rangle) = n^*. \tag{5}$$

The theorem asserts that if F is selected so that the dimension of $\langle A + DFC|\mathcal{B}\rangle$ is as large as possible, then this dimension must equal the dimension of $\mathcal{X}$ less the degree of the product of the last $t-r$ polynomials in the list of transmission polynomials of (C, A, B). The solution to the main problem is now clearly as follows.

COROLLARY 1. *Under the hypotheses of Theorem* 1, *there exists a map F such that*

$$\langle A+DFC|\mathcal{B}\rangle = \mathcal{X}$$

if and only if the number of transmission polynomials of (C, A, B) is no greater than the rank of the transfer matrix of (C, A, B).

Remark 2. We call (C, A, B) *complete* if $t \leqq r$. Since the rank of a transfer matrix is invarient under transpösition, and since the t.p. of (C, A, B) are the same as the t.p. of the dual system (B', A', C') (see Remark 6),[3] it follows that (C, A, B) is complete if and only if its dual system is complete.

Corollary 1 implies that if (A, B) is controllable, then (C, A, B) is complete. This and duality allow us to assert that either controllability of (A, B) or

[1] This variety can be defined in matrix terms as the set of all points F in the real parameter space of $d(\mathcal{V}) \times d(\mathcal{Y})$ matrices for which all qth order minors of the controllability matrix of $(A+DFC, B)$ equal zero, where $q = \max_F d(\langle A+DFC|\mathcal{B}\rangle)$.

[2] D plays no role in what follows, provided $\langle A|\mathcal{D}+\mathcal{B}\rangle = \mathcal{X}$.

[3] Prime denotes dual.

observability of (C, A) implies completeness of (C, A, B). From this and Corollary 1, we recover the main result of [2], as follows.

COROLLARY 2. *If (C, A) is observable and $\langle A|\mathscr{B}+\mathscr{D}\rangle = \mathscr{X}$, there exists a map F such that $(A+DFC, B)$ is controllable.*

If (C, A, B) is incomplete (i.e., if $r<t$), then Corollary 1 implies that $(A+DFC, B)$ cannot be made controllable. In this case, the following theorem provides a characterization of the "uncontrollable spectrum" of $A+DFC$ in terms of the list $\{\alpha_{r+1}, \cdots, \alpha_t\}$ (i.e., the *remnant polynomials* of (C, A, B)).

THEOREM 2. *Suppose (C, A, B) is incomplete (i.e., $t>r$), and, for arbitrary F, let $\{\rho_i^F, i \in \mathbf{q}^F\}$ denote the list of invariant factors of the map induced by $A+DFC$ in $\mathscr{X}/\langle A+DFC|\mathscr{B}\rangle$. Under the hypotheses of Theorem* 1,

$$q^F \geqq t-r \quad \text{and} \quad \alpha_{r+i}|\rho_i^F, \quad i \in \{1, 2, \cdots, t-r\}, \quad \text{for all } F. \tag{6}$$

In addition, if $d(\langle A+DFC|\mathscr{B}\rangle) = n^*$, then

$$q^F = t-r \quad \text{and} \quad \alpha_{r+i} = \rho_i^F, \quad i \in \{1, 2, \cdots, t-r\}. \tag{7}$$

Let Λ_0 denote the symmetric set of complex roots of the polynomial $\prod_{i=r+1}^{t} \alpha_i$, repeated according to multiplicity (i.e., the *remnant zeros* of (C, A, B)). Theorem 2 implies that Λ_0 is a fixed subset of the spectrum of $A+DFC$ for all F, and that Λ_0 equals the spectrum of $A+DFC$ mod $\langle A+DFC|\mathscr{B}\rangle$ (i.e., the *uncontrollable spectrum* of $A+DFC$) whenever $d(\langle A+DFC|\mathscr{B}\rangle)$ is as large as possible. From this, we immediately obtain the following corollary.

COROLLARY 3. *Under the hypotheses of Theorem* 1, *there exists a map F such that the pair $(A+DFC, B)$ is stabilizable if and only if Λ_0 is a stable set.*[4]

The principal problem involved in proving Theorems 1 and 2 is that unless (C, A, B) is complete, the family of subspaces $\{\langle A+DFC|\mathscr{B}\rangle : F : \mathscr{Y} \to \mathscr{V}\}$ does not contain a unique largest member relative to a partial ordering by inclusion. To proceed, it is first necessary to show, for any F, that $\mathscr{R}_F \equiv \langle A+DFC|\mathscr{B}\rangle$ can be viewed as the controllable space of another system.

Below, $\mathscr{T}$ denotes the unique smallest subspace satisfying $\mathscr{B} \subset \mathscr{T}$ and $A(\mathscr{T} \cap \ker C) \subset \mathscr{T}$ (cf. [6]); in addition, $\mathbf{K} \equiv \{K : (A+KC)\mathscr{T} \subset \mathscr{T}\}$, and $\mathbf{K}^*$ is the (nonempty) subset $\mathbf{K}^* \equiv \{K : \operatorname{im} K \subset \mathscr{D}+\mathscr{T}+A\mathscr{T}, K \in \mathbf{K}\}$.

PROPOSITION 1. *For each map $F : \mathscr{Y} \to \mathscr{V}$, there exist maps $F_0 : \mathscr{Y} \to \mathscr{V}$ and $K \in \mathbf{K}^*$ such that*

$$\mathscr{R}_F = \langle A+KC|\mathscr{T}+(DF_0-K)C\mathscr{T}\rangle, \tag{8a}$$

$$(A+DFC)\|(\mathscr{X}/\mathscr{R}_F) = (A+KC)\|(\mathscr{X}/\mathscr{R}_F). \tag{8b}$$

Conversely, if $F_0 : \mathscr{Y} \to \mathscr{V}$ and $K \in \mathbf{K}^*$ are arbitrary, there exists a map $F : \mathscr{Y} \to \mathscr{V}$ such that (8) *holds.*

Proof. First, observe that if $K \in \mathbf{K}$, then $\mathscr{T}+A\mathscr{T} = \mathscr{T}+(A+KC-KC)\mathscr{T} = \mathscr{T}+KC\mathscr{T}$, so that

$$\mathscr{T}+A\mathscr{T}+\mathscr{D} = \mathscr{T}+KC\mathscr{T}+\mathscr{T}, \qquad K \in \mathbf{K}. \tag{9}$$

[4] Λ_0 is stable, if each of its elements is a point in the open left-half complex plane.

It will now be shown that (i) for any $F: \mathcal{Y} \to \mathcal{V}$, there exist $K \in \mathbf{K}^*$ and $F_0: \mathcal{Y} \to \mathcal{V}$ such that

$$\text{im}\,(DF-K) \subset \mathcal{T}+(DF_0-K)C\mathcal{T}, \tag{10}$$

$$F|C\mathcal{T} = F_0|C\mathcal{T}, \tag{11}$$

and conversely, that (ii) for any $K \in \mathbf{K}^*$ and $F_0: \mathcal{Y} \to \mathcal{V}$, there exists $F: \mathcal{Y} \to \mathcal{V}$ satisfying (10) and (11).

(i) Let F be fixed; then (11) holds with $F_0 \equiv F$. To construct K, let $\mathcal{T}_0$ be any subspace satisfying $\mathcal{T}_0 \oplus \mathcal{T} \cap \ker C = \mathcal{T}$, and write T_0 for the insertion of $\mathcal{T}_0$ in $\mathcal{X}$; then CT_0 is monic with left inverse L. If

$$K \equiv DF-(A+DFC)T_0L, \tag{12}$$

then $KCT_0 = DFCT_0-(A+DFC)T_0LCT_0 = -AT_0$, so that $(A+KC)T_0 = 0$. It follows that $(A+KC)\mathcal{T} = (A+KC)(\mathcal{T}_0+\mathcal{T}\cap\ker C) = A(\mathcal{T}\cap\ker C) \subset \mathcal{T}$; thus $K \in \mathbf{K}$. From (12), $\text{im}\,(DF-K) \subset \text{im}\,(A+DFC)T_0 = \text{im}\,(A+DFC)T_0LCT_0 = \text{im}\,(DF-K)CT_0$, so $\text{im}\,(DF-K) \subset (DF-K)C\mathcal{T}$, and since $F = F_0$, (10) is true. But (10) implies $\text{im}\,K \subset \mathcal{D}+\mathcal{T}+KC\mathcal{T}$; thus, from (9), $\text{im}\,K \subset \mathcal{T}+A\mathcal{T}+\mathcal{D}$, and since $K \in \mathbf{K}$, $K \in \mathbf{K}^*$.

(ii) Let F_0 and $K \in \mathbf{K}^*$ be fixed. Then $K \in \mathbf{K}$, so from (9) and the definition of $\mathbf{K}^*$, $\text{im}\,K \subset KC\mathcal{T}+\mathcal{T}+\mathcal{D} = (DF_0-K)C\mathcal{T}+\mathcal{T}+\mathcal{D}$. Hence there exists a map F_1 such that

$$\text{im}\,(DF_1-K) \subset (DF_0-K)C\mathcal{T}+\mathcal{T}. \tag{13}$$

Let $\mathcal{Y}_0$ be any subspace satisfying $\mathcal{Y}_0 \oplus C\mathcal{T} = \mathcal{Y}$, and define F so that $F|\mathcal{Y}_0 = F_1|\mathcal{Y}_0$ and $F|C\mathcal{T} = F_0|C\mathcal{T}$; then F satisfies (11). Since $\text{im}\,(DF-K) = (DF-K)\cdot(\mathcal{Y}_0+C\mathcal{T}) = (DF_1-K)\mathcal{Y}_0+(DF_0-K)C\mathcal{T}$, (13) implies that F satisfies (10) as well.

Now let $K \in \mathbf{K}^*$, F and F_0 be any maps satisfying (10) and (11). To complete the proof, it is sufficient to show that (8) follows.

Set $\mathcal{R}_F \equiv \langle A+DFC|\mathcal{B}\rangle$, then $\mathcal{B} \subset \mathcal{R}_F$ and $A((\ker C)\cap\mathcal{R}_F) \subset \mathcal{R}_F$. Since $\mathcal{T}$ is the smallest subspace with these two properties, $\mathcal{T} \subset \mathcal{R}_F$. Thus $\mathcal{R}_F = \langle A+DFC|\mathcal{B}\rangle \subset \langle A+DFC|\mathcal{T}\rangle \subset \langle A+DFC|\mathcal{R}_F\rangle = \mathcal{R}_F$, or $\mathcal{R}_F = \langle A+DFC|\mathcal{T}\rangle$. Clearly,

$$\mathcal{R}_F = \langle A+DFC|\mathcal{T}+(A+DFC)\mathcal{T}\rangle \tag{14}$$

But $\mathcal{T}+(A+DFC)\mathcal{T} = \mathcal{T}+(A+KC+DFC-KC)\mathcal{T} = \mathcal{T}+(DF-K)C\mathcal{T}$; hence, from (11), $\mathcal{T}+(A+DFC)\mathcal{T} = \mathcal{T}+(DF_0-K)C\mathcal{T}$. This and (14) imply that $\mathcal{R}_F = \langle A+KC+(DF-K)C|\mathcal{T}+(DF_0-K)C\mathcal{T}\rangle$. It now follows from (10) that (8a) is true.

Let $P: \mathcal{X} \to \mathcal{X}/\mathcal{R}_F$ be the canonical projection, and write $\bar{A}_F$ and $\bar{A}_K$ for the maps induced by $A+DFC$ and $A+KC$, respectively, in $\mathcal{X}/\mathcal{R}_F$. From (8a) and (10), $\text{im}\,(DF-K) \subset \mathcal{R}_F$; thus $P(DF-K) = 0$, so $P(A+DFC) = P(A+KC)$. By definition, $\bar{A}_FP = P(A+DFC)$ and $\bar{A}_KP = P(A+KC)$; since P is epic, it follows that $\bar{A}_F = \bar{A}_K$.

LEMMA 1. *Let $\mathcal{W} \subset \mathcal{X}$ be A-invariant, write $\{\sigma_i, i \in \mathbf{k}\} \equiv$ i.f. A, $\{\mu_i, i \in \mathbf{j}\}$*

$\equiv$ i.f. $A|\mathscr{W}$ *and* $\{\beta^i, i \in \mathbf{q}\} \equiv$ i.f. $A\|(\mathscr{X}/\mathscr{W})$. Then

(i) $j \leqq k$ *and* $\mu_i | \sigma_i, i \in \mathbf{j}$;

(ii) $k - j \leqq q$ *and if* $j < k$, *then* $\sigma_{j+i} | \beta_i, i \in \{1, 2, \cdots, k-j\}$.

A proof of this lemma appears in the Appendix.

LEMMA 2. *Let* (C, A, B) *be a controllable system with* $[C|A] \neq 0$; *write* $\{\alpha_i, i \in \mathbf{k}\} \equiv$ i.f. $A|[C|A]$. *There exists a map* F *and a polynomial* δ *such that* i.f. $(A + BFC) = \{\delta\alpha_1, \alpha_2, \cdots, \alpha_k\}$.

Proof. In the quotient space $\bar{\mathscr{X}} \equiv \mathscr{X}/[C|A]$, the system $(\bar{C}, \bar{A}, \bar{B})$ induced by (C, A, B) is both controllable and observable. Hence, by the main result of [1], there exist maps $g : \mathbb{R} \to \mathscr{U}$, $h : \mathscr{Y} \to \mathbb{R}$ and $F_0 : \mathscr{Y} \to \mathscr{U}$ such that $(h\bar{C}, \bar{A} + \bar{B}F_0\bar{C}, \bar{B}g)$ is single-input controllable and single-output observable. As $\mathbb{R}$ is an infinite field, root locus analysis provides a scalar $\mu \in \mathbb{R}$ for which c.p. $(\bar{A} + \bar{B}F_0\bar{C} + \bar{B}g\mu h\bar{C})$ is coprime with α_1. Fix μ at this value, let δ be the corresponding c.p. and define $F = F_0 + g\mu h$; then $\bar{A} + \bar{B}F\bar{C}$ is cyclic with c.p. δ. If $\{\beta_i, i \in \mathbf{j}\} \equiv$ i.f. $(A + BFC)$, then $\beta_1 =$ minimal polynomial (m.p.) of $A + BFC$. Since $\delta =$ m.p. $(A + BFC)$ mod $[C|A]$, $\alpha_1 =$ m.p. $(A + BFC)\|[C|A]$, and α_1 and δ are coprime, it follows that $\beta_1 = \delta\alpha_1$. In addition, $\prod_{i=1}^{j} \beta_i =$ c.p. $(A + BFC) = \delta(\prod_{i=1}^{k} \alpha_i)$ and, by Lemma 1, $k \leqq j$ and $\alpha_i | \beta_i$, $i \in \mathbf{j}$. This is possible only if $j = k$ and $\beta_i = \alpha_i$, $i \in \{2, \cdots, k\}$.

Remark 3. In the sequel, use will be made of the following interpretation of the transmission polynomials of (C, A, B), which by Remark 6 are the same as the transmission polynomials of the dual system (B', A', C').

If $\mathscr{S}$ is the largest (A, B)-invariant subspace in ker C, and if $\mathscr{S}^\perp$ and $\mathscr{T}^\perp$ are the annihilators of $\mathscr{S}$ and $\mathscr{T}$, respectively, then, as noted in [6], $\mathscr{T}^\perp$ (resp. $\mathscr{T}^\perp \cap \mathscr{S}^\perp$) is the largest (A', C')-invariant (resp. controllability) subspace in ker B'. Hence, by definition, t.p. $(B', A', C') =$ i.f. $(A + KC)'\|(\mathscr{T}^\perp/(\mathscr{T}^\perp \cap \mathscr{S}^\perp))$ for any $K \in \mathbf{K}$. Since $(A + KC)'\|(\mathscr{T}^\perp/(\mathscr{T}^\perp \cap \mathscr{S}^\perp)$ is similar to $(A + KC)\|((\mathscr{S} + \mathscr{T})/\mathscr{T})$, it follows that

$$\text{t.p. } (C, A, B) = \text{i.f. } (A + KC)\|((\mathscr{S} + \mathscr{T})/\mathscr{T}), \qquad K \in \mathbf{K}. \tag{15}$$

For future reference, we also note here that if H is defined so that ker HC = ker $C + \mathscr{T} + \mathscr{S}$, then

$$[HC|A + KC] = \mathscr{T} + \mathscr{S}, \qquad K \in \mathbf{K}. \tag{16}$$

For, with H so defined, im $C'H' = \mathscr{T}^\perp \cap \mathscr{S}^\perp \cap$ (im C'), and since $\mathscr{T}^\perp \cap \mathscr{S}^\perp = \{A' + C'K' | \mathscr{T}^\perp \cap \mathscr{S}^\perp \cap (\text{im } C')\rangle$, $K \in \mathbf{K}$, (i.e., $\mathscr{T}^\perp \cap \mathscr{S}^\perp$ is a controllability space), (16) follows by duality.

Proof of Theorem 1. We first construct a polynomial δ and a map $K \in \mathbf{K}^*$ such that

$$\text{i.f.}(A + KC)\|(\mathscr{X}/\mathscr{T}) = \{\delta\alpha_1, \alpha_2, \cdots, \alpha_t\}, \tag{17}$$

where $\{\alpha_i, i \in \mathbf{t}\} =$ t.p. (C, A, B).[5] *For this, let* $K_0 \in \mathbf{K}^*$ be fixed; then im $K_0 \subset \mathscr{T} + A\mathscr{T} + \mathscr{D}$, so $\langle A + K_0C | \mathscr{T} + A\mathscr{T} + \mathscr{D}\rangle = \langle A | \mathscr{T} + A\mathscr{T} + \mathscr{D}\rangle$

[5] If (C, A, B) has no t.p., replace the right side of (17) by $\{\delta\}$.

$= \langle A|\mathcal{T}+\mathcal{D}\rangle \supset \langle A|\mathcal{B}+\mathcal{D}\rangle = \mathcal{X}$. Therefore, if M is the insertion of $\mathcal{T}+A\mathcal{T}+\mathcal{D}$ in $\mathcal{X}$, then $(A+K_0C, M)$ is controllable. If H is defined as in (16) so that $[HC|A+K_0C] = \mathcal{T}+\mathcal{S}$, then the system $(\bar{C}, \bar{A}_0, \bar{M})$ induced in $\bar{\mathcal{X}} = \mathcal{X}/\mathcal{T}$ by $(HC, A+K_0C, M)$ is also controllable; in addition, since $[\bar{C}|\bar{A}_0] = (\mathcal{S}+\mathcal{T})/\mathcal{T}$, (15) implies i.f. $\bar{A}_0|[\bar{C}|\bar{A}_0] = \{\alpha_i, i \in \mathbf{t}\}$. Use Lemma 2 to construct d and a map L so that i.f. $(\bar{A}_0 + \bar{M}L\bar{C}) = \{\delta\alpha_1, \alpha_2, \cdots, \alpha_t\}$. Thus, if $K \equiv K_0 + MLH$, then $(A+KC)\|(\mathcal{X}/\mathcal{T}) = \bar{A}_0 + \bar{M}L\bar{C}$, so (17) is true. In addition, the definition of K implies im $K \subset$ im $K_0 \times$ im $M \subset \mathcal{T}+A\mathcal{T}+\mathcal{D}$, and since $HC\mathcal{T} = 0$ and $K_0 \in \mathbf{K}$, it must be that $K \in \mathbf{K}$. Thus $K \in \mathbf{K}^*$.

It will now be shown that there exists a map F such that

$$d(\langle A+DFC|\mathcal{B}\rangle) \geqq n^*. \tag{18}$$

In view of Proposition 1, it is enough to find a map F_0 such that $d\langle\langle A+KC)|\mathcal{T}+(DF_0-K)C\mathcal{T}\rangle) \geqq n^*$. This in turn will be true if F_0 is selected so that

$$d(\langle\bar{A}|(\bar{D}F_0-\bar{K})C\mathcal{T}\rangle) \geqq n^* - d(\mathcal{T}), \tag{19}$$

where $\bar{A} \equiv (A+KC)\|\bar{\mathcal{X}}$, $\bar{K} \equiv PK$, $\bar{D} \equiv PD$ and $P : \mathcal{X} \to \bar{\mathcal{X}}$ is the canonical projection.

If $r \geqq t$, let $\bar{\mathcal{U}}$ be the zero subspace in $\bar{\mathcal{X}}$; if $r < t$, let $\bar{\mathcal{U}}$ be the sum of the $t-r$ cyclic subspaces of a rational canonical decomposition of $\bar{\mathcal{X}}$ corresponding to the invariant factors $\{\alpha_{r+1}, \cdots, \alpha_t\}$ of $\bar{A}$ as shown in (17). In either case, the definition of n^* implies $n^* = n - d(\bar{\mathcal{U}})$; hence (19) can be rewritten as $d(\langle\bar{A}|(\bar{D}F_0-\bar{K})C\mathcal{T}\rangle \geqq d(\bar{\mathcal{X}}/\bar{\mathcal{U}})$. In $\hat{\mathcal{X}} \equiv \bar{\mathcal{X}}/\bar{\mathcal{U}}$, this inequality is equivalent to

$$\langle\hat{A}|(\hat{D}F_0-\hat{K})C\mathcal{T}\rangle = \hat{\mathcal{X}}, \tag{20}$$

where $\hat{A} \equiv \bar{A}\|\hat{\mathcal{X}}$, $\hat{D} \equiv \hat{P}\bar{D}$, $\hat{K} \equiv \hat{P}\bar{K}$ and $\hat{P} : \bar{\mathcal{X}} \to \hat{\mathcal{X}}$ is the canonical projection. Hence it is enough to select F_0 so that (20) is true.

Since $K \in \mathbf{K}^*$, (9) and the definition of $\mathbf{K}^*$ imply im $K \subset \mathcal{T}+\mathcal{D}+KC\mathcal{T}$; thus $\langle A+KC|\mathcal{T}+\mathcal{D}+KC\mathcal{T}\rangle = \langle A|\mathcal{T}+\mathcal{D}+KC\mathcal{T}\rangle \supset \langle A|\mathcal{B}+\mathcal{D}\rangle = \mathcal{X}$. It follows that $\langle A+KC|\mathcal{T}+\mathcal{D}+KC\mathcal{T}\rangle = \mathcal{X}$, $\langle\bar{A}|\bar{\mathcal{D}}+\bar{K}C\mathcal{T}\rangle = \bar{\mathcal{X}}$, and thus

$$\langle\hat{A}|\hat{\mathcal{D}}+\hat{K}C\mathcal{T}\rangle = \hat{\mathcal{X}}. \tag{21}$$

But the definition of $\bar{\mathcal{U}}$ and (17) imply that $\hat{A}$ has $k \equiv \min\{r, t\}$ invariant factors $\{\delta\alpha_1, \alpha_2, \cdots, \alpha_k\}$; hence, by the main result of [4], there exists a k-dimensional subspace $\hat{\mathcal{W}} \subset \hat{\mathcal{D}} + \hat{K}C\mathcal{T}$ such that

$$\langle\hat{A}|\hat{\mathcal{W}}\rangle = \hat{\mathcal{X}}. \tag{22}$$

By Remark 7, $d(C\mathcal{T}) = r$; since $r \geqq k$, there exist maps $R : C\mathcal{T} \to C\mathcal{T}$, $S : C\mathcal{T} \to \mathcal{V}$ such that $\mathcal{W} =$ im $(\hat{D}S + \hat{K}JR)$, where J is the insertion of $C\mathcal{T}$ in $\mathcal{Y}$. From this and (22), it follows that if I is the identity on $C\mathcal{T}$, then

$$\langle\hat{A}|\text{Im}\,(\hat{D}S + \hat{K}J(R+\mu I))\rangle = \hat{\mathcal{X}} \tag{23}$$

for $\mu = 0$. Thus, by the well-known generic property of controllable pairs over $\mathbb{R}$ [7, Thm. 11, p. 100], (23) remains true for all but a finite set of values of $\mu \in \mathbb{R}$.

Since in addition, $\det[\mu I + R]$ has only a finite number of zeros, there exists a value $\mu = \hat{\mu} \in \mathbb{R}$ for which (23) holds and $[R + \hat{\mu} I]$ is invertible.

Set $F_0 \equiv -S(R+\hat{\mu}I)^{-1}J^{-1}$, where J^{-1} is a left inverse of J. It follows that $(\hat{D}F_0 - \hat{K})C\mathcal{T} = \text{im}\,(\hat{D}S(R+\hat{\mu}I)^{-1}J^{-1} + K)J = \text{im}\,(\hat{D}S + \hat{K}J(R+\hat{\mu}I))$. From this, the choice of $\hat{\mu}$ and (23), it now follows that (20) is true. Hence (18) is correct.

If $r \geqq t$, then $n^* = n$, and (18) clearly implies (5). To complete the proof for the case $r < t$, it is enough to show that for arbitrary F, $d(\langle A + DFC | \mathcal{B} \rangle) \leqq n^*$, but this follows from the definition of n^* and (6) in Theorem 2, which will be proved below.

Remark 4. In both the preceding proof and in the proof of Lemma 2, constructions are used which depend on the fact that $\mathbb{R}$ has infinitely many elements. Thus we cannot assert that Theorem 1 holds for arbitrary fields, and we expect that for finite fields the theorem is false. On the other hand, the proofs of Proposition 1, Lemma 1 and Theorem 2 which follow do not involve field-dependent constructions and are therefore valid for arbitrary fields.

Proof of Theorem 2. Let F be arbitrary, and choose F_0 and $K \in \mathbf{K}^*$ according to Proposition 1 so that (8) holds. If $P : \mathcal{X} \to \mathcal{X}/\mathcal{T}$ is the canonical projection, and $\bar{A} \equiv (A+KC) \| (\mathcal{X}/\mathcal{T})$, then $(A+KC) \| (\mathcal{X}/\mathcal{R}_F)$ is similar to $\bar{A} \| ((\mathcal{X}/\mathcal{T})/(\mathcal{R}_F/\mathcal{T}))$. From this, (8b) and the definition of $\{\rho_i^F, i \in \mathbf{q}^F\}$, it follows that

$$\{\rho_i^F, i \in \mathbf{q}^F\} = \text{i.f. } \bar{A} \| (P\mathcal{X}/P\mathcal{R}_F). \tag{24}$$

If $\{\alpha_i, i \in \mathbf{t}\} \equiv \text{t.p. } (C, A, B)$ and $\{\beta_i, i \in \mu\} \equiv \text{i.f. } \bar{A}$, then from (15) and (i) of Lemma 1 applied to the data $P\mathcal{X}$, $P\mathcal{S}$ and $\bar{A}$, we have

$$t \leqq \mu; \quad \alpha_i | \beta_i, \quad i \in \mathbf{t}. \tag{25}$$

Since (8a) implies $P\mathcal{R}_F = \langle \bar{A} | P(DF_0 - K)C\mathcal{T} \rangle$, $P\mathcal{R}_F$ is generated by a subspace of dimension at most equal to $d(C\mathcal{T})$. Thus, if $\{\gamma_i, i \in \boldsymbol{\sigma}\} \equiv \text{i.f. } \bar{A} | P\mathcal{R}_F$, then $\sigma \leqq d(C\mathcal{T})$. But $d(C\mathcal{T}) = r$ (Remark 7), so $\sigma \leqq r$. Thus, from (25) and the hypothesis $r < t$,

$$\sigma \leqq r < t \leqq \mu. \tag{26}$$

In addition, $\sigma \leqq r$ implies $\beta_{r+i} | \beta_{\sigma+i}$, $i \in \{1, 2, \cdots, t-r\}$, which, with (25), yields

$$\alpha_{r+i} | \beta_{\sigma+i}, \qquad i \in \{1, 2, \cdots, t-r\}. \tag{27}$$

From (26), $\sigma < t$; hence (ii) of Lemma 1 can be applied to the data $P\mathcal{X}$, $P\mathcal{R}_F$ and $\bar{A}$ to obtain,

$$(\mu - \sigma) \leqq q^F; \quad \beta_{\sigma+i} | \rho_i^F, \qquad i \in \{1, 2, \cdots, \mu - \sigma\}. \tag{28}$$

But (26) implies $(t-r) \leqq (\mu - \sigma)$. Thus, from (28) and (27), there follow $(t-r) \leqq q^F$ and $\alpha_{r+i} | \rho_i^F$, $i \in \{1, 2, \cdots, t-r\}$, so (6) is true.

Now, suppose that $d(\mathcal{R}_F) = n^* = n - \deg\,(\prod_{i=r+1}^{t} \alpha_i)$; then $d(\mathcal{X}/\mathcal{R}_F) = \deg\,(\prod_{i=r+1}^{t} \alpha_i)$. Since $d(\mathcal{X}/\mathcal{R}_F) = d(P\mathcal{X}/P\mathcal{R}_F)$, it follows from (24) that

$$\deg\left(\prod_{i=r+1}^{t} \alpha_i\right) = \deg\left(\prod_{i=1}^{q^F} \rho_i^F\right),$$

but this and (6) can be true only if $q^F = t - r$ and $\rho_i^F = \alpha_{r+1}$, $i \in \{1, 2, \cdots, t-r\}$.

2. Smith–McMillan form. As has just been shown, the transmission polynomials of (C, A, B) play an important role in determining the extent to which the pair $(A+DFC, B)$ can be made controllable or stabilizable with F. In this section, we relate the transmission polynomials to familiar invariants of the transfer matrix of (C, A, B).

If I is the identity on $\mathscr{X}$, then (I, A) is clearly observable; it follows from Corollary 1 and duality that there exists a state feedback map F such that $(C, A+BF)$ is observable just in case (C, A, B) is complete; and if (A, B) is controllable, then by Remark 2, (C, A, B) is necessarily complete. In this case, we have the following theorem, which asserts that the transmission polynomials of (C, A, B) coincide with the numerator polynomials of the rational functions appearing in the Smith–McMillan form of $C(\lambda I - A - BF)^{-1}B$.

THEOREM 3.[6] *Let (C, A, B) be a controllable system with nonzero transfer matrix. The set $\mathbf{F} \equiv \{F : (C, A+BF)$ is observable$\}$ is nonempty, and for each $F \in \mathbf{F}$, the Smith–McMillan form of $C(\lambda I - A - BF)^{-1}B$ has the structure PQ^{-1} where $Q = \operatorname{diag}[\beta_1, \cdots, \beta_\mu, 1, \cdots, 1]_{m\times m}$,*

$$P = \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix}_{p\times m},$$

$R = \operatorname{diag}\ [1, \cdots, 1, \alpha_k, \cdots, \alpha_1]_{r\times r}$, $\{\beta_i,\ i \in \mu\} = \text{i.f.}\ (A+BF)$, $\{\alpha_i, i \in \mathbf{k}\} = \text{t.p.}\ (C, A, B)$ *and* $r = \operatorname{rank} C(\lambda I - A - BF)^{-1}B$.

It is, of course, well known [9] that for $F \in \mathbf{F}$, the β_i are the invariant factors of $A+BF$ and r is the rank of $C(\lambda I - A - Bf)^{-1}B$. It is also known [10, Thm. 4.1, p. 111] that for $F \in \mathbf{F}$, the α_i in the Smith–McMillan form are the same as the invariant polynomials (i.p.) of the $\mathbb{R}[\lambda]$-matrix

$$M_F = \begin{bmatrix} \lambda I - A - BF & B \\ C & 0 \end{bmatrix}.$$

Thus the remaining assertion of Theorem 3 (i.e., $\{\alpha_i,\ i \in \mathbf{k}\} = \text{t.p.}\ (C, A, B)$) is a direct consequence of the following lemma.

LEMMA 3. *Let (C, A, B) be fixed. For all F,*

$$\operatorname{rank}(M_F) = n + d(C\mathscr{T}) \tag{29}$$

and

$$\text{i.p.}\ (M_F) = \text{t.p.}\ (C, A, B). \tag{30}$$

Remark 5. Although Lemma 3 holds for arbitrary triples (C, A, B), Theorem 4.1 in [9, p. 111] holds only in the case where $(C, A+BF, B)$ is controllable and observable. If $(C, A+BF, B)$ is not controllable or observable, it is not clear how the list of polynomials $\{\alpha_i, i \in \mathbf{k}\}$ in the Smith–McMillan form of $C(\lambda I - A - BF)^{-1}B$ is related to the transmission polynomials of (C, A, B). Note in particular that if (C, A, B) is incomplete, the number of transmission polynomials of (C, A, B) exceeds the rank of $C(\lambda I - A)^{-1}B$, which in turn is an upper bound for k.

[6] An alternative version of this theorem has recently appeared in [8].

Remark 6. Since the invariant polynomials of an $\mathbb{R}[\lambda]$-matrix are invariant under matrix transposition, it follows from (30) that the transmission polynomials of (C, A, B) are the same as the transmission polynomials of the dual system (B', A', C').

Remark 7. From the matrix identity

$$\begin{bmatrix} I & 0 \\ -C(\lambda I - A)^{-1} & I \end{bmatrix} M_0 = \begin{bmatrix} \lambda I - A & B \\ 0 & -C(\lambda I - A)^{-1}B \end{bmatrix}$$

and (29) (with $F = 0$), it follows that r, the rank of $C(\lambda I - A)^{-1}B$, must satisfy $r = d(C\mathcal{T})$.

The preceding remark together with (29) show that the rank of the matrix $N(\lambda) \equiv M_0$ equals $n + r$ over $\mathbb{R}[\lambda]$. From this, (30) and the structure of the Smith form of $N(\lambda)$, it follows that for all $\mu \in \mathbb{C}$ (= the field of complex numbers), rank $N(\mu) \geqq n + r - t$. Thus (C, A, B) is a complete system (i.e., $r \geqq t$), just in case rank $N(\mu) \geqq n$ for all $\mu \in \mathbb{C}$. But, since $\det[\mu I - A]$ is an nth order minor of $N(\mu)$, rank $N(\mu)$ cannot be less than n, except possibly on the spectrum of A. We are led to the following corollary, which provides an alternative test for the completeness of (C, A, B).

COROLLARY 4. *A triple* (C, A, B), *with* $C(\lambda I - A)^{-1}B \neq 0$, *is complete if and only if*

$$\operatorname{rank} \begin{bmatrix} \mu I - A & B \\ C & 0 \end{bmatrix} \geqq n$$

for all $\mu \in \sigma(A)$.

Proof of Lemma 3. First, note that if $H : \mathcal{Y} \to \mathcal{Y}$, $G : \mathcal{U} \to \mathcal{U}$ and $T : \mathcal{X} \to \mathcal{X}$ are automorphisms, and if $F : \mathcal{X} \to \mathcal{U}$ and $K : \mathcal{Y} \to \mathcal{X}$ are arbitrary, then the matrices

$$Q \equiv \begin{bmatrix} T & -TK \\ 0 & H \end{bmatrix}, \qquad P \equiv \begin{bmatrix} T^{-1} & 0 \\ -FT^{-1} & G \end{bmatrix}$$

are invertible in $\mathbb{R}[\lambda]$, and M_0 is equivalent to the matrix

$$QM_0P = \begin{bmatrix} \lambda I - T(A + BF + KC)T^{-1} & TBG \\ HCT^{-1} & 0 \end{bmatrix}.$$

Hence i.p. (M_F) and rank (M_F) are independent of F. In addition, we can assume that (C, A, B), is in the $\mathfrak{S}^*$-canonical form of [6].[7] Thus, after suitable row and column permutations, M_0 admits the more detailed representation:

$$M_0 = \begin{bmatrix} \lambda I - A_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda I - A_2 & B_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda I - A_3 & 0 & 0 \\ 0 & 0 & 0 & C_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda I - A_4 & B_4 \\ 0 & 0 & 0 & 0 & C_4 & 0 \end{bmatrix},$$

[7] The $\mathfrak{S}^*$-canonical form constructed in [6] is for systems (C, A, B) with C epic and B monic. Extension to the more general case treated here (C_3 and B_2 are no longer required to be full rank matrices) is a simple matter and therefore will not be discussed.

where i.f. (A_1) = t.p. (C, A, B), (A_2, B_2) is controllable, (C_3, A_3) is observable, and (C_4, A_4, B_4) is in prime canonical form. Observe that $\mathcal{M}_0 \equiv$ column span (c.s.) of M_0 over $\mathbb{R}[\lambda]$ is a free submodule of $\mathcal{X} \equiv \mathbb{R}[\lambda]^{n+p}$; in addition, $\mathcal{M}_0 = \mathcal{M}_1 \oplus \mathcal{M}_2 \oplus \mathcal{M}_3 \oplus \mathcal{M}_4$, where $\mathcal{M}_1 =$ c.s. $[\lambda I - A_1]$, $\mathcal{M}_2 =$ c.s. $[\lambda I - A_2, B_2]$, $\mathcal{M}_3 =$ c.s. $[\lambda I - A_3', C_3']'$, and

$$\mathcal{M}_4 = \text{c.s.} \begin{bmatrix} \lambda I - A_4 & B_4 \\ C_4 & 0 \end{bmatrix}.$$

Since (A_2, B_2) is controllable, $[\lambda I - A_2, B_2]$ has Smith form $[I, 0]$ (cf. [9]), implying $\mathcal{M}_2$ is a direct summand of $\mathcal{X}$. By similar reasoning, $\mathcal{M}_3$ is also a direct summand.

Next, observe that since (C_4, A_4, B_4) is in prime canonical form, $C_4(\lambda I - A_4)^{-1}B_4 = \text{diag}[1/\gamma_1, \cdots, 1/\gamma_j]$, where $\{\gamma_i, i \in \mathbf{j}\} =$ i.f. (A_4). It follows that

$$\det \begin{bmatrix} \lambda I - A_4 & B_4 \\ C_4 & 0 \end{bmatrix} = -\det(\lambda I - A_4) \det(C_4(\lambda I - A_4)^{-1}B_4) = -1,$$

so $\mathcal{M}_4$ is the span of an invertible $\mathbb{R}[\lambda]$ matrix and therefore a direct summand of $\mathcal{X}$.

Hence, if we define $n_i =$ size A_i and $\rho =$ rank B_4, then $\mathcal{M}_0 = \mathcal{M}_1 \oplus \hat{\mathcal{M}}$, where $\mathcal{M}_1$ is a free n_1-dimensional submodule with invariant polynomials equal to t.p. (C, A, B), and $\hat{\mathcal{M}} \equiv \mathcal{M}_2 \oplus \mathcal{M}_3 \oplus \mathcal{M}_4$ is an $(n_2 + n_3 + n_4 + \rho)$-dimensional direct summand of $\mathcal{X}$. It thus follows that i.p. $(M_0) =$ i.p. $(\mathcal{M}_1)$ (cf. [10, Chap. X, § 8]) and rank $M_0 = n_1 + n_2 + n_3 + n_4 + \rho$. Thus (30) is true. In addition, $n_1 + n_2 + n_3 + n_4 = n$ and $\rho = d(C\mathcal{T})$ (cf. [6, pp. 358–361]), so (29) follows.

Concluding remarks. The main results of this paper (Theorems 1 and 2) are basic to the analysis of interconnected linear systems and therefore should prove useful in a variety of applications. For example, Theorem 1 provides necessary and sufficient conditions for the existence of a control map F which will make the cascade connection of two systems Σ_1 and Σ_2 (see introduction) controllable.[8] The theorems are also applicable to the analysis of feedback interconnections of Σ_1 and Σ_2, the design of decentralized control systems, and the study of the generic solvability of various control problems [13]. Some of these applications will be treated in a future paper.

Appendix.

Proof of Lemma 1. Without loss of generality, assume m.p. A is a power of an irreducible polynomial π (cf. [14, Ch. VII]). Since m.p. $A \| \mathcal{W}$ and m.p. $A \| (\mathcal{X}/\mathcal{W})$

---

[8] In this case, the conditions can be restated more explicitly in terms of properties of minimal realizations $(\bar{C}, \bar{A}, \bar{B})$ and $(\bar{L}, \hat{A}, \bar{D})$ of Σ_1 and Σ_2 respectively. Since Σ_1 and Σ_2 are assumed to be noninteracting, the set of *transmission divisors* of (C, A, B) (i.e., the elementary divisors determined by t.p. (C, A, B) when the latter is viewed as a list of invariant factors) equals the disjoint union of the set of transmission divisors of $(\bar{C}, \bar{A}, \bar{B})$ together with the set of elementary divisors of $\hat{A}$. From this, Corollary 1 and some lengthy but straightforward combinatorics, it can be shown [12] that if $\{\alpha_i, i \in \mathbf{t}\} \equiv$ t.p. $(\bar{C}, \bar{A}, \bar{B})$ and $\{\beta_i, i \in \mathbf{k}\} \equiv$ i.f. $\hat{A}$, then the cascade connection of Σ_1 and Σ_2 can be made controllable if and only if $k \leqq r$ and g.c.d. $(\beta_i, \alpha_{r+1-i}) = 1$, $i \in \mathbf{r}$, where $\beta_i \equiv \alpha_i \equiv 1$ for $i > k$ and $j > t$. If no transmission zero of Σ_1 (i.e., $(\bar{C}, \bar{A}, \bar{B})$) is a pole of Σ_2 (i.e., an eigenvalue of $\hat{A}$), these conditions can be replaced by the still simpler requirement that the number of invariant factors of $\hat{A}$ be no greater than the rank of the transfer matrix of Σ_1, and in a neighborhood of Σ_1 or Σ_2, all of these conditions are bound to hold.

each divide m.p. A, both m.p. $A|\mathcal{W}$ and m.p. $A\|(\mathcal{X}/\mathcal{W})$ are also powers of π. Thus the invariant factors $\{\sigma_i,\ i\in\mathbf{k}\}$, $\{\mu_i,\ i\in\mathbf{j}\}$ and $\{\beta_i,\ i\in\mathbf{q}\}$ are also the elementary divisors of A, $A|\mathcal{W}$ and $A\|(\mathcal{X}/\mathcal{W})$, respectively. From the Jordan decomposition theorem applied separately to A, $A|\mathcal{W}$ and $A\|(\mathcal{X}/\mathcal{W})$,

$$\begin{aligned}&(k)\ \deg(\pi)=d(\ker\pi(A))\\ &(j)\ \deg(\pi)=d(\ker\pi(A|\mathcal{W}))\\ &(q)\ \deg(\pi)=d(\ker\pi(A\|(\mathcal{X}/\mathcal{W})))\end{aligned}\tag{A.1}$$

(i) The relations $\ker\pi(A|\mathcal{W})=\mathcal{W}\cap(\ker\pi(A))\subset\ker\pi(A)$ and (A.1) imply

$$k\geqq j.\tag{A.2}$$

Since m.p. $A|\mathcal{W}$ divides m.p. A, there follows $\sigma_1|\mu_1$. If $j=1$, the proof of assertion (i) is complete.

Assume $j>1$, and fix $\rho\in\{2,3,\cdots,j\}$. Set $\tilde{\mathcal{X}}=\mathcal{X}_1\oplus\mathcal{X}_2\oplus\cdots\oplus\mathcal{X}_{\rho-1}$ and $\mathcal{V}=\tilde{\mathcal{X}}\cap\mathcal{W}_1\oplus\tilde{\mathcal{X}}\cap\mathcal{W}_2\oplus\cdots\oplus\tilde{\mathcal{X}}\cap\mathcal{W}_\rho$, where $\{\mathcal{X}_i,\ i\in\mathbf{k}\}$ and $\{\mathcal{W}_i,\ i\in\mathbf{j}\}$ are sets of component subspaces of rational canonical decompositions (r.c.d.) of $\mathcal{X}$ and $\mathcal{W}$, respectively. Thus the map $\tilde{A}\equiv A|\tilde{\mathcal{X}}$ has $\tilde{k}=\rho-1$ invariant factors. Since $\tilde{\mathcal{X}}\cap\mathcal{W}_i\subset\mathcal{X}_i$ and $\mathcal{W}_i$ is cyclic, either $\tilde{\mathcal{X}}\cap\mathcal{W}_i=0$, or $\tilde{\mathcal{X}}\cap\mathcal{W}_i$ is cyclic with m.p. equal to a power of π. Therefore the $\tilde{\jmath}$ nonzero component subspaces of $\mathcal{V}$ form an r.c.d. of $\mathcal{V}$, so $\tilde{A}|\mathcal{V}$ has $\tilde{\jmath}$ invariant factors. Since (A.2) is valid for the data $\tilde{\mathcal{X}}$, $\mathcal{V}$, $\tilde{A}$, it follows that $\tilde{k}\geqq\tilde{\jmath}$ or $\tilde{\jmath}\leqq\rho-1$. Hence there exists an integer $\delta\in\boldsymbol{\rho}$ such that $\mathcal{W}_\delta\cap\tilde{\mathcal{X}}=0$. Thus, if $Q:\mathcal{X}\to\mathcal{X}/\tilde{\mathcal{X}}$ is the canonical projection and B the map induced by A in $\mathcal{X}/\mathcal{V}$, then clearly m.p. $A|\mathcal{W}_\delta=$ m.p. $B|Q\mathcal{W}_\delta$ and m.p. $A|\sum_{i=\rho}^{k}\mathcal{H}_i=$ m.p. B. Since m.p. $B|Q\mathcal{W}_\delta$ divides m.p. B and m.p. $A|\sum_{i=\rho}^{k}\mathcal{H}_i=$ m.p. $A|\mathcal{H}_\rho$, it follows that $\mu_\delta|\sigma_\rho$. But $\delta\leqq\rho$, so $\mu_\rho|\sigma_\rho$. Since $\rho\in\{2,3,\cdots,j\}$ is arbitrary and $\mu_1|\sigma_1$, clearly $\mu_i|\sigma_i$, $i\in\mathbf{j}$, so assertion (i) is true.

(ii) If $P:\mathcal{X}\to\mathcal{X}/\mathcal{W}$ is the canonical projection, then $\pi(A\|(\mathcal{X}/\mathcal{W}))P=P\pi(A)$; thus $\ker(\pi(A\|(\mathcal{X}/\mathcal{W})))\supset\ker P\pi(A)$. But $\ker P\pi(A)=(\mathcal{W}+\ker\pi(A))/\mathcal{W}$, which in turn is isomorphic to $(\ker\pi(A))/(\mathcal{W}\cap(\ker\pi(A)))$. Since $\mathcal{W}\cap(\ker\pi(A))=\ker\pi(A|\mathcal{W})$, it follows that $d(\ker\pi(A\|(\mathcal{X}/\mathcal{W})))\geqq d(\ker\pi(A))-\mathrm{d}(\ker\pi(A|\mathcal{W})$. From this and (A.1),

$$q\geqq k-j.\tag{A.3}$$

Assume $k>j$, and fix $\rho\in\{1,2,\cdots,k-j\}$. Set $\hat{\mathcal{X}}=P^{-1}(\bar{\mathcal{W}}_0\oplus\bar{\mathcal{W}}_1\oplus\cdots\oplus\bar{\mathcal{W}}_{\rho-1})$, where $\bar{\mathcal{W}}_0\equiv\mathcal{W}/\mathcal{W}$ and $\{\bar{\mathcal{W}}_i,\ i\in\mathbf{q}\}$ is an r.c.d. of $P\mathcal{X}$ relative to $A\equiv A\|(\mathcal{X}/\mathcal{W})$. The definition of $\hat{\mathcal{X}}$ implies $\mathcal{W}\subset\hat{\mathcal{X}}$ and $P\hat{\mathcal{X}}=\bar{\mathcal{W}}_0\oplus\bar{\mathcal{W}}_1\oplus\cdots\oplus\bar{\mathcal{W}}_{\rho-1}$. Thus $\hat{\mathcal{X}}$ is A-invariant and $A\|(\hat{\mathcal{X}}/\mathcal{W})$ (i.e., $\bar{A}|P\hat{\mathcal{X}}$) has $\rho-1$ invariant factors. Clearly, m.p. $\bar{A}\|(P\mathcal{X}/P\hat{\mathcal{X}})=$ m.p. $\bar{A}|\bar{\mathcal{W}}_\rho$, but since m.p. $\bar{A}|\bar{\mathcal{W}}_\rho=\beta_\rho$ and m.p. $A\|(\mathcal{X}/\hat{\mathcal{X}})=$ m.p. $\bar{A}\|(P\mathcal{X}/P\hat{\mathcal{X}})$,

$$\text{m.p. } A\|(\mathcal{X}/\hat{\mathcal{X}})=\beta_\rho.\tag{A.4}$$

Let $\hat{k}$ denote the number of invariant factors of $\hat{A}\equiv A|\hat{\mathcal{X}}$. Note that the numbers of invariant factors of $\hat{A}|\mathcal{W}$ and $\hat{A}\|(\hat{\mathcal{X}}/\mathcal{W})$ are j and $\rho-1$, respectively.

Since (A.3) is valid for the data $\hat{\mathscr{X}}$, $\mathscr{W}$, $\hat{A}$, it follows that $(\rho-1)\leqq\hat{k}-j$, or

$$\hat{k}\leqq j+\rho-1. \tag{A.5}$$

Set $\hat{\mathscr{W}}=\hat{\mathscr{X}}\cap\mathscr{X}_1\oplus\hat{\mathscr{X}}\cap\mathscr{X}_2\oplus\cdots\oplus\hat{\mathscr{X}}\cap\mathscr{X}_{j+\rho}$. Since $\hat{\mathscr{X}}\cap\mathscr{X}_i\subset\mathscr{X}_i$ and $\mathscr{X}_i$ is cyclic, either $\hat{\mathscr{X}}\cap\mathscr{X}_i=0$, or $\hat{\mathscr{X}}\cap\mathscr{X}_i$ is cyclic with m.p. equal to a power of π. Thus the $\hat{\jmath}$ nonzero component subspaces of $\hat{\mathscr{W}}$ form an r.c.d. of $\hat{\mathscr{W}}$, so that $\hat{A}\,|\,\hat{\mathscr{W}}$ has $\hat{\jmath}$ invariant factors. Since (A.2) is valid for the data $\hat{\mathscr{X}}$, $\hat{\mathscr{W}}$, $\hat{A}$, it follows that $\hat{k}\geqq\hat{\jmath}$, or, from (A.5), $\hat{\jmath}\leqq j+\rho-1$. This and the definition of $\hat{\mathscr{W}}$ ensure the existence of an integer $\delta\in\{1,2,\cdots,j+\rho\}$ such that $\hat{\mathscr{X}}\cap\mathscr{X}_\delta=0$. Clearly, m.p. $A\|((\mathscr{X}_\delta\oplus\hat{\mathscr{X}})/\hat{\mathscr{X}})$ = m.p. $A|\mathscr{X}_\delta=\sigma_\delta$. In addition, since m.p. $A\|((\mathscr{X}_\delta\oplus\hat{\mathscr{X}})/\hat{\mathscr{X}})$ divides m.p. $A\|(\mathscr{X}/\hat{\mathscr{X}})$, it follows from (A.4) that $\sigma_\delta|\beta_\rho$. But $\delta\leqq j+\rho$, so $\sigma_{j+\rho}|\beta_\rho$. As $\rho\in\{1,2,\cdots,k-j\}$ is arbitrary, assertion (ii) is true.

Acknowledgment. The potential application of Theorems 1 and 2 to the cascade connection of systems was suggested by the work of E. J. Davison and S. H. Wang [15].

REFERENCES

[1] F. M. Brasch, Jr. and J. B. Pearson, *Pole placement using dynamic compensators*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 34–43.

[2] C. Y. Ding, F. M. Brasch, Jr. and J. B. Pearson, *On multivariable linear systems*, Ibid., AC-15 (1970), pp. 96–97.

[3] W. M. Wonham, *On pole assignment in multi-input controllable linear systems*, Ibid., AC-12 (1967), pp. 660–665.

[4] M. Heymann, *On the input and output reducibility of multivariable linear systems*, Ibid., AC-15 (1970), pp. 563–570.

[5] J. P. Corfmat and A. S. Morse, *Spectrum assignment with decentralized feedback control*, Proc. 7th Ann. Princeton Conf. on Information Sciences and Systems, Princeton Univ., Princeton, N.J., 1973, pp. 228–231.

[6] A. S. Morse, *Structural invariants of linear multivariable systems*, this Journal, 11 (1973), pp. 446–465.

[7] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[8] B. C. Moore and L. M. Silverman, *A time domain characterization of the invariant factors of a system transfer function*, Proc. 1974 Joint Automatic Control Conf., Univ. of Texas at Austin, pp. 186–193.

[9] R. E. Kalman, *Irreducible realizations and the degree of a rational matrix*, this Journal, 13 (1965), pp. 530–544.

[10] H. H. Rosenbrock, *State-Space and Multivariable Theory*, John Wiley, New York, 1970.

[11] S. MacLane and G. Birkhoff, *Algebra*, Macmillan, New York, 1967.

[12] J. P. Corfmat, *Decentralized control of linear multivariable systems*, Doctoral dissertation, Yale Univ., New Haven, Conn., 1974.

[13] B. Francis, O. A. Sebakhy and W. M. Wonham, *Synthesis of multivariable regulators: the internal model principle*, Appl. Math. and Opt., 1 (1974), no. 1.

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

[15] E. J. Davison and S. H. Wang, *New results on the controllability and observability of general composite systems*, IEEE Trans. Automatic Control, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 1, January 1976

THE SEPARATION PRINCIPLE IN STOCHASTIC CONTROL VIA GIRSANOV SOLUTIONS*

M. H. A. DAVIS†

Abstract. This paper deals with the separation of estimation and control for linear systems with additive Gaussian white noise and nonquadratic cost function. All measurable functions of the observations are admissible as controls, the corresponding solutions being defined by the Girsanov measure transformation. The separation principle is established, under certain conditions, if the dimension of the observation process is equal to that of the state; if there are fewer observations, then additional ones of arbitrarily low signal-to-noise ratio can be adjoined such that there is a separated policy based on the augmented observations which is superior to any policy using the original observations.

1. Introduction. Recently a number of papers, for example [1]–[5], have appeared in which the theory of control of nonlinear systems with additive white noise is developed using the concept of "Girsanov" or "weak" solutions of stochastic differential equations. This allows solutions to be defined for a very large class of control laws, and using it, various existence results and conditions for optimality have been obtained. Mostly, however, the results apply only to the "complete observation" case, where the entire history of the state process is available to the controller. A standard idea, going back at least to [6], for dealing with partially-observable problems is that of the *hyperstate* or *information state.* The conditional distribution of the state represents all the relevant information gained from the observations, and one can therefore, at least in principle, replace the original problem by a completely observable one whose state is the conditional distribution function. One technical problem is, of course, that this is in general infinite-dimensional. There are special cases where it is not, for example, if the state space is a finite set or if the system is linear with Gaussian noise, when all conditional distributions are normal and hence parametrized by the conditional mean and covariance. In this paper, we consider the latter case in order to examine the applicability of the hyperstate idea in the Girsanov solution context. The system dynamics are represented by the stochastic differential equations

$$\begin{aligned} dx_t &= A(t)x_t\,dt + \beta(u(t))\,dt + G(t)\,dw_t^1, \\ dy_t &= F(t)x_t\,dt + R^{1/2}(t)\,dw_t^2, \end{aligned} \tag{1.1}$$

where the control $\{u_t\}$ is to be chosen as a function of the observations $\{y_s, 0 \leqq s \leqq t\}$ so as to minimize a cost criterion of the form

$$J(u) = E\int_0^1 L(t, x_t, u_t)\,dt. \tag{1.2}$$

* Received by the editors June 7, 1974, and in revised form February 3, 1975.

† Department of Computing and Control, Imperial College, London SW7 2BT, England. This work was carried out at Harvard University, where the author was supported by the Joint Services Electronics Program (Contract N00014-67-A-0298-0006) of the U.S. Office of Naval Research.

In (1.1), w^1 and w^2 are vectors of independent Wiener processes so that (1.1) is similar to the Kalman–Bucy filter model except for the control term β. The effect of β will be simply to shift the mean $\hat{x}_t$ of the conditional distribution of x_t, which remains normal with nonrandom variance. Thus the hyperstate is in this case $\hat{x}_t$, and one therefore expects that an optimal controller will first compute $\hat{x}_t$ and then implement a policy based on $\hat{x}_t$, i.e., that the optimal policy will be of the "separated" form $u_t^0 = u^0(t, \hat{x}_t)$. This is the assertion of the separation principle. The principle is most easily established when the cost rate L is quadratic, when an explicit solution to the control problem can be worked out; see [11]. This solution has the additional feature that the function u^0 is the same as in the complete observation case: the optimal control based on $\{x_s, 0 \leqq s \leqq t\}$ is $u^0(t, x_t)$. This feature is known as the *certainty-equivalence principle* and will not hold for more general cost functions.

In [11], Wonham proved the separation principle for general cost functions under certain conditions, notably that x_t and y_t had to be of the same dimension to insure the uniform ellipticity of a certain differential operator. The problem (1.1), (1.2) is replaced by an equivalent one involving the completely observable hyperstate $\hat{x}_t$, and the existence of a solution to the resulting Hamilton–Jacobi equation of dynamic programming is established, thus defining a separated control policy which is easily seen to be optimal. Here we consider the same problem in the framework of Girsanov solutions. Unfortunately, the condition of equal dimension of x_t and y_t seems indispensable to obtaining the existence of an optimal separated policy via the results of [2], [3]. The reason for this is that otherwise, even using the Girsanov method, the existence of solutions to (1.1) is not guaranteed for a sufficiently wide class of separated controls. However, for the case $\dim(y_t) < \dim(x_t)$, we prove the following result: if we allow ourselves some additional observations *of arbitrarily low signal-to-noise ratio*, then there is a separated policy based on the augmented observations whose performance is at least as good as that of any policy based on the original observations. This is almost as good a result as could be desired since the additional observations are, for all practical purposes, just noise. It is perhaps worth remarking that a similar type of argument can be carried through in the converse, less typical, case where $\dim(y_t) > \dim(x_t)$.

2. Problem formulation. Let $(\Omega, \mathcal{F}, P)$ be a probability space carrying four independent separable Brownian motion processes ξ^1, ξ^2, ξ^3, B of dimensions n, $m, n-m, n-m$ respectively (here $m \leqq n$), and a normal random n-vector a which is independent of these processes. Let $\mathcal{F}_t$ be the sub-σ-field of $\mathcal{F}$ generated by $\{a, \xi_s^1, \xi_s^2, \xi_s^3, B_s, 0 \leqq s \leqq t\}$. The stochastic processes x, y are defined for the rest of this paper by

$$\begin{aligned} dx_t &= A(t)x_t\,dt + G(t)\,d\xi_t^1, \qquad x_0 = a, \\ dy_t &= F(t)x_t\,dt + R^{1/2}(t)\,d\xi_t^2, \qquad y_0 = 0. \end{aligned} \tag{2.1}$$

Here $x_t \in R^n$, $y_t \in R^m$ and A, F, G, R are matrices whose elements are piecewise continuous functions of time. G and $R^{1/2}$ are assumed to be nonsingular and F to have rank m for all t. The time-dependence of these matrices will generally not be explicitly expressed. Now let $\mathcal{Y}_t = \sigma\{y_s, 0 \leqq s \leqq t\}$, U be a compact subset of

some Euclidean space and $\beta: U \to R^n$ be a continuous function. The set of admissible controls $\mathscr{U}$ is the set of measurable functions $u:[0,1] \times \Omega \to U$ such that $u(t, \cdot)$ is $\mathscr{Y}_t$-measurable for each $t \in [0,1]$. For $u \in \mathscr{U}$, let P_u be the measure on $(\Omega, \mathscr{F}_1)$ defined by the Radon–Nikodym derivative

$$\rho(u) = \frac{dP_u}{dP} = \exp\left(\int_0^1 (G^{-1}\beta(u_t))' \, d\xi_t^1 - \frac{1}{2}\int_0^1 |G^{-1}\beta(u_t)|^2 \, dt\right), \tag{2.2}$$

where $'$ is the matrix transpose and $|\cdot|$ the Euclidean norm in R^n. By Girsanov's theorem ([1, Lemmas 0, 1, Thm. 1]) P_u is a probability measure and (w^1, w^2) defined as follows are independent Wiener processes under P_u:

$$w_t^1 = \xi_t^1 - \int_0^t G^{-1}\beta(u_s)\, ds,$$

$$w_t^2 = \xi_t^2.$$

Thus the processes $\{x_t, y_t\}$ defined by (2.1) satisfy an equation of the form (1.1) under the new measure P_u. Let $L:[0,1] \times R^n \times U \to R$ be a measurable function satisfying

(i) $0 \leqq L(t,x,u) \leqq K$ for all $(t,x,u) \in [0,1] \times R^n \times U$, for some $K \in R$,

(ii) $L(t,x,\cdot)$ is continuous on U, uniformly in (t,x).

The cost corresponding to $u \in \mathscr{U}$ is

$$J(u) = E_u \int_0^1 L(t, x_t, u_t)\, dt, \tag{2.3}$$

where E_u represents integration with respect to measure P_u.

Let $\hat{x}_t = E_u(x_t|\mathscr{Y}_t)$ and define the process v_t by[1]

$$v_t = y_t - \int_0^t F\hat{x}_t \, dt.$$

This is the *innovations process* which is, as is well known, a Wiener process with respect to $\{\mathscr{Y}_t\}$. Equation (1.1) is similar to the Kalman–Bucy filter model except for the additive control term $\beta(u_t)$, which is, however, adapted to $\mathscr{Y}_t$ and therefore affects the estimation of x_t from $\mathscr{Y}_t$ simply by shifting the conditional mean. This is the content of the following theorem which is proved in [7, Thms. 1 and 2].

THEOREM 1. *The conditional distribution of x_t given $\mathscr{Y}_t$ is normal. The covariance $P_t = E[(x_t - \hat{x}_t)(x_t - \hat{x}_t)'|\mathscr{Y}_t]$ is nonrandom and is the unique solution of the matrix Riccati equation*

$$\dot{P} = AP + PA' + GG' - PF'R^{-1}FP,$$

$$P(0) = \operatorname{cov} a.$$

[1] There exist measurable versions of $\hat{x}_t$ and similar processes introduced below; see [12, Lemma 1].

The mean $\hat{x}_t$ satisfies the stochastic differential equation

$$\begin{aligned} d\hat{x} &= A\hat{x}_t\,dt + \beta(u_t)\,dt + \sigma_t\,dv_t, \\ \hat{x}_0 &= Ea + \bar{a}, \end{aligned} \tag{2.4}$$

where $\sigma_t = P_t F_t'$.

Let $g(\cdot\,; t, x)$ be the n-dimensional normal density function with mean x and covariance P_t. Define the function $\hat{L}$ by

$$\hat{L}(t, x, u) = \int_{R^n} L(t, z, u) g(z; t, x)\,dz.$$

Then for $u \in \mathscr{U}$, the cost $J(u)$ can be expressed as

$$\begin{aligned} J(u) &= E_u \int_0^1 E_u[L(t, x_t, u_t) | \mathscr{Y}_t]\,dt \\ &= E_u \int_0^1 \hat{L}(t, \hat{x}_t, u_t)\,dt. \end{aligned} \tag{2.5}$$

Let $\mathscr{S}$ be the set of measurable functions $v : [0, 1] \times R^n \to U$. The idea of the separation principle is that since the entire conditional distribution of x_t is specified by $\hat{x}_t$, the optimal control should be of the form $u_t = v(t, \hat{x}_t)$ for some $v \in \mathscr{S}$; such control policies will be called "separated". Notice, however, that one cannot define a solution to (1.1) for such policies by standard application of the Girsanov transformation, because the random variable $\hat{x}_t$ depends on $\{u_s, s \leqq t\}$, and so $u(t, \hat{x}_t)$ is not specified *as a function of* $\{y_s, s \leqq t\}$ unless (1.1) already has a solution in the Ito sense. It is therefore necessary to introduce a new definition of the solution of (1.1) for separated policies. This is done by not considering the problem (1.1), (2.3) directly, but switching attention instead to the equivalent problem (2.4), (2.5), which is one of complete observations with state $\hat{x}_t$. (This is effectively what Wonham [11] does). Problems of this type were considered in [3], where an optimal policy was shown to exist under very weak conditions when σ is invertible, which in the present context means that x_t and y_t must have the same dimension (this condition was required, for related reasons, in [11]). If $m < n$, the measure corresponding to the $\hat{x}$ process cannot be directly defined by the Girsanov formula and it is necessary artificially to adjoin some extra observations. This idea is considered in the next section.

3. Control with augmented observations. Let

$$\bar{y}_t = \xi_t^3 \quad \text{and} \quad \tilde{y}_t' = (y_t', \bar{y}_t').$$

Define $\tilde{\mathscr{Y}}_t = \sigma\{\tilde{y}_s, 0 \leqq s \leqq t\}$ and let $\tilde{\mathscr{U}}$ be the set of controls which are adapted to $\tilde{\mathscr{Y}}_t$ (instead of to $\mathscr{Y}_t$ as before). Let $\bar{F}_t$ be an $(n - m) \times m$ matrix with piecewise continuous elements such that $\tilde{F}_t' = [F_t' \vdots \bar{F}_t']$ is nonsingular. For $k = 1, 2, 3, \cdots$ define $\bar{\rho}_k \in L_1(\Omega, \mathscr{F}_1, P)$ by

$$\bar{\rho}_k = \exp\left(k^{-1/2} \int_0^1 (\bar{F}x_t)'\,d\xi_t^3 - \frac{1}{2k} \int_0^1 |\bar{F}x_t|^2\,dt\right).$$

Then for $u \in \mathscr{U}$, the formula

$$\frac{dP_u^k}{dP} = \rho(u)\bar{\rho}_k \tag{3.1}$$

defines a probability measure P_u^k with respect to which $dw^3 = d\xi^3 - (1/\sqrt{k})\bar{F}x\,dt$ is Brownian; thus under P_u^k the processes $(x_t, y_t, \bar{y}_t)$ satisfy

$$\begin{aligned} dx_t &= Ax_t\,dt + \beta(u_t)\,dt + G\,dw_t^1, \\ dy_t &= Fx_t\,dt + R^{1/2}\,dw_t^2, \\ d\bar{y}_t &= (1/\sqrt{k})\bar{F}x_t\,dt + dw_t^3. \end{aligned} \tag{3.2}$$

The cost for $u \in \tilde{\mathscr{U}}$ is given by

$$J^k(u) = E\left(\rho(u)\bar{\rho}_k \int_0^1 L(t, x_t, u_t)\,dt\right).$$

It is clear that for practical purposes the new information $\bar{y}$ is useless for large k; indeed, the covariance of the conditional distribution of x_t given $\tilde{\mathscr{Y}}_t$ is P_t^k satisfying

$$\dot{P}^k = AP^k + P^kA' + GG' - P^k\tilde{F}'(R^k)^{-1}\tilde{F}P^k,$$

where

$$(R^k)^{-1} = \left[\begin{array}{c|c} R^{-1} & 0 \\ \hline 0 & (1/k)I \end{array}\right]$$

so that $P_t^k \to P_t$ as $k \to \infty$. The conditional mean $\hat{x}_t = E_u^k(x_t|\tilde{\mathscr{Y}}_t)$ satisfies, as in (2.4) above,

$$d\hat{x}_t = A\hat{x}_t\,dt + \beta(u_t)\,dt + \sigma_k(t)\,dv_t^k, \tag{3.3}$$

where $\sigma_k = P^k\tilde{F}'$ and

$$dv_t^k = d\tilde{y} - \left[\begin{array}{c} F_t \\ (1/\sqrt{k})\bar{F}_t \end{array}\right]\hat{x}_t\,dt.$$

In this framework, it is possible to calculate directly the cost corresponding to a separated control $u \in \mathscr{S}$. Let B_t be an n-dimensional separable Wiener process on some probability space $(\Omega', \mathscr{A}, \mu)$, and let X_t be the solution of

$$dX_t = AX_t\,dt + \sigma_k(t)\,dB_t, \qquad X_0 = \bar{a}. \tag{3.4}$$

Then for $u \in \mathscr{S}$, a measure μ_u^k is defined by

$$\frac{d\mu_u^k}{d\mu} = \exp\left(\int_0^1 (\sigma_k^{-1}\beta(u(t, X_t)))'\,dB_t - \frac{1}{2}\int_0^1 |\sigma_k^{-1}\beta(u)|^2\,dt\right). \tag{3.5}$$

Under μ_u^k, X_t satisfies

$$dX_t = AX_t\,dt + \beta(u(t, X_t))\,dt + \sigma_k\,d\tilde{B}_t, \tag{3.6}$$

which is of the form of (3.3)[2] with $u_t = u(t, \hat{x}_t)$, and it follows from Lemma 2 of [4] that all solutions of such an equation have the measure given by (3.5) above.[3] Thus the cost corresponding to u is (compare (2.5))

$$M^k(u) = \int_{\Omega'} \frac{d\mu_u^k}{d\mu} \int_0^1 \hat{L}^k(t, X_t, u(t, X_t))\, dt\, d\mu, \tag{3.7}$$

where

$$\hat{L}^k(t, x, u) = \int_{R^n} L(t, z, u) g^k(z; t, x)\, dz,$$

$g^k(\cdot\,; t, x)$ being the normal density with mean x and covariance P_t^k.

Denote by (C^n, B^n) the measurable space of all continuous functions from $[0, 1]$ to R^n with the Borel σ-field. Let μ_X be the measure on (C^n, B^n) generated by $\{X_t\}$, the solution of (3.4), and for any fixed $u \in \tilde{\mathscr{U}}$, let $\mu_{\hat{x}}$ be the measure generated by the process $\hat{x}_t = E_u^k(x_t | \tilde{\mathscr{Y}}_t)$ of (3.3).

LEMMA 1. $\mu_{\hat{x}} \simeq \mu_X$.

Proof. $\{\hat{x}_t\}$ satisfies (3.3), where $\{v_t^k\}$ is a Wiener process with respect to $\tilde{\mathscr{Y}}_t$. Let $\hat{\mathscr{X}}_t = \sigma\{\hat{x}_s, 0 \leqq s \leqq t\}$ and let $\theta_t : C^n \to R^n$ be the function such that

$$\bar{\theta}_t(x) = E_u^k[\beta(u_t) | \hat{\mathscr{X}}_t] \quad \text{a.s.}$$

Now define

$$d\hat{v}_t = \sigma_k^{-1}(d\hat{x}_t - (A\hat{x}_t + \theta_t(\hat{x}))\, dt).$$

Then $(\hat{v}_t, \hat{\mathscr{X}}_t)$ is a Wiener process (this is the standard innovations theorem; see, for example, [8, Lemma 2.1]) and $\hat{x}_t$ satisfies

$$d\hat{x}_t = A\hat{x}_t\, dt + \theta_t(\hat{x})\, dt + \sigma_k\, d\hat{v}_t,$$

where all terms are adapted to $\hat{\mathscr{X}}_t$. It now follows from the Girsanov theorem, since θ is bounded, that $\mu_{\hat{x}} \simeq \mu_X$ with Radom–Nikodym derivative

$$\frac{d\mu_{\hat{x}}}{d\mu_X}(X) = \exp\left(\int_0^1 (\sigma_k^{-1}\theta(X))' \sigma_k^{-1}\, dX - \frac{1}{2}\int_0^1 |\sigma_k^{-1}\theta(X)|^2\, dt \right).$$

THEOREM 2. *There exists* $u^k \in \mathscr{S}$ *such that*

$$M^k(u^k) = \inf_{u \in \mathscr{S}} M^k(u).$$

Furthermore,

$$M^k(u^k) = \inf_{u \in \tilde{\mathscr{U}}} J^k(u). \tag{3.8}$$

---

[2] I.e., (3.6) defines a μ_u^k-Wiener process $\{\tilde{B}_t\}$.

[3] The referee has pointed out that Lemma 2 of [4] was established by appeal to Girsanov's Lemma 7, whose proof is incomplete, but that the special case needed here can be proved by Girsanov's argument. Also, Lemma 2 of [4] needs the stronger hypothesis $|\phi(t, x)| \leqq f(\sup_{0 \leqq s \leqq t} |x_s|)$ rather than $\leqq f(\|x\|)$ as assumed by the authors of [4].

Proof. The first statement is a direct application of Theorem 3 of [3]. One has only to check that for fixed $(t, x, p) \in [0, 1] \times R^n \times R^n$, the Hamiltonian function

$$p'(Ax + \beta(u)) + L(t, x, u)$$

achieves its minimum over $u \in U$. But this is immediate since L and β are continuous in U, and U is compact.

Let $\phi(t, x)$ be the value function [8] for the control problem (3.6), (3.7), i.e., the minimum cost over $[t, 1]$ starting at $X_t = x$. Thus, in particular,

$$\begin{aligned} &\phi(0, \bar{a}) = \inf_{u \in \mathscr{S}} M^k(u), \\ &\phi(1, x) \equiv 0. \end{aligned} \tag{3.9}$$

It is shown in Lemma 6.2, Theorem 6.1 of [8] that there exist measurable functions $\Lambda\phi : [0, 1] \times R^n \to R$ and $\nabla\phi : [0, 1] \times R^n \to R^n$ such that the process $\phi(t, X_t)$ satisfies

$$\phi(t, X_t) - \phi(0, \bar{a}) = \int_0^t \Lambda\phi(s, X_s)\, ds + \int_0^t \nabla\phi(s, X_s)\, dX_s \quad \text{a.s.}\,(\mu_X).$$

The optimality condition states that

$$\Lambda\phi(t, x) + \nabla\phi(t, x)(Ax + \beta(v)) + \hat{L}^k(t, x, v) \geqq 0 \tag{3.10}$$

for almost all $(t, x, v) \in [0, 1] \times R^n \times U$, and $u \in \mathscr{S}$ is optimal if and only if equality holds in (3.10) a.e. for $v = u(t, x)$.

Fix $u \in \tilde{\mathscr{U}}$. From Lemma 1, $\mu_{\hat{x}} \simeq \mu_X$, i.e., these measures have the same null sets, so that the process $\phi(t, \hat{x}_t)$ satisfies

$$\phi(t, \hat{x}_t) - \phi(0, \bar{a}) = \int_0^t \Lambda\phi(s, \hat{x}_s)\, ds + \int_0^t \nabla\phi(s, \hat{x}_s)\, d\hat{x}_s \quad \text{a.s.}\,(\mu_{\hat{x}}).$$

Now $\hat{x}_t$ satisfies (3.3), so that

$$\phi(1, \hat{x}_1) - \phi(0, \bar{a}) = \int_0^1 (\Lambda\phi(s, \hat{x}_s) + \nabla\phi(s, \hat{x}_s)(A\hat{x}_s + \beta(u_s)))\, ds + \int_0^1 \nabla\phi(s, \hat{x}_s)\sigma_k\, dv_s^k.$$

Thus, in view of (3.9) and (3.10),

$$M^k(u^k) = \phi(0, \bar{a}) = -E_u^k \int_0^1 (\Lambda\phi(s, \hat{x}_s) + \nabla\phi(s, \hat{x}_s)(A\hat{x}_s + \beta(u_s)))\, ds$$

$$\leqq E_u^k \int_0^1 \hat{L}^k(t, \hat{x}_s, u_s)\, ds = J^k(u),$$

and consequently,

$$M^k(u^k) \leqq \inf_{u \in \tilde{\mathscr{U}}} J^k(u).$$

To get the opposite inequality, we start with arbitrary $v \in \mathscr{S}$ and $\varepsilon > 0$ and produce a control $u \in \tilde{\mathscr{U}}$ such that

$$|M^k(v) - J^k(u)| < \varepsilon. \tag{3.11}$$

This shows that

$$\inf_{u \in \tilde{\mathscr{U}}} J^k(u) \leqq \inf_{v \in \mathscr{S}} M_k(v) + \varepsilon,$$

which gives the desired result. The control u is constructed by delaying v slightly and then using the fact that (1.1), with the delayed control, always has a solution in the Ito sense. Indeed, let $v \in \mathscr{S}$, fix $\delta > 0$ and define

$$\begin{aligned} v^\delta(t, X_s, s \leqq t) &= v(t - \delta, X_{t-\delta}) \quad \text{for } t \geqq \delta, \\ v^\delta(t, \cdot) &= 0 \qquad\qquad\qquad\quad \text{for } t < \delta \end{aligned}$$

(v^δ is not actually in $\mathscr{S}$, but this causes no problems in terms of the framework of (3.4)–(3.7)). Now let $\{\xi_t, \eta_t\}$ be the solutions on $(\Omega, \mathscr{F}, P)$ of the equations

$$\begin{aligned} d\zeta_t &= A\zeta_t\, dt + \beta(v(t - \delta, \hat{\zeta}_{t-\delta}))\, dt + G\, d\xi_t^1, \\ d\eta_t &= \tilde{F}\zeta_t\, dt + R^{1/2}\, d\xi_t^2, \end{aligned} \tag{3.12}$$

with $\zeta_0 = a$, $\eta_0 = 0$, $v(t, \cdot) = 0$ for $t < 0$ and $\hat{\zeta}_t = E[\zeta_t | \eta_s, 0 \leqq s \leqq t]$. (3.12) has a unique solution, constructed successively on intervals of length δ: for $t \in [0, \delta]$, $\beta(v(t - \delta, \hat{\zeta}_{t-\delta})) = \beta(0)$ so that (3.12) can be solved on $[0, \delta]$; then $\{\hat{\zeta}_s, s \in [0, \delta]\}$ is known and (3.12) can be solved on $[\delta, 2\delta] \cdots$. Now let $u' : [0, 1] \times C^n \to U$ be a function such that

$$v(t - \delta, \hat{\zeta}_{t-\delta}) = u'(t, \{\eta_s, s \leqq t\}) \quad \text{a.s.}$$

and define $u^\delta \in \tilde{\mathscr{U}}$ by

$$u^\delta(t, \omega) = u'(t, \{y_s, s \leqq t\}),$$

where $\{y_t\}$ is defined by (2.1). In view of the uniqueness of the measure given by the Girsanov formula ([4, Lemma 2]),

$$\begin{aligned} J^k(u^\delta) &= E \int_0^1 L(t, \zeta_t, u'_t)\, dt \\ &= E \int_0^1 \hat{L}^k(t, \hat{\zeta}_t, v(t - \delta, \hat{\zeta}_{t-\delta}))\, dt. \end{aligned}$$

On the other hand, writing down the Kalman filter equation corresponding to (3.11), we see that

$$M^k(v^\delta) = E \int_0^1 \hat{L}^k(t, \hat{\zeta}_t, v^\delta)\, dt,$$

so that $M^k(v^\delta) = J^k(u^\delta)$, and (3.12) will be established if it is possible to choose δ such that

$$|M^k(v) - M^k(v^\delta)| < \varepsilon. \tag{3.13}$$

It follows from a standard result of Lebesgue integration ([8, p. 91]) that for almost all $\omega' \in \Omega'$,

$$\int_0^1 |v_t^\delta - v_t|\, dt \to 0, \qquad \delta \to 0, \tag{3.14}$$

and hence, since U is compact, that

$$\int_0^1 |\sigma_k^{-1}(\beta(v^\delta) - \beta(v))|^2\, dt \to 0 \quad \text{a.s.}$$

as $\delta \to 0$. From [9] we can choose a sequence $\delta_n \to 0$ such that

$$P \limsup_n \left[\int_0^1 |\sigma_k^{-1}(\beta(v^{\delta_n}) - \beta(v))|^2\, dt \geqq \frac{1}{2^n} \right] = 0,$$

and for this sequence,

$$\sup_{t\in[0,1]} \int_0^t (\sigma_k^{-1}(\beta(v^{\delta_n}) - \beta(v)))'\, dB \to 0 \quad \text{a.s.}$$

so that

$$\frac{d\mu^k_{v^{\delta_n}}}{d\mu} \to \frac{d\mu^k_v}{d\mu} \quad \text{a.s.} \tag{3.15}$$

By fixing $\omega' \in \Omega'$ and considering a sequence of continuous functions v_t^n converging to $v_t(\omega')$, one can show that

$$\int_0^1 |\hat{L}^k(t, X_t, v_t^\delta) - \hat{L}^k(t, X_t, v_t)|\, dt \to 0 \tag{3.16}$$

a.s. as $\delta \to 0$. Now, since the set of densities

$$\left\{ \frac{d\mu^k_v}{d\mu} : v \in \mathscr{S} \right\}$$

is uniformly integrable ([1, Lemma 1]) and $\hat{L}^k$ is bounded, it follows from (3.7), (3.15) and (3.16) that

$$M^k(v^{\delta_n}) \to M^k(v).$$

This establishes (3.13) and completes the proof.

COROLLARY. *Suppose $n = m$, i.e., that the state and observations are of the same dimension, and that the observation matrix F_t is nonsingular for all $t \in [0, 1]$. Then the separation principle holds.*

As k increases the additional observations $\bar{y}$ get increasingly noisy, so one expects the cost associated with the optimal policies u^k to increase with k. At the same time, the original observations y are always retained, so u^k should approximate the performance of the best control in $\mathscr{U}$ for large k. The following theorem establishes these assertions.

THEOREM 3. (i) *The sequence* $M^k(u^k)$ *is monotone increasing.*
(ii) *Let* $M^* = \lim_k M^k(u^k)$. *Then*

$$M^* = \inf_{u \in \mathcal{U}} J(u).$$

Proof. (i) Fix k and $u \in \tilde{\mathcal{U}}$. We are going to show that there exists $u_1 \in \tilde{\mathcal{U}}$ such that $J^k(u_1) \leqq J^{k+1}(u)$. Then $\inf_{\tilde{\mathcal{U}}} J^k(u) \leqq \inf_{\tilde{\mathcal{U}}} J^{k+1}(u)$, which gives the result in view of (3.8). The idea of the proof is that the observations in case $k+1$ can be regarded as the observations in case k plus an additional independent noise component. We can thus construct a superior k-policy by selecting a better-than-average sample function from the additional noise. Indeed, let $\{B_t\}$ be the $(n-m)$-dimensional Brownian motion specified in § 2 and define

$$z_t = \sqrt{\frac{k}{k+1}}\,\xi_t^3 + \sqrt{\frac{1}{k+1}}\,B_t.$$

Then under measure P_u^k defined by (3.1), z_t satisfies

$$\begin{aligned} dz_t &= \sqrt{\frac{k}{k+1}}\,d\bar{y}_t + \sqrt{\frac{1}{k+1}}\,dB_t \\ &= \frac{1}{\sqrt{k+1}}\bar{F}x\,dt + \left(\sqrt{\frac{k}{k+1}}\,dw^3 + \sqrt{\frac{1}{k+1}}\,dB\right), \end{aligned} \tag{3.17}$$

where w^3, B are independent Brownian motions under P_u^k, so that the bracketed term on the right of (3.17) is itself a standard Brownian motion. The control (u_t), being adapted to $(\tilde{\mathcal{Y}}_t)$, can be regarded as a function of the sample path of $(y, \bar{y})$, i.e., $u_t(\omega) = f(t, y(\omega), \bar{y}(\omega))$. Now construct a new control $\bar{u}$ by replacing $\bar{y}$ by z, giving

$$\bar{u}_t(\omega) = f(t, y(\omega), z(\omega)).$$

In view of (3.2) and (3.17), the sample space measure of (x, y, z) under $P_{\bar{u}}^k$ is the same as that of $(x, y, \bar{y})$ under P_u^{k+1}. Putting

$$\gamma(\bar{u}) = \int_0^1 L(t, x, \bar{u})\,dt,$$

we have

$$J^{k+1}(u) = E_u^k \gamma(\bar{u}).$$

Now let $\mathcal{B} = \sigma\{B_s, 0 \leqq s \leqq t\}$ and $\alpha : C^{n-m} \to R$ be a measurable function such that

$$E_{\bar{u}}^k[\gamma(\bar{u}) | \mathcal{B}] = \alpha(B) \quad \text{a.s.}$$

Then

$$J^{k+1}(u) = E_u^k \alpha(B) = \int_{C^{n-m}} \alpha(B)\mu_w(dB),$$

μ_w being Wiener measure on C^{n-m}. Thus there exists $B^0 \in C^{n-m}$ such that

$$\alpha(B^0) \leqq J^{k+1}(u).$$

Now define

$$z_t^0 = \sqrt{\frac{k}{k+1}}\,\bar{y}_t + \sqrt{\frac{1}{k+1}}\,B_t^0 \tag{3.18}$$

and

$$u_1(t, y, \bar{y}) = u(t, y, z^0).$$

It is clear from (3.17) that the conditional measure of (x, y, z) given is equal to the P_u^k-measure of $(x, y, \bar{y})$ with $\bar{y}$ translated as in (3.18). Thus

$$J^k(u_1) = \alpha(B_0) \leqq J^{k+1}(u).$$

It follows that $M^k(u^k) = \inf_{u \in \tilde{\mathscr{U}}} J^k(u) \leqq \inf_{u \in \tilde{\mathscr{U}}} J^{k+1}(u) = M^{k+1}(u^{k+1})$ as claimed. Since $M^k(u^k) \leqq K$ for all k, there is a least upper bound M^*.

(ii) Here we show that for large k, the cost of a policy $u(t, y, \bar{y}) \in \tilde{\mathscr{U}}$ is close to the cost achieved if $\bar{y}$ is replaced by w^3. The new policy $u(t, y, w^3)$ is just a "randomized" policy in $\mathscr{U}$, and we pick a "good" noise sample function as in (i) above to produce a policy $u_1 \in \mathscr{U}$ whose cost is close to that of $u \in \tilde{\mathscr{U}}$.

Recall that u^k is the optimal policy in $\mathscr{S}$ for case k. Fix $\varepsilon > 0$. In view of (3.8) there exists, for each k, $\tilde{u}^k \in \tilde{\mathscr{U}}$ such that

$$J^k(\tilde{u}^k) \leqq M^* + \varepsilon/4.$$

For any $u \in \tilde{\mathscr{U}}$, the cost in case k is

$$J^k(u) = E(\rho(u)\bar{\rho}_k \gamma(u)).$$

It follows from [1, Lemma 1] that $\{\rho(u)\bar{\rho}_k : k = 1, 2, \cdots, u \in \tilde{\mathscr{U}}\}$ is a uniformly integrable subset of $L_1(\Omega_1, \mathscr{F}_1, P)$. Since L is bounded, the subset

$$\mathscr{H} = \{\rho(u)\bar{\rho}_k(u)\gamma(u) : u \in \tilde{\mathscr{U}}, k = 1, 2, \cdots\}$$

is also uniformly integrable. By the definition of the stochastic integral in [9], there is a subsequence k_n such that

$$P \limsup_n \left[\int_0^1 \frac{1}{k_n} |\bar{F}x|^2 \, dt \geqq \frac{1}{2^n}\right] = 0,$$

and for this subsequence,

$$\sup_t \frac{1}{\sqrt{k_n}} \int_0^t (\bar{F}x)' \, d\xi^3 \to 0 \quad \text{a.s.},$$

so that

$$\bar{\rho}_{k_n} \to 1 \quad \text{a.s.}.$$

Choose δ such that $h \in \mathscr{H} \Rightarrow \int_E h < \frac{1}{4}\varepsilon$ for any $E \in \mathscr{F}_1$ with $PE < \delta$. Now by Egorov's theorem, there exists $E \in \mathscr{F}_1$ such that $PE < \delta$ and $\bar{\rho}_{k_n} \to 1$ uniformly on

$\Omega - E$. Choose $k = k_n$ such that $|\bar{\rho}_k(\omega) - 1| < \varepsilon/(4K)$ for $\omega \in \Omega - E$. Then

$$\left|\int_\Omega \rho(\tilde{u}^k)\gamma(\tilde{u}^k)(\bar{\rho}_k - 1)\right| \leqq \left|\int_E \rho(\tilde{u}^k)\gamma(\tilde{u}^k)(\bar{\rho}_k - 1)\right| + \frac{\varepsilon}{4k}\left|\int_{\Omega-E} \rho(\tilde{u}^k)\gamma(\tilde{u}^k)\right| \tag{3.19}$$

$$\leqq \frac{2\varepsilon}{4} + \frac{\varepsilon}{4K}K = \tfrac{3}{4}\varepsilon.$$

Now $E\rho(\tilde{u}^k)\gamma(\tilde{u}^k)$ is the cost of policy $\tilde{u}^k$ when $\bar{y}$ is replaced by w^3. Let $\mathscr{W} = \sigma\{w_s^3, 0 \leqq s \leqq t\}$ and $\eta : C^{n-m} \to R$ be a function such that $E[\rho(u^k)\gamma(u^k)|\mathscr{W}] = \eta(w^3)$ a.s. Then

$$E\rho(u^k)\gamma(u^k) = \int_{C^n} \eta(w^3)\mu_w(dw^3)$$

and we can choose $w_0 \in C^{n-m}$ such that

$$\eta(w_0) \leqq E\rho(u^k)\gamma(u^k).$$

If we now define $u_2 \in \mathscr{U}$ by

$$u_2(t, y) = u^k(t, y, w_0),$$

then

$$J(u_2) = \eta(\omega) \leqq E\rho(u^k)\gamma(u^k). \tag{3.20}$$

From (3.19) and (3.20),

$$J(u_2) \leqq J(\tilde{u}^k) + \tfrac{3}{4}\varepsilon.$$

Since $J(\tilde{u}^k) \leqq M^* + \frac{1}{4}\varepsilon$, we have $J(u_2) \leqq M^* + \varepsilon$. Hence

$$J^* \triangleq \inf_{u \in \mathscr{U}} J(u) \leqq M^*.$$

For the reverse inequality, observe that

$$\inf_{u \in \tilde{\mathscr{U}}} J^k(u) \leqq J^* \tag{3.21}$$

since $\mathscr{U} < \tilde{\mathscr{U}}$. If $J^* < M^*$, then $J^* < M^k(u^k)$ for some k, which is a contradiction in view of (3.8) and (3.21).

Theorem 3 is our main result: if $m = n$ and F is nonsingular, then there is a policy in $\mathscr{S}$ whose cost is minimal in $\mathscr{U}$. If $m < n$ and F has rank m, then we can augment the observations to achieve the same result while giving the controller negligible additional information. The condition rank $F = m$ is harmless since this property is generic, i.e., can be achieved by arbitrarily small perturbations of the elements of F which are irrelevant from the information-gathering point of view.

REFERENCES

[1] V. E. BENEŠ, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.
[2] M. H. A. DAVIS, *On the existence of optimal policies in stochastic control*, this Journal, 11 (1973), pp. 587–594.
[3] ———, *Optimal control of a degenerate Markovian system*, Recent Mathematical Developments in Control, D. J. Bell, ed., Academic Press, New York, 1973.
[4] T. E. DUNCAN AND P. P. VARAIYA, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.
[5] R. W. RISHEL, *Weak solutions of a partial differential equation of dynamic programming*, this Journal, 9 (1971), pp. 519–528.
[6] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, N.J., 1957.
[7] M. H. A. DAVIS AND P. P. VARAIYA, *Information states for linear stochastic systems*, J. Math. Anal. Appl., 37 (1972), pp. 384–402.
[8] ———, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.
[9] H. P. MCKEAN, *Stochastic Integrals*, Academic Press, New York, 1969.
[10] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, 1968.
[11] W. M. WONHAM, *On the separation theorem of stochastic control*, this Journal, 6 (1968), pp. 312–326.
[12] M. P. ERSHOV, *Representations of Ito processes*, Theor. Probability Appl., 17 (1972), pp. 165–169.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

SINGULAR PERTURBATIONS OF TWO-POINT BOUNDARY VALUE PROBLEMS ARISING IN OPTIMAL CONTROL*

MARVIN I. FREEDMAN AND JAMES L. KAPLAN†

Abstract. This paper considers a two-point boundary value problem which arises from an application of the Pontryagin maximal principle to some underlying optimal control problem. The system depends singularly upon a small parameter, ε. It is assumed that there exists a continuous solution of the system when $\varepsilon = 0$, known as the reduced solution. Conditions are given under which there exists an "outer solution", and "left and right boundary-layer solutions" whose sum constitutes a solution of the system which degenerates uniformly on compact sets to the reduced solution. The principal tool used in the proof is a Banach space implicit function theorem.

1. Introduction. In this paper, we study a two-point boundary value problem which arises by applying the Pontryagin maximal principle to a nonlinear optimal control problem in which a small parameter multiplies derivatives in the state equation. Specifically, we shall be concerned with the system

$$\dot{\xi} = \phi(t, \xi, \chi, \rho, \nu, u, \varepsilon), \tag{1a}$$

$$\dot{\chi} = \pi(t, \xi, \chi, \rho, \nu, u, \varepsilon), \tag{1b}$$

$$\varepsilon\dot{\rho} = \gamma(t, \xi, \chi, \rho, \nu, u, \varepsilon), \tag{1c}$$

$$\varepsilon\dot{\nu} = \psi(t, \xi, \chi, \rho, \nu, u, \varepsilon), \tag{1d}$$

$$0 = H_u(t, \xi, \chi, \rho, \nu, u, \varepsilon), \tag{1e}$$

on the interval $[O, T]$, together with the boundary conditions

$$\xi(0) = a(\varepsilon), \tag{1f}$$

$$\chi(T) = b(\varepsilon), \tag{1g}$$

$$\rho(0) = c(\varepsilon), \tag{1h}$$

$$\nu(T) = d(\varepsilon). \tag{1i}$$

In the above, " $\dot{}$ " denotes d/dt, ε is a small, positive, real parameter, $\xi, \chi, \phi, \pi \in E^{n_1}$, $\rho, \nu, \gamma, \psi \in E^{n_2}$, u is a measurable function on $[0, T]$ with values in E^{n_3}, H is a scalar-valued Hamiltonian function, and H_u denotes the partial derivative of H with respect to u. We may imagine system (1), hereinafter referred to as the *full system*, arising as the result of an application of the maximal principle to some underlying optimal control problem in the variables ξ and ρ. The variables χ and ν may be thought of as the "costate variables" corresponding to ξ and ρ. Alternatively, this may be expressed by saying that $(\chi(t), \nu(t))$ represents the adjoint response to the underlying problem (and hence the reason for the boundary conditions on χ and ν at $t = T$). The function $u(t, \varepsilon)$ appearing in (1a)–(1d) may be viewed, as a consequence of (1e), as the optimal choice for the underlying control problem; that is, it represents that choice of admissible controller which maximizes the Hamiltonian function. (See Lee and Markus [9] for a discussion of the

* Received by the editors May 14, 1974, and in revised form January 5, 1975.
† Department of Mathematics, Boston University, Boston, Massachusetts 02215.

maximal principle. See Freedman and Granoff [5] for an example of a control problem which gives rise to a system of the form (1).).

We present sufficient conditions which enable us to construct an "outer solution," a "left boundary layer solution" and a "right boundary layer solution" whose sum is a solution of the full system. Moreover, as $\varepsilon \to 0+$, this solution will degenerate uniformly on compact subsets of $(0, T)$ to the solution of the system (1) with ε set equal to 0. System (1) with ε set equal to 0 and with (1h) and (1i) omitted is known as the *reduced system.* The reduced system has no boundary conditions related to (1h) and (1i), because the number of derivatives of the full system has been reduced by setting $\varepsilon = 0$. Specifying additional boundary conditions would, in general, overdetermine the reduced system. Thus, unless the reduced solution happens by chance to satisfy $\rho_0(0) = c(0)$, $\nu_0(T) = d(0)$ (which, in general, it will not), we cannot expect that the solution of the full system (1) will converge to the solution of the reduced system as $\varepsilon \to 0+$, especially near the boundary $t = 0$ and $t = T$. This phenomenon is known as boundary layer behavior. It necessitates the construction of the left and right boundary layer solutions which explain the behavior of the full system near $t = 0$ and $t = T$, respectively.

The principal motivation for our approach to the problem is a paper of Hoppensteadt [8] in which he establishes similar results for a singularly perturbed initial value problem not involving a controller. We will rely heavily upon this previous work in § 5. Nevertheless, our work differs from that of Hoppensteadt in several respects. The most obvious differences are that we treat a boundary value problem instead of an initial value problem, and that we accommodate the presence of the additional functional equation (1e). Additionally, our proofs are somewhat more concise, due to a utilization of the Banach space implicit function theorem. This eliminates many of the estimates required by the method of successive approximations employed in [8]. We mention that Fife [16] has previously used a Banach space implicit function theorem in the singular perturbation context.

Wilde and Kokotovic [14] have examined this type of problem for linear systems with quadratic cost functionals.

O'Malley has investigated similar questions for linear and restricted nonlinear systems [11], [12].

A recent paper which treats a problem closer to our own is that of Hadlock [6]. He also treats a singularly perturbed two-point boundary value problem, although it does not involve a controller. Although his existence theorem establishes a solution which degenerates regularly as $\varepsilon \to 0+$, he does not explicitly exhibit the form of either the outer solution, or the left and right boundary layer solutions.

In [13], Sannuti gave an asymptotic analysis for a class of nonlinear problems in which the variables which are multiplied by the small parameter appear linearly. The control appears linearly in the state equations and quadratically in the performance criterion. In this situation, the control can be directly solved for as a linear function of the costate variables.

Finally, in a recent paper, Freedman and Granoff [5] develop formally the asymptotic series for the solution of the full problem whose existence is rigorously established here.

2. Notation and a statement of the problem. It will be convenient for us to introduce the following notation:

$$x = \begin{pmatrix} \xi \\ \chi \end{pmatrix}, \quad y = \begin{pmatrix} \rho \\ \nu \end{pmatrix}, \quad f = \begin{pmatrix} \phi \\ \pi \end{pmatrix} \quad \text{and} \quad g = \begin{pmatrix} \gamma \\ \psi \end{pmatrix},$$

with corresponding notation for x_0 and y_0.

The full system may now be written as

$$\dot{x} = f(t, x, y, u, \varepsilon), \tag{2a}$$

$$\varepsilon\dot{y} = g(t, x, y, u, \varepsilon), \tag{2b}$$

$$0 = H_u(t, x, y, u, \varepsilon), \tag{2c}$$

$$\xi(0) = a(\varepsilon), \tag{2d}$$

$$\chi(T) = b(\varepsilon), \tag{2e}$$

$$\rho(0) = c(\varepsilon), \tag{2f}$$

$$\nu(\mathrm{T}) = d(\varepsilon), \tag{2g}$$

while the reduced system becomes

$$\dot{x}_0 = f(t, x_0, y_0, u_0, 0), \tag{3a}$$

$$0 = g(t, x_0, y_0, u_0, 0), \tag{3b}$$

$$0 = H_u(t, x_0, y_0, u_0, 0), \tag{3c}$$

$$\xi_0(0) = a(0), \tag{3d}$$

$$\chi_0(T) = b(0). \tag{3e}$$

We will assume throughout that there exists a *reduced solution* $x_0(t)$, $y_0(t)$, $u_0(t)$ of system (3).

DEFINITION 2.1. A solution

$$x^*(t, \varepsilon) = \begin{pmatrix} \xi^*(t, \varepsilon) \\ \chi^*(t, \varepsilon) \end{pmatrix}, \qquad y^*(t, \varepsilon) = \begin{pmatrix} \rho^*(t, \varepsilon) \\ \nu^*(t, \varepsilon) \end{pmatrix}, \qquad u^*(t, \varepsilon)$$

which satisfies (2a)–(2c) together with

$$\xi^*(0, \varepsilon) = a^*(\varepsilon), \tag{4a}$$

$$\chi^*(T, \varepsilon) = b^*(\varepsilon), \tag{4b}$$

where $a^*(0) = a(0)$, $b^*(0) = b(0)$, is called an *outer solution of order K* if x^*, y^* and u^* are $K+1$ times continuously differentiable with respect to ε (in particular, $x^*(t, 0) = x_0(t)$, $y^*(t, 0) = y_0(t)$, and $u^*(t, 0) = u_0(t)$). The system (2a)–(2c), (4a), (4b) will be called the *outer system*. We remark that an outer solution is a solution of the differential equations (2a)–(2c), possessing additional smoothness in ε, which is close to the reduced solution and such that no boundary conditions have been imposed on $y^*(t, \varepsilon)$.

Of course, there is no reason to expect that any particular outer solution will also be a solution of the full system (2), as the boundary conditions on $y(t, \varepsilon)$, (2f) and (2g), will not, in general, be satisfied. In order to compensate for this deficiency of the outer solution, the solution of the full system will have to incorporate terms which will correct behavior at the boundary. These terms are known as the left and right boundary layer solutions, respectively.

In order to obtain these corrections, let

$$X = \begin{pmatrix} \xi^L \\ \chi^L \end{pmatrix}, \qquad Y = \begin{pmatrix} \rho^L \\ \nu^L \end{pmatrix},$$

where $\xi^L, \chi^L \in E^{n_1}$ and $\rho^L, \nu^L \in E^{n_2}$, and perform the change of variables

$$X(t, \varepsilon) = x(t, \varepsilon) - x^*(t, \varepsilon),$$

$$Y(t, \varepsilon) = y(t, \cdot\varepsilon) - y^*(t, \varepsilon),$$

$$U(t, \varepsilon) = u(t, \varepsilon) - u^*(t, \varepsilon),$$

$$\tau = \frac{t}{\varepsilon},$$

in (2a), (2b), (2c). This results in a system of the form

$$\frac{dX}{d\tau} = \varepsilon \hat{f}(\varepsilon\tau, X, Y, U, \varepsilon), \tag{5a}$$

$$\frac{dY}{d\tau} = \hat{g}(\varepsilon\tau, X, Y, U, \varepsilon), \tag{5b}$$

$$0 = \hat{H}_u(\varepsilon\tau, X, Y, U, \varepsilon), \tag{5c}$$

where we have used the notation $\hat{f}(\varepsilon\tau, X, Y, U, \varepsilon)$ to mean

$$\hat{f}(\varepsilon\tau, X, Y, U, \varepsilon) = f(\varepsilon\tau, X(\tau, \varepsilon) + x^*(t, \varepsilon), Y(\tau, \varepsilon) + y^*(t, \varepsilon), U(\tau, \varepsilon) + u^*(t, \varepsilon), \varepsilon) - f(\varepsilon\tau, x^*(t, \varepsilon), y^*(t, \varepsilon), u^*(t, \varepsilon), \varepsilon).$$

The function $\hat{g}$ and $\hat{H}_u$ are defined in an analogous manner.

DEFINITION 2.2. A solution $X^L(\tau, \varepsilon)$, $Y^L(\tau, \varepsilon)$, $U^L(\tau, \varepsilon)$ of (5a), (5b), (5c) on $0 \leqq \tau \leqq T/\varepsilon$, and satisfying an initial condition

$$\rho^L(0, \varepsilon) = \hat{c}(\varepsilon), \tag{5d}$$

is called a *left boundary layer solution of order K* if $X^L(\tau, \varepsilon)$, $Y^L(\tau, \varepsilon)$ and $U^L(\tau, \varepsilon)$ are $K+1$ times continuously differentiable with respect to ε. Moreover, we require that there exist some positive constants C, δ such that for all $\tau, 0 \leqq \tau \leqq T/\varepsilon$,

$$|X^L(\tau, \varepsilon)| + |Y^L(\tau, \varepsilon)| + |U^L(\tau, \varepsilon)| \leqq C e^{-\delta\tau}. \tag{6}$$

For now, we will leave the choice of $\hat{c}(\varepsilon)$ in (5d) unspecified. Our solution to this stage consists of the sum

$$(x^*(t, \varepsilon), y^*(t, \varepsilon), u^*(t, \varepsilon)) + (X^L(\tau, \varepsilon), Y^L(\tau, \varepsilon), U^L(\tau, \varepsilon)).$$

Whatever our choice of $\hat{c}(\varepsilon)$ in (5d) (even if it is such that the above sums will

satisfy (2d) and (2f)), we do not expect the right-hand boundary conditions, (2e) and (2g), to be satisfied. This necessitates the inclusion of an additional (right) boundary layer solution, in order to obtain a solution of the full system. Towards this end, let us introduce

$$X(t, \varepsilon) = x(t, \varepsilon) - x^*(t, \varepsilon) - X^L\left(\frac{t}{\varepsilon}, \varepsilon\right)$$

and similar expressions for $Y(t, \varepsilon)$, $U(t, \varepsilon)$ in (2a), (2b), (2c). Setting $\sigma = (T-t)/\varepsilon$ results in the system

$$\frac{dX}{d\sigma} = -\varepsilon \hat{f}(T-\varepsilon\sigma, X, Y, U, \varepsilon), \tag{7a}$$

$$\frac{dY}{d\sigma} = -\hat{g}(T-\varepsilon\sigma, X, Y, U, \varepsilon), \tag{7b}$$

$$0 = \hat{H}_u(T-\varepsilon\sigma, X, Y, U, \varepsilon). \tag{7c}$$

Here we have used the notation

$$\hat{f}(T-\varepsilon\sigma, X, Y, U, \varepsilon) = f(T-\varepsilon\sigma, X+X^L, Y+Y^L, U+U^L, \varepsilon) - f(T-\varepsilon\sigma, X^L, Y^L, U^L, \varepsilon),$$

where, in terms of σ,

$$X^L(\tau, \varepsilon) = X^L(t/\varepsilon, \varepsilon) = X^L(T/\varepsilon - \sigma, \varepsilon), \text{ etc.}$$

The functions $\hat{g}$ and $\hat{H}$ are defined in a similar manner.

DEFINITION 2.3. A solution

$$X^R = \begin{pmatrix} \xi^R(\sigma, \varepsilon) \\ \chi^R(\sigma, \varepsilon) \end{pmatrix}, \qquad Y^R = \begin{pmatrix} \rho^R(\sigma, \varepsilon) \\ \nu^R(\sigma, \varepsilon) \end{pmatrix}, \qquad U^R(\sigma, \varepsilon)$$

of (7a), (7b), (7c) on $0 \leqq \sigma \leqq T/\varepsilon$, and satisfying an initial condition

$$\nu^R(T, \varepsilon) = \hat{d}(\varepsilon), \tag{7d}$$

will be called a *right boundary layer solution of order K* if $X^R(\sigma, \varepsilon)$, $Y^R(\sigma, \varepsilon)$, $U^R(\sigma, \varepsilon)$ are $K+1$ times continuously differentiable with respect to ε, and if there are positive constants C, δ such that for all σ, $0 \leqq \sigma \leqq T/\varepsilon$,

$$|X^R(\sigma, \varepsilon)| + |Y^R(\sigma, \varepsilon)| + |U^R(\sigma, \varepsilon)| \leqq C e^{-\delta\sigma}. \tag{8}$$

In the next section, we will discuss the manner in which one formally constructs the asymptotic series for the outer solution and the left and right boundary layer solutions. In §§ 4 and 5, we will consider the more difficult problem of the selection of $a^*(\varepsilon)$, $b^*(\varepsilon)$, $\hat{c}(\varepsilon)$ and $\hat{d}(\varepsilon)$ so as to ensure the existence of these solutions. Finally, in § 6, we will combine these results to establish our principal theorem (Theorem 6.1). Roughly speaking, we will show

that under suitable hypotheses it is possible to select $a^*(\varepsilon)$, $b^*(\varepsilon)$, $\hat{c}(\varepsilon)$, $\hat{d}(\varepsilon)$ so that the sums

$$x(t,\varepsilon)=x^*(t,\varepsilon)+X^L\left(\frac{t}{\varepsilon},\varepsilon\right)+X^R\left(\frac{T-t}{\varepsilon},\varepsilon\right),$$

$$y(t,\varepsilon)=y^*(t,\varepsilon)+Y^L\left(\frac{t}{\varepsilon},\varepsilon\right)+Y^R\left(\frac{T-t}{\varepsilon},\varepsilon\right),$$

$$u(t,\varepsilon)=u^*(t,\varepsilon)+U^L\left(\frac{t}{\varepsilon},\varepsilon\right)+U^R\left(\frac{T-t}{\varepsilon},\varepsilon\right)$$

will constitute a solution of the full system (2).

3. The formal expansions. In order to show how to construct the formal asymptotic expansions of the outer solution and the left and right boundary layer solutions, it is necessary to make several assumptions concerning our data. Throughout the remainder of the paper, we will assume that the following conditions hold:

(9) the reduced system (3) has a continuous solution $x_0(t)$, $y_0(t)$, $u_0(t)$ on the interval $0\leqq t\leqq T$,

(10a) there exists $\varepsilon_0>0$ such that f, g are $K+2$ times continuously differentiable and H is $K+3$ times continuously differentiable with respect to t, x, y, u and ε, for all $0\leqq t\leqq T$, $0\leqq\varepsilon\leqq\varepsilon_0$ and (x, y, u) in a neighborhood of the reduced solution,

(10b) $a(\varepsilon)$, $b(\varepsilon)$, $c(\varepsilon)$, $d(\varepsilon)$ are $K+2$ times continuously differentiable with respect to ε on $0\leqq\varepsilon\leqq\varepsilon_0$,

(10c) the $n_3\times n_3$ matrix $H_{uu}(t, x(t), y_0(t), u_0(t), 0)$ is invertible for $0\leqq t\leqq T$.

Let us now proceed to develop the formal expansion of the outer solution. We will assume that the functions $a^*(\varepsilon)$, $b^*(\varepsilon)$ appearing in (4) have $K+1$ continuous derivatives with respect to ε and satisfy $a^*(0)=a(0)$, $b^*(0)=b(0)$. We further assume the existence of an outer solution of order K. Such a solution is $K+1$ times continuously differentiable with respect to ε. Under these smoothness hypotheses, the outer solution has a finite Taylor series expansion in powers of ε, given by

$$(x^*(t,\varepsilon), y^*(t,\varepsilon), u^*(t,\varepsilon))=(x_0(t), y_0(t), u_0(t))$$

$$+\sum_{k=1}^{K}(x_k(t), y_k(t), u_k(t))\varepsilon^k+R_1(t,\varepsilon), \tag{11}$$

where $R_1(t,\varepsilon)=O(\varepsilon^{K+1})$. Observe that $(x_0(t), y_0(t), u_0(t))$ in (11) must coincide with the solution of the reduced system.

We may now substitute series (11) into (2a), (2b), (2c). Taking note of smoothness hypothesis (10a), the resulting equations may be differentiated k times with respect to ε, for $1\leqq k\leqq K$. Upon setting $\varepsilon=0$, we obtain

$$\dot{x}_k(t)=f_x(t)x_k(t)+f_y(t)y_k(t)+f_u(t)u_k(t)+p_k(t), \tag{12a}$$

$$\dot{y}_{k-1}(t)=g_x(t)x_k(t)+g_y(t)y_k(t)+g_u(t)u_k(t)+q_k(t), \tag{12b}$$

$$0=H_{ux}(t)x_k(t)+H_{uy}(t)y_k(t)+H_{uu}(t)u_k(t)+s_k(t). \tag{12c}$$

Here we have used the notation

$$f_x(t) = \frac{\partial f}{\partial x}(t, x_0(t), y_0(t), u_0(t), 0),$$

the other partial derivatives being similarly abbreviated. The expressions $p_k(t)$, $q_k(t)$, $s_k(t)$ are polynomials in $x_1(t), \cdots, x_{k-1}(t)$, $y_1(t), \cdots, y_{k-1}(t)$, $u_1(t), \cdots, u_{k-1}(t)$, with coefficients depending on $(t, x_0(t), y_0(t), u_0(t))$. The appropriate boundary conditions for (12a), (12b) and (12c) are readily found to be

$$\xi_k^*(0) = a_k^* = \frac{1}{k!}\frac{d^k}{d\varepsilon^k}(a^*(\varepsilon))|_{\varepsilon=0}, \tag{12d}$$

$$\chi_k^*(T) = b_k^* = \frac{1}{k!}\frac{d^k}{\varepsilon^k}(b^*(\varepsilon))|_{\varepsilon=0}, \tag{12e}$$

for $1 \leqq k \leqq K$. Here we have used a_k^*, b_k^* to denote the coefficient of ε^k in the finite Taylor series expansion of $a^*(\varepsilon)$, $b^*(\varepsilon)$, respectively.

We may now make a crucial observation with regard to system (12). If we assume that $x_0(t), y_0(t), u_0(t), \cdots, x_{k-1}(t), y_{k-1}(t), u_{k-1}(t)$ have been previously determined, then (12b) and (12c) may be regarded as a pair of simultaneous inhomogeneous linear functional equations in the unknown functions $y_k(t)$ and $u_k(t)$. These equations will determine $y_k(t)$ and $u_k(t)$ uniquely whenever the coefficient matrix

$$\begin{pmatrix} g_y(t) & g_u(t) \\ H_{uy}(t) & H_{uu}(t) \end{pmatrix}$$

is nonsingular. Since, in (10c), we assumed that $H_{uu}(t) = H_{uu}(t, x_0(t), y_0(t), u_0(t), 0)$ is invertible on $0 \leqq t \leqq T$, we see that (12b) and (12c) will determine $y_k(t)$ and $u_k(t)$ uniquely whenever the $2n_2 \times 2n_2$ matrix $L(t)$, defined by

$$L(t) = g_y(t) - g_u(t)[H_{uu}(t)]^{-1} H_{uy}(t), \qquad 0 \leqq t \leqq T, \tag{13}$$

is invertible.

Upon substitution of the resulting expressions for $y_k(t)$, $u_k(t)$ into (12a), we are faced with the problem of solving a linear, inhomogeneous, two-point boundary value problem in $x_k(t)$. It is well known that this problem will have a unique solution, provided the corresponding homogeneous problem

$$\dot{\tilde{x}} = M(t)\tilde{x}, \qquad \tilde{\xi}(0) = 0, \qquad \tilde{\chi}(T) = 0 \tag{14}$$

has only the trivial solution. The $2n_1 \times 2n_1$ matrix function $M(t)$ is given explicitly by

$$M(t) = f_x(t) - [f_y(t), f_u(t)]\begin{pmatrix} g_y(t) & g_u(t) \\ H_{uy}(t) & H_{uu}(t) \end{pmatrix}^{-1}\begin{pmatrix} g_x(t) \\ H_{ux}(t) \end{pmatrix}. \tag{15}$$

This discussion suggests the following definition.

DEFINITION 3.1. The outer system (i.e., (2a)–(2c), (4a), (4b)) will be said to be *formally solvable* if

(i) $L(t)$ is invertible for $0 \leqq t \leqq T$, and

(ii) the homogeneous boundary value problem (14) has only the trivial solution.

Our discussion has now established the following result.

THEOREM 3.1. *Let hypotheses* (9) *and* (10) *be satisfied. Then system* (12) *has a unique solution* $x_k(t)$, $y_k(t)$, $u_k(t)$ *for* $1 \leqq k \leqq K$ *provided the outer system is formally solvable.*

We now turn our attention to the problem of determining the form of the left and right boundary layer solutions. Consider first the left boundary layer solution $X^L(\tau, \varepsilon)$, $Y^L(\tau, \varepsilon)$, $U^L(\tau, \varepsilon)$. This solution, if it exists, possesses $K+1$ continuous derivatives with respect to ε. It therefore has a finite Taylor series expansion

$$(16)\quad (X^L(\tau, \varepsilon), Y^L(\tau, \varepsilon), U^L(\tau, \varepsilon)) = \sum_{k=1}^{K} (X_k^L(\tau), Y_k^L(\tau), U_k^L(\tau))\varepsilon^k + R_2(t, \varepsilon),$$

where $R_2(t, \varepsilon) = O(\varepsilon^{K+1})$. Substitution of series (16) into (5) yields, when we set $\varepsilon = 0$,

$$(17a)\qquad \frac{dX_0^L}{d\tau} = 0,$$

$$(17b)\qquad \frac{dY_0^1}{d\tau} = g(0, X_0^L, Y_0^L, U_0^L, 0),$$

$$(17c)\qquad 0 = H_u(0, X_0^L, Y_0^L, U_0^L, 0).$$

The appropriate initial condition for (17a), (17b) and (17c) is found from (5d) to be

$$(17d)\qquad \hat{\rho}_0^L(0) = \hat{c}_0.$$

The equations satisfied by the remaining coefficients are determined by differentiating (5) with respect to ε k times and then setting $\varepsilon = 0$, for $1 \leqq k \leqq K$. This yields

$$(18a)\qquad \frac{dX_k^L}{d\tau} = p_k^L(\tau),$$

$$(18b)\qquad \frac{dY_k^L}{d\tau} = g_x(\tau)X_k^L + g_y(\tau)Y_k^L + g_u(\tau)U_k^L + q_k^L(\tau),$$

$$(18c)\qquad 0 = H_{ux}(\tau)X_k^L + H_{uy}(\tau)Y_k^L + H_{uu}(\tau)U_k^L(\tau) + s_k^L(\tau),$$

with initial condition

$$(18d)\qquad \rho_k^L(0) = \hat{c}_k.$$

In the above, we have used $\hat{g}_x(\tau)$ to denote $(\partial g/\partial X^L)$ $(0, X_0^L(\tau), Y_0^L(\tau), U_0^L(\tau), 0)$; $p_k^L(\tau)$, $q_k^L(\tau)$, and $s_k^L(\tau)$ are polynomials in $X_1^L(\tau), \cdots, X_{k-1}^L(\tau)$, $Y_1^L(\tau), \cdots, Y_{k-1}^L(\tau)$, $U_1^L(\tau), \cdots, U_{k-1}^L(\tau)$, with coefficients depending on $(\tau, X_0^L(\tau), Y_0^L(\tau), U_0^L(\tau))$. Moreover, we see from Definition 2.2 that $p_k^L(\tau)$, $q_k^L(\tau)$

and $s_k^L(\tau)$ decay exponentially as $\tau \to \infty$. The numbers $c_0, c_1, \cdots, c_k$ represent the coefficients in the finite Taylor series expansion of $c(\varepsilon)$.

We remark that condition (8) implies, when we set $\varepsilon = 0$, that

$$|X_0^L(\tau)| + |Y_0^L(\tau)| + |U_0^L(\tau)| \leqq C e^{-\delta\tau} \quad \text{for } 0 \leqq \tau < \infty. \tag{19}$$

Before we can prove a result concerning the formal solvability of systems (17) and (18), we will require a preliminary lemma. It is an adaptation of results of Hartman [7, Chap. IX], which employs topological techniques for the study of the behavior of solutions of an autonomous system in the vicinity of a stationary point. This lemma will enable us to establish the existence of certain locally invariant manifolds from which the left and right boundary layer solutions decay exponentially, as required in (6) and (8), respectively. The proof is omitted. It may be supplied by the reader, or the details may be found in [15].

LEMMA 3.2. *Consider the differential system*

$$\frac{dz}{d\tau} = h(z(\tau)), \qquad z \in E^n, \quad z = (z_1, \cdots, z_n).$$

Let h be twice continuously differentiable, and assume that the Jacobian matrix $A = h_z(0)$ is nonsingular. Suppose that A has k eigenvalues, $1 \leqq k \leqq n$, with negative real part, and $n - k$ with positive real part. Let P be a $2n_2 \times 2n_2$ nonsingular matrix such that

$$P^{-1}AP = \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix},$$

where B is a $k \times k$ matrix whose eigenvalues all have negative real part, while the eigenvalues of C have positive real part. Suppose, further, that

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix},$$

where the $k \times k$ matrix P_{11} satisfies $\det P_{11} \neq 0$. Then there exists $\delta > 0$ such that if $(\alpha_1, \alpha_2, \cdots, \alpha_k) \in E^k$, with $|\alpha_i| < \delta$, then there exists a solution $z(\tau)$ of (21) satisfying $z_i(0) = \alpha_i$, $1 \leqq i \leqq k$, which decays exponentially to zero as $\tau \to \infty$.

In order to be able to apply Lemma 3.2, we are forced to make an additional technical assumption on the matrix $L(t)$ in (13). We require the following:

$$\begin{aligned} &\text{for each } t,\ 0 \leqq t \leqq T,\ L(t) \text{ has } n_2 \text{ eigenvalues } \lambda_1(t), \lambda_2(t), \cdots, \lambda_{n_2}(t), \text{ with} \\ &\text{Re } \lambda_i(t) \leqq -\gamma < 0, \quad i = 1, 2, \cdots, n_2, \quad \text{and} \quad n_2 \quad \text{eigenvalues} \\ &\lambda_{n_2+1}(t), \cdots, \lambda_{2n_2}(t) \text{ with Re } \lambda_i(t) \geqq \gamma > 0,\ i = n_2 + 1, \cdots, 2n_2. \end{aligned} \tag{20}$$

Remark 3.1. It is well known that under the smoothness assumptions of this paper, eigenvalue condition (20) is equivalent to the existence of

(i) an $n_2 \times n_2$ matrix $B(t)$, $0 \leqq t \leqq T$, with all eigenvalues having negative real parts $\leqq -\gamma < 0$;

(ii) an $n_2 \times n_2$ matrix $C(t)$, $0 \leqq t \leqq T$, with all eigenvalues having positive real parts $\geqq \gamma > 0$; and

(iii) a $2n_2 \times 2n_2$ continuously differentiable matrix $P(t)$ such that

$$P^{-1}(t)L(t)P(t) = \begin{pmatrix} B(t) & 0 \\ 0 & C(t) \end{pmatrix}.$$

Hypotheses essentially the same as (20) have previously been used by numerous other authors: Flatto and Levinson [4], Hadlock [6], Hoppensteadt [8], Levin [10], Chang [1] and Chang and Coppel [2].

We are now prepared to prove the following result.

THEOREM 3.3. *Suppose that hypotheses* (9) *and* (10) *hold. Assume, further, that* $L(t)$ *is invertible and that eigenvalue condition* (20) *holds. Assume that the matrix* $P(t)$ *of Remark* 3.1 (iii) *satisfies* $\det P_{11}(0) \neq 0$, *where*

$$P(t) = \begin{pmatrix} P_{11}(t) & P_{12}(t) \\ P_{21}(t) & P_{22}(t) \end{pmatrix}.$$

Then, for sufficiently small vectors $\mu, \eta \in E^{n_2}$, *the following are true*:

(i) *System* (17) *has a unique solution* $X_0^L(\tau)$, $Y_0^L(\tau)$, $U_0^L(\tau)$ *satisfying* $\rho_0^L(0) = \mu$, *and the additional condition*

$$\lim_{\tau\to\infty} X_0^L(\tau) = \lim_{\tau\to\infty} Y_0^L(\tau) = \lim_{\tau\to\infty} U_0^L(\tau) = 0. \tag{21}$$

Moreover, $X_0^L(\tau) = 0$ *and* $Y_0^L(\tau)$, $U_0^L(\tau)$ *decay exponentially.*

(ii) *For each* k, $1 \leqq k \leqq K$, *system* (18) *has a unique solution* $X_k^L(\tau)$, $Y_k^L(\tau)$, $U_k^L(\tau)$ *satisfying* $\rho_k^L(0) = \eta$ *and*

$$\lim_{\tau\to\infty} X_k^L(\tau) = \lim_{\tau\to\infty} Y_k^L(\tau) = \lim_{\tau\to\infty} U_k^L(\tau) = 0. \tag{22}$$

Moreover, the solution $X_k^L(\tau)$, $Y_k^L(\tau)$, $U_k^L(\tau)$ *decays exponentially.*

Remark 3.2. Under suitable conditions, as in O'Malley [12], the smallness conditions on μ and η may be eliminated. However, in the general nonlinear context with which we are dealing, those conditions are necessary and related to the general problem of estimation of the size of the region of attraction.

Proof of (i). We first note that (17a), together with the condition $\lim_{\tau\to\infty} X_0^L(\tau) = 0$, imply that

$$X_0^L(\tau) \equiv 0. \tag{23a}$$

Utilizing (23a), we reduce (17b) and (17c) to:

$$\begin{aligned} \frac{dY_0^L(\tau)}{d\tau} &= \hat{g}(0, 0, Y_0^L(\tau), U_0^L(\tau), 0) \\ &= g(0, x_0(0), y_0(0) + Y_0^L(\tau), u_0(0) + U_0^L(\tau), 0) \\ &\quad - g(0, x_0(0), y_0(0), u_0(0), 0), \end{aligned} \tag{23b}$$

$$0 = \hat{H}_u(0, 0, Y_0^L(\tau), U_0^L(\tau), 0) = H_u(0, x_0(0), y_0(0) + Y_0^L(\tau), u_0(0) + U_0^L(\tau), 0). \tag{23c}$$

In (23c) we have used the fact that $H_u(0, x_0(0), y_0(0), u_0(0), 0) = 0$. In addition to this, (10c) implies that $\det H_{uu}(0, x_0(0), y_0(0), u_0(0), 0) \neq 0$. By the standard implicit function theorem in Euclidean space, it follows that (23c) may be solved for $U_0^L(\tau)$ in terms of $Y_0^L(\tau)$ in some sufficiently small neighborhood of 0 in E^{2n_2}. More precisely, there exists a continuously differentiable map Ψ from a neighborhood of 0 in E^{2n_2} into E^{n_3} satisfying

$$\Psi(0) = 0,$$

$$\Psi_{y(0)} = -H_{uu}^{-1}(0)[H_{uy}(0)],$$

and

$$0 = H_u(0, x_0(0), y_0(0)+\alpha, u_0(0)+\Psi(\alpha), 0)$$

for any α in this neighborhood.

It will now suffice to prove that the equation

$$\frac{dY_0^L(\tau)}{d\tau} = g(0, x_0(0), y_0(0)+Y_0^L(\tau), u_0(0)+\Psi(Y_0^L(\tau)), 0) - g(0, x_0(0), y_0(0), u_0(0), 0) \tag{24}$$

has a unique solution $Y_0^L(\tau)$ which decays exponentially and which satisfies $\rho_0^L(0) = \mu$ for any sufficiently small $\mu \in E^{n_2}$.

Now, define

$$h(Y_0^L(\tau)) = g(0, x_0(0), y_0(0)+Y_0^L(\tau), u_0(0)+\Psi(Y_0^L(\tau)), 0) - g(0, x_0(0), y_0(0), u_0(0), 0).$$

We easily compute

$$\left.\frac{dh}{dY_0^L}\right|_{Y_0^L(\tau)=0} = g_y(0)+g_u(0)[-H_{uu}(0)]^{-1}H_{uy}(0) = L(0).$$

Thus (24) satisfies the requirements of Lemma 3.2. The conclusion of that lemma completes the proof of Theorem 3.3 (i).

Proof of (ii). Consider now system (18). By induction, it is easy to establish that for each k, $1 \leqq k \leqq K$, the function $p_k^L(\tau)$ decays exponentially as $\tau \to \infty$. It then follows that the only solution of (18a) which satisfies $\lim_{\tau\to\infty} X_k^L(\tau) = 0$ must be given by

$$X_k^L(\tau) = -\int_\tau^\infty p_k^L(s)\, ds.$$

Equation (18c) may be written as

$$U_k^L(\tau) = -[\hat{H}_{uu}(\tau)]^{-1}[\hat{H}_{ux}(\tau)X_k^L + H_{uy}(\tau)y_k^L + s_k^L(\tau)]. \tag{25}$$

From this, we see that it will suffice for us to show that if $\eta \in E^{n_2}$ is sufficiently small, then the solution $Y_k^L(\tau)$ of (18b) satisfying $\rho_k^L(0) = \eta$ decays exponentially.

Towards this end, let us insert expression (25) into (18b) to obtain

$$\begin{aligned}\frac{dY_k^L}{d\tau} &= (\hat{g}_x(\tau) - \hat{g}_u(\tau)[\hat{H}_{uu}(\tau)]^{-1}[\hat{H}_{ux}(\tau)])X_k^L \\ &\quad + (\hat{g}_y(\tau) - \hat{g}_u(\tau)[\hat{H}_{uu}(\tau)]^{-1}[\hat{H}_{uy}(\tau)])Y_k^L \\ &\quad + q_k^L(\tau) - \hat{g}_u(\tau)[\hat{H}_{uu}(\tau)]^{-1}s_k(\tau).\end{aligned}$$

If we now utilize the fact that $X_k^L(\tau)$ is a known function of τ which decays exponentially, we see that this equation is a linear, inhomogeneous differential equation of the form

$$\frac{dY_k^L}{d\tau} = A(\tau)Y_k^L(\tau) + B(\tau), \tag{26}$$

where $A(\tau)$, $B(\tau)$ are known functions. The function $B(\tau)$ decays exponentially as $\tau \to \infty$, due to the presence of the terms $X_k^L(\tau)$, $q_k^L(\tau)$, $s_k^L(\tau)$, all of which decay exponentially. (The decay of $q_k^L(\tau)$, $s_k^L(\tau)$ is established by induction.) On the other hand, the matrix $A(\tau)$ is given by

$$A(\tau) = \hat{g}_y(\tau) - \hat{g}_u(\tau)[\hat{H}_{uu}(\tau)]^{-1}[\hat{H}_{uy}(\tau)],$$

where $\hat{g}_y(\tau) = g_y(0, x_0(0), y_0(0) + Y_0^L(\tau), u_0(0) + U_0^L(\tau), 0)$, and $\hat{g}_u$, $\hat{H}_{uu}$ and $\hat{H}_{uy}$ are similarly defined. In part (i) of this theorem, we established the fact that if $\alpha = \rho_0^L(0)$ is sufficiently small, then $Y_0^L(\tau)$ (and hence $U_0^L(\tau) = \Psi(Y_0^L(\tau))$) decay exponentially. Combining this with the continuous dependence of solutions upon initial conditions, if α is sufficiently small, we may ensure that $Y_0^L(\tau)$, $U_0^L(\tau)$ will be uniformly small on $[0, \infty)$. It now follows, by choosing α sufficiently small, that we can make $\|A(\tau) - L(0)\|$ as small as we desire for all $\tau \in [0, \infty)$. Moreover, $A(\tau) - L(0)$ decays exponentially. By rewriting (26) as

$$\frac{dY_k^L}{d\tau} = L(0)\, Y_k^L(\tau) + [A(\tau) - L(0)]\, Y_k^L(\tau) + B(\tau),$$

we may now therefore view it as a perturbation of the linear, homogeneous, constant coefficient equation

$$\frac{dz}{d\tau} = L(0)z(\tau). \tag{27}$$

The proof that $Y_k^L(\tau)$ decays exponentially for any choice of $\eta = \rho_k^L(0)$ in a suitably small neighborhood of the origin will now follow directly from an application of the following lemma.

LEMMA 3.4. *Consider the linear system $\dot{y} = Ay + f(t, y)$, where f is continuous for $|y|$ small and $t \geqq 0$. We assume that, given $\alpha > 0$, there exist β, T such that for all $t \geqq 0$,*

$$|f(t, x_1) - f(t, x_2)| \leqq \alpha |x_1 - x_2| \quad \text{for } |x_1|, |x_2| \leqq \beta.$$

Let A have k eigenvalues with negative real part and $n - k$ with positive real part. Then for any t_0 there exists a real k-dimensional manifold S containing the origin such that any solution ϕ of $\dot{y} = Ay + f(t, y)$ with $\phi(t_0) \in S$ satisfies $\phi(t) \to 0$ exponentially as $t \to \infty$.

Lemma 3.4 appears in Coddington and Levinson [3, Thm. 41, p. 330]. A careful examination of the proof will reveal that it is not necessary that for any $\alpha > 0$ there exist δ such that $|f(t, x_1) - f(t, x_2)| \leqq \alpha |x_1 - x_2|$ for $|x_1|, |x_2| \leqq \beta$, but merely that it hold for β sufficiently small, which is the case here. Moreover, the change of variables $y = Pz$ in (27) shows that the k-dimensional manifold whose existence is assured consists of the first k components of y. Since $\det P_{11}(0) \neq 0$, an implicit function theorem argument allows us to choose the first k components of z in some suitably small neighborhood of 0.

Let us now turn our attention to the determination of the right boundary layer solution of order K. In a manner paralleling the previous procedure, we find that such a right boundary layer solution $X^R(\sigma, \varepsilon)$, $Y^R(\sigma, \varepsilon)$, $U^R(\sigma, \varepsilon)$, which is

$K+1$ times continuously differentiable with respect to ε, will of necessity satisfy

$$dX_0^R/d\sigma = 0, \tag{28a}$$

$$dY_0^R/d\sigma = \hat{g}(T, X_0^R, Y_0^R, U_0^R, 0), \tag{28b}$$

$$0 = \hat{H}_u(T, X_0^R, Y_0^R, U_0^R, 0), \tag{28c}$$

together with initial condition

$$\nu_0^R(0) = \hat{d}_0. \tag{28d}$$

For $1 \leqq k \leqq K$ we have

$$dX_k^R/d\sigma = p_k^R(\sigma), \tag{29a}$$

$$dY_k^R/d\sigma = -\hat{g}_x(\sigma)X_k^R - \hat{g}_y(\sigma)Y_k^R - \hat{g}_u(\sigma)u_k^R + q_k^R(\sigma), \tag{29b}$$

$$0 = \hat{H}_{ux}(\sigma)X_k^R(\sigma) + \hat{H}_{uy}(\sigma)Y_k^R(\sigma) + \hat{H}_{uu}(\sigma)U_k^R(\sigma) + s_k^R(\sigma), \tag{29c}$$

with initial condition

$$\nu_k^R(0) = \hat{d}_k. \tag{29d}$$

In the above equations, X_k^R, Y_k^R, U_k^R, $\hat{d}_k$, $0 \leqq k \leqq K$, denote the coefficients of the finite Taylor series expansions in powers of ε of X^R, Y^R, U^R, $\hat{d}$, respectively. The symbol $\hat{g}_x(\sigma)$ denotes $(\partial g/\partial X^R)(T, X_0^R(\sigma), Y_0^R(\sigma), U_0^R(\sigma), 0)$, the other partial derivatives being similarly defined. We require, of course, that

$$\lim_{\sigma\to\infty} X^R(\sigma, \varepsilon) = \lim_{\sigma\to\infty} Y^R(\sigma, \varepsilon) = \lim_{\sigma\to\infty} U^R(\sigma, \varepsilon) = 0,$$

which is equivalent to requiring that

$$\lim_{\varepsilon\to 0} X^R\left(\frac{T-t}{\varepsilon}, \varepsilon\right) = \lim_{\varepsilon\to 0} Y^R\left(\frac{T-t}{\varepsilon}, \varepsilon\right) = \lim_{\varepsilon\to 0} U^R\left(\frac{T-t}{\varepsilon}, \varepsilon\right) = 0.$$

Note that, by induction, these conditions imply the exponential decay of $p_k^R(\sigma)$, $q_k^R(\sigma)$, $s_k^R(\sigma)$ as $\sigma \to \infty$.

By analogy with Theorem 3.3, we now have the following.

THEOREM 3.5. *Suppose that hypotheses* (9) *and* (10) *hold. Assume, further, that* $L(t)$ *is invertible and that eigenvalue condition* (20) *holds. Assume that the matrix* $P(t)$ *of Remark* 3.1 (iii) *satisfies* $\det P_{22}(T) \neq 0$, *where*

$$P(t) = \begin{pmatrix} P_{11}(t) & P_{12}(t) \\ P_{21}(t) & P_{22}(t) \end{pmatrix}.$$

Then, for sufficiently small vectors $\alpha, \eta \in E^{n_2}$, *the following are true*:

(i) *System* (28) *has a unique solution* $X_0^R(\sigma)$, $Y_0^R(\sigma)$, $U_0^R(\sigma)$ *satisfying* $\nu_0^R(0) = \alpha$ *and the additional conditions*

$$\lim_{\sigma\to\infty} X_0^R(\sigma) = \lim_{\sigma\to\infty} Y_0^R(\sigma) = \lim_{\sigma\to\infty} U_0^R(\sigma) = 0.$$

Moreover, $X_0^R(\sigma) \equiv 0$ *and* Y_0^R, U_0^R *decays exponentially.*

(ii) *For each* k, $1 \leqq k \leqq K$, *the system* (29) *has a unique solution* $X_k^R(\sigma)$, $Y_k^R(\sigma)$, $U_k^R(\sigma)$ *satisfying* $\nu_k^R(0) = \eta$ *and*

$$\lim_{\sigma \to \infty} X_k^R(\sigma) = \lim_{\sigma \to \infty} Y_k^R(\sigma) = \lim_{\sigma \to \infty} U_k^R(\sigma) = 0.$$

Moreover, the solution X_k^R, Y_k^R, U_k^R *decays exponentially.*

In this section, we have shown how one formally constructs an outer solution, and a left and right boundary layer solution. We have deliberately avoided, to this point, any discussion of how one might select $a^*(\varepsilon)$, $b^*(\varepsilon)$, $\hat{c}(\varepsilon)$, $\hat{d}(\varepsilon)$ so that

$$(x^*(t, \varepsilon) + X^L(\tau, \varepsilon) + X^R(\sigma, \varepsilon),\ y^*(t, \varepsilon) + Y^L(\tau, \varepsilon) + Y^R(\sigma, \varepsilon),\ u^*(t, \varepsilon) + U^L(\tau, \varepsilon) + U^R(\sigma, \varepsilon))$$

will constitute a solution of the full system. In the following sections, we will show that such a selection is possible.

4. Existence of an outer solution. In this section, we shall see that hypotheses (9) and (10), plus the assumption of formal solvability of the outer system, appear insufficient for a rigorous proof of the existence of outer solutions. We will again require the use of eigenvalue condition (20).

THEOREM 4.1. *Suppose that* (9), (10), (20) *hold, and the outer system is formally solvable. Let* $a^*(\varepsilon)$, $b^*(\varepsilon) \in E^{n_1}$ *be as in Definition* 2.1, *i.e.,* $a^*(0) = a(0)$, $b^*(0) = b(0)$. *Suppose, further, that* $a^*(\varepsilon)$, $b^*(\varepsilon)$ *are* $K+1$ *times continuously differentiable on* $0 \leqq \varepsilon < \varepsilon_0$. *Then there exists* $\varepsilon_1 > 0$ *such that for all* ε, $0 \leqq \varepsilon < \varepsilon_1$, *the outer system possesses a solution* $x^*(t, \varepsilon)$, $y^*(t, \varepsilon)$, $u^*(t, \varepsilon)$ *satisfying*

$$\begin{aligned} x^*(t, \varepsilon) - \sum_{k=0}^{K} x_k(t)\varepsilon^k &= O(\varepsilon^{K+1}), \\ y^*(t, \varepsilon) - \sum_{k=0}^{K} y_k(t)\varepsilon^k &= O(\varepsilon^{K+1}), \\ u^*(t, \varepsilon) - \sum_{k=0}^{K} u_k(t)\varepsilon^k &= O(\varepsilon^{K+1}), \end{aligned} \tag{30}$$

where the $O(\varepsilon^{K+1})$ *is taken to hold uniformly for* $0 \leqq t \leqq T$.

Proof of Theorem 4.1. We will establish our result under the assumption that $L(t)$ is in block diagonal form, i.e., $L(t) = \operatorname{diag}[B(t), C(t)]$, where $B(t)$ is an $n_2 \times n_2$ matrix having all eigenvalues with real parts $\leqq -\gamma < 0$, and $C(t)$ is an $n_2 \times n_2$ matrix whose eigenvalues all have real parts $\geqq \gamma > 0$. Upon completion of the proof in this case, we shall show how the general case may be reduced to this one.

We consider first the case $K = 0$. Let $(r(t), s(t), v(t))$ lie in a suitably small neighborhood of $(0, 0, 0)$ in $E^{2n_1} \times E^{2n_2} \times E^{n_3}$ for $0 \leqq t \leqq T$. For $0 \leqq \varepsilon < \varepsilon_o$, define

$$\mathscr{F}(t, r, s, v, \varepsilon) = \begin{cases} \dfrac{f(t, x_0(t) + \varepsilon r(t), y_0(t) + \varepsilon s(t), u_0(t) + \varepsilon v(t), \varepsilon) - f(t, x_0(t), y_0(t), 0)}{\varepsilon} & \text{for } \varepsilon \neq 0, \\ f_x(t)r(t) + f_y(t)s(t) + f_u(t)v(t) + f_\varepsilon(t) & \text{for } \varepsilon = 0. \end{cases}$$

In a similar manner, define $\mathcal{G}(t, r, s, v, \varepsilon)$ and $\mathcal{H}_u(t, r, s, v, \varepsilon)$. Then $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}_u$ are continuous at $\varepsilon = 0+$.

Also, define

$$A^*(\varepsilon) = \begin{cases} \dfrac{a^*(\varepsilon) - a(0)}{\varepsilon}, & \varepsilon \neq 0, \\ a_1^*, & \varepsilon = 0, \end{cases} \qquad B^*(\varepsilon) = \begin{cases} \dfrac{b^*(\varepsilon) - b(0)}{\varepsilon}, & \varepsilon \neq 0, \\ b_1^*, & \varepsilon = 0. \end{cases}$$

It will be convenient for us to adopt the following notational convention throughout the remainder of the proof. If w denotes any variable with values in E^{2n}, for some integer n, then

$$w = \begin{pmatrix} w^1 \\ w^2 \end{pmatrix},$$

where w^1 and w^2 denote the first n and the last n components, respectively, of w.

It is readily verified that any bounded, continuous solution $(\alpha(t, \varepsilon), \beta(t, \varepsilon), \gamma(t, \varepsilon))$ of the system

$$\dot{\alpha} = \mathcal{F}(t, \alpha, \beta, \gamma, \varepsilon), \tag{31a}$$

$$\varepsilon\dot{\beta} = \mathcal{G}(t, \alpha, \beta, \gamma, \varepsilon) - \dot{y}_0, \tag{31b}$$

$$0 = \mathcal{H}_u(t, \alpha, \beta, \gamma, \varepsilon), \tag{31c}$$

$$\alpha^1(0, \varepsilon) = A^*(\varepsilon), \tag{31d}$$

$$\alpha^2(0, \varepsilon) = B^*(\varepsilon), \tag{31e}$$

yields an outer solution of order 0 (i.e., a solution of (2a)–(2c), (4a), (4b)), given by $(x^*(t, \varepsilon), y^*(t, \varepsilon), u^*(t, \varepsilon)) = (x_0(t), y_0(t), u_0(t)) + (\alpha(t, \varepsilon), \beta(t, \varepsilon), \gamma(t, \varepsilon))\varepsilon$.

In the light of the above remark, we will focus our attention on solving system (31). As a preliminary step, we first consider the auxiliary system

$$\dot{\alpha} = \mathcal{F}(t, r, s, v, \varepsilon), \tag{32a}$$

$$\varepsilon\dot{\beta} = g_y(t)\beta + g_u(t)\gamma + \mathcal{G}^*(t, r, s, v, \varepsilon) - \dot{y}_0, \tag{32b}$$

$$0 = H_{uy}(t)\beta + H_{uu}(t) + \mathcal{H}_u^*(t, r, s, v, \varepsilon), \tag{32c}$$

together with boundary conditions

$$\alpha^1(0, \varepsilon) = A^*(\varepsilon), \tag{32d}$$

$$\alpha^2(0, \varepsilon) = B^*(\varepsilon). \tag{32e}$$

The function $\mathcal{G}^*$ is defined by

$$\mathcal{G}^*(t, r, s, v, \varepsilon) = \mathcal{G}(t, r, s, v, \varepsilon) - g_y(t)s - g_u(t)v,$$

and $\mathcal{H}_u^*$ is similarly defined. System (32) may be thought of as a partial linearization of system (31), in which the nonlinear terms are treated as an inhomogeneous forcing function.

From the definitions of $\mathcal{G}^*$ and $\mathcal{H}^*$, it follows that

$$\begin{aligned} &\frac{\partial \mathcal{G}^*}{\partial s}(t, r, s, v, 0) = 0, \qquad \frac{\partial \mathcal{G}^*}{\partial v}(t, r, s, v, 0) = 0, \\ &\frac{\partial \mathcal{H}_u^*}{\partial s}(t, r, s, v, 0) = 0, \qquad \frac{\partial \mathcal{H}_u^*}{\partial v}(t, r, s, v, 0) = 0. \end{aligned} \tag{33}$$

Let $M(t, r, s, v, \varepsilon)$ denote the expression

$$-g_u(t)[H_{uu}(t)]^{-1}\mathcal{H}_u^*(t, r, s, v, \varepsilon) + \mathcal{G}^*(t, r, s, v, \varepsilon) - \dot{y}_0.$$

It is easily verified that M is $K+2$ times continuously differentiable with respect to ε. Moreover, (32c) may be solved explicitly for γ in terms of β and then substituted into (32b) to yield

$$\varepsilon\dot{\beta} = L(t)\beta + M(t, r, s, v, \varepsilon). \tag{34}$$

Consider now the solution of system (32) satisfying the additional boundary conditions

$$\beta^1(0, \varepsilon) = -[B(0)]^{-1}[M(0, r(0), s(0), v(0), 0)]^1, \tag{35a}$$

$$\beta^2(T, \varepsilon) = -[C(T)]^{-1}[M(T, r(T), s(T), v(T), 0)]^2. \tag{35b}$$

Admittedly, reasons for this choice of additional boundary conditions are obscure. The role that they play will only become apparent later in the proof.

Given any $\varepsilon > 0$, the solution of (32), (35) may now be written down explicitly. It is given by

$$\alpha^1(t, \varepsilon) = A^*(\varepsilon) + \int_0^t [\mathcal{F}(\tau, r(\tau), s(\tau), v(\tau), \varepsilon)]^1 \, d\tau, \tag{36a}$$

$$\alpha^2(t, \varepsilon) = B^*(\varepsilon) - \int_t^T [\mathcal{F}(\tau, r(\tau), s(\tau), v(\tau), \varepsilon)]^2 \, d\tau, \tag{36b}$$

$$\beta^1(t, \varepsilon) = \phi(t; 0, \varepsilon)\beta^1(0, \varepsilon) + \frac{1}{\varepsilon}\int_0^t \phi(t; \tau, \varepsilon)[M(\tau, r(\tau), s(\tau), v(\tau), \varepsilon)]^1 \, d\tau, \tag{36c}$$

$$\beta^2(t, \varepsilon) = \psi(t; T, \varepsilon)\beta^2(T, \varepsilon) - \frac{1}{\varepsilon}\int_t^T \psi(t; \tau, \varepsilon)[M(\tau; r, s, v, \varepsilon)]^2 \, d\tau, \tag{36d}$$

$$\gamma(t, \varepsilon) = [H_{uu}(t)]^{-1}[H_{uy}(t)\beta + \mathcal{H}_u^*(t, r, s, v, \varepsilon)]. \tag{36e}$$

In equations (36), $\beta^1(0, \varepsilon)$ and $\beta^2(T, \varepsilon)$ are as given in (35). The functions $\phi(t; \tau, \varepsilon)$ and $\psi(t; \tau, \varepsilon)$ are the fundamental matrix solutions of the systems

$$\varepsilon z' = B(t)z \quad \text{and} \quad \varepsilon z' = C(t)z,$$

respectively, satisfying $\phi(\tau; \tau, \varepsilon) = I = \psi(\tau; \tau, \varepsilon)$.

It is well known that our eigenvalue assumptions on the matrices B and C imply the following:

There exists $C>0$, $\delta>0$ such that

$$\text{(37a)}\qquad \begin{aligned} &\|\phi(t;\tau,\varepsilon)\| \leqq C e^{-\delta(\tau-t)/\varepsilon} \quad \text{for } 0<\varepsilon<\varepsilon_0 \text{ and } 0\leqq\tau\leqq t\leqq T,\\ &\|\psi(t;\tau,\varepsilon)\| \leqq C e^{-\delta(\tau-t)/\varepsilon} \quad \text{for } 0<\varepsilon<\varepsilon_0 \text{ and } 0\leqq t\leqq\tau\leqq T, \end{aligned}$$

$$\text{(37b)}\qquad \begin{aligned} &\int_0^t \phi(t;\tau,\varepsilon)\,d\tau = 0(\varepsilon),\\ &\int_t^T \psi(t;\tau,\varepsilon)\,d\tau = 0(\varepsilon), \end{aligned}$$

uniformly for $0\leqq t\leqq T$.

Now, denote by Φ the mapping of (r, s, v, ε) to $(\alpha, \beta, \gamma, \varepsilon)$ given by (36). Let $\mathcal{B}$ denote the Banach space of continuous functions on $[0, T]$ taking values in $E^{2n_1}\times E^{2n_2}\times E^{n_3}$. Then Φ is defined in some neighborhood θ of $(0, 0, 0)$ in $\mathcal{B}$, and for $0<\varepsilon\leqq\varepsilon_0$, that is,

$$\Phi : \theta\times(0, \varepsilon_0]\to\mathcal{B}.$$

For $\varepsilon=0$, given $(r, s, v)\in\theta$, define $\Phi(r, s, v, 0)$ to be the unique solution $\alpha_0, \beta_0, \gamma_0$ of the system

$$\text{(38a)}\qquad \dot{\alpha}_0 = f_x(t)r + f_y(t)s + f_u(t)v + f_\varepsilon(t),$$

$$\text{(38b)}\qquad 0 = g_x(t)r + g_y(t)\beta_0 + g_u(t)\gamma_0 + g_\varepsilon(t) - \dot{y}_0,$$

$$\text{(38c)}\qquad 0 = H_{ux}(t)r + H_{uy}(t)\beta_0 + H_{uu}(t)\gamma_0 + H_{u\varepsilon}(t),$$

with associated boundary conditions

$$\text{(38d)}\qquad \alpha_0^1(0) = A^*(0) = a_1^*,$$

$$\text{(38e)}\qquad \alpha_0^2(T) = B^*(0) = b_1^*.$$

We note that our assumption of formal solvability suffices to ensure the existence of a solution to (38). In fact, given (r, s, v), α_0 may be determined from (38a) by quadrature, while β_0 and γ_0 can then be solved for in terms of (r, s, v) using the invertibility of $H_{uu}(t)$ and $L(t)$. We have thus extended the domain of definition of Φ, so that

$$\Phi : \theta\times[0, \varepsilon_0]\to\mathcal{B}.$$

Claim 1. The map Φ is continuous.

The continuity of Φ on $\theta\times(0, \varepsilon_0]$ is obvious. All that we must verify is that Φ is continuous at $\varepsilon=0$.

Note first that the continuity of $\mathcal{F}$ at $\varepsilon=0$ implies that $\alpha(t, \varepsilon)\to\alpha_0(t)$ uniformly on $0\leqq t\leqq T$ as $\varepsilon\to 0$. Moreover, $\gamma(t, \varepsilon)$, given by (36e), will reduce, as $\varepsilon\to 0+$, to (37c), provided we can first establish the fact that $\beta(t, \varepsilon)\to\beta_0(t)$ uniformly on $0\leqq t\leqq T$ as $\beta\to 0+$.

Let us therefore examine the behavior of $\beta(t, \varepsilon)$. Recall that we have previously shown that (32b) and (32c) may be solved for γ in terms of β, to yield (34). It follows that for $\varepsilon=0$,

$$\text{(39)}\qquad L(t)\beta(t, 0) = -M(t, r, s, v, 0).$$

By direct computation, we can verify that (38b) and (38c) may be similarly solved for β_0 to yield

$$L(t)\beta_0(t) = -M(t, r, s, v, 0).$$

Thus, since $L(t)$ is invertible, $\beta(t, 0) = \beta_0(t)$. If we now use the assumed representation $L(t) = \operatorname{diag}[B(t), C(t)]$, then (39) can be rewritten as

$$\beta_0^1(t) = \beta^1(t, 0) = -[B(t)]^{-1}[M(t, r, s, v, 0)]^1,$$
$$\beta_0^2(t) = \beta^2(t, 0) = -[C(t)]^{-1}[M(t, r, s, v, 0)]^2.$$

If we now employ (35) and (36), we have that for $\varepsilon > 0$,

$$\beta^1(t, \varepsilon) = \phi(t; 0, \varepsilon)\beta_0^1(0) + \frac{1}{\varepsilon}\int_0^t \phi(t; \tau, \varepsilon)[M(\tau, r, s, v, \varepsilon)]^1\, d\tau,$$

$$\beta^2(t, \varepsilon) = \psi(t; T, \varepsilon)\beta_0^2(T) - \frac{1}{\varepsilon}\int_t^T \psi(t; \tau, \varepsilon)[M(\tau, r, s, v, \varepsilon)]^2\, d\tau.$$

Let us now show that $\beta^1(t, \varepsilon) \to \beta_0^1(t)$ uniformly on $0 \leqq t \leqq T$ as $\varepsilon \to 0+$. The proof for $\beta^2(t, \varepsilon)$ is similar.

Define $W(t, \varepsilon) = \beta^1(t, \varepsilon) - \beta_0^1(t)$. It is then easy to show that $W(t, \varepsilon)$ satisfies the equation

$$W(t, \varepsilon) = \int_0^t \phi(t; \tau, \varepsilon)\left[\frac{M(\tau, r, s, v, \varepsilon) - M(\tau, r, s, v, 0)}{\varepsilon} - \beta_0^1(\tau)\right] d\tau.$$

(Note that it is in the derivation of this equation that the additional boundary condition (35a) is crucial.)

Now the bracketed term on the right in the above equation is $O(1)$, while $\int_0^t \phi(t; \tau, \varepsilon) = o(\varepsilon)$, $\varepsilon \to 0$. It follows that $\beta^1(t, \varepsilon) \to \beta_0^1(t)$ uniformly on $[0, T]$, thus establishing our claim of the continuity of Φ.

We remark that it is easily verified that $\Phi : \theta \times [0, \varepsilon_0] \to \mathcal{B}$ has a continuous Fréchet derivative for each $(r, s, v) \in 0$ and $0 \leqq \varepsilon < \varepsilon_0$. The continuity of these derivatives at $\varepsilon = 0$ may be demonstrated by a proof analogous to the one just given for the continuity of Φ at $\varepsilon = 0$.

In particular, let us write down the Fréchet derivative at $\varepsilon = 0$. At any point $(r, s, v) \in \theta$, the Fréchet derivative $D_{(r,s,v)}\Phi|_{\varepsilon=0}$ must be a bounded linear transformation in $\mathcal{B}$:

$$D_{(r,s,v)}\Phi|_{\varepsilon=0} : (\eta_1, \eta_2, \eta_3) \to (\mu_1, \mu_2, \mu_3).$$

It may be shown that $\mu_1(t)$, $\mu_2(t)$, $\mu_3(t)$ must be differentiable, and they will satisfy the equations

$$\dot{\eta}_1(t) = f_x(t)\eta_1 + f_y(t)\eta_2 + f_u(t)\eta_3, \tag{40a}$$

$$0 = g_x(t)\eta_1 + g_y(t)\mu_2 + g_u(t)\mu_3, \tag{40b}$$

$$0 = H_{ux}(t)\eta_1 + H_{uy}(t)\mu_2 + H_{uu}(t)\mu_3, \tag{40c}$$

together with boundary conditions

$$\mu_1^1(0) = \mu_1^2(T) = 0. \tag{40d}$$

It may be useful to observe that (33) plays a crucial role in the derivation of system (40). Not surprisingly, equations (40) are simply the "variational equations" corresponding to (38).

To complete our proof, under the assumption $L(t) = \operatorname{diag}[B(t), C(t)]$, we must show that there exist $\alpha(t, \varepsilon)$, $\beta(t, \varepsilon)$, $\gamma(t, \varepsilon)$, jointly continuous in t and ε for $0 \leqq t \leqq T$, $0 \leqq \varepsilon < \varepsilon_0$, satisfying

$$\Phi(\alpha, \beta, \gamma, \varepsilon) = (\alpha, \beta, \gamma).$$

Our tool for the demonstration of this fact will be the Banach space form of the implicit function theorem.

Define $\Psi : \theta \times [0, \varepsilon_0] \to \mathscr{B}$ by

$$\Psi(r, s, v, \varepsilon) = \Phi(r, s, v, \varepsilon) - (I(r, s, v), \varepsilon),$$

where $I : \mathscr{B} \to \mathscr{B}$ is the identity map. Then Ψ is a continuous mapping which has a continuous Fréchet derivative for each $(r, s, v, \varepsilon) \in \theta \times [0, \varepsilon_0]$. Moreover, a comparison of (12) with (36) shows that for $(r(t), s(t), v(t)) = (x_1(t), y_1(t), u_1(t))$ (where $x_1(t)$, $y_1(t)$, $u_1(t)$ denote the appropriate coefficients which were formally derived in § 3), we must have

$$\Psi(x_1, y_1, u_1, 0) = 0.$$

The Banach space implicit function theorem will now establish the existence of the desired $(\alpha(t, \varepsilon), \beta(t, \varepsilon), \gamma(t, \varepsilon))$ satisfying

$$0 = \psi(\alpha, \beta, \gamma, \varepsilon)$$

or, equivalently,

$$(\alpha, \beta, \gamma) = \Phi(\alpha, \beta, \gamma, \varepsilon),$$

provided we can demonstrate that the Fréchet derivative of Ψ at $(x_1, y_1, u_1, 0)$ is a topological linear isomorphism. But

$$D_{(r,s,v)}\Psi|_{(x_1,y_1,u_1,0)} = D_{(r,s,v)}\Phi|_{(x_1,y_1,u_1,0)} - (I(x_1, y_1, u_1), 0).$$

Thus, by (40), for $(\eta_1, \eta_2, \eta_3) \in \mathscr{B}$, the mapping

$$D_{(r,s,v)}\Psi|_{(x_1,y_1,u_1,0)}(\eta_1, \eta_2, \eta_3) \to (\mu_1, \mu_2, \mu_3)$$

is given explicitly by

$$(\eta_1 + \mu_1)^1(t) = \int_0^t [f_x(\tau)\eta_1(\tau) + f_y(\tau)\eta_2(\tau) + f_u(\tau)\eta_3(\tau)]^1 \, d\tau,$$

(41a)

$$(\eta_1 + \mu_1)^2(t) = -\int_t^T [f_x(\tau)\eta_1(\tau) + f_y(\tau)\eta_2(\tau) + f_u(\tau)\eta_3(\tau)]^2 \, d\tau,$$

(41b) $$0 = g_x(t)\eta_1 + g_y(t)(\eta_2 + \mu_2) + g_u(t)(\eta_3 + \mu_3),$$

(41c) $$0 = H_{ux}(t)\eta_1 + H_{uy}(t)(\eta_2 + \mu_2) + H_{uu}(t)(\eta_3 + \mu_3).$$

In differential form, (41a) may be rewritten as

(42a) $$(\dot{\eta}_1 + \dot{\mu}_1) = f_x(t)\eta_1 + f_y(t)\eta_2 + f_u(t)\eta_3,$$

with boundary conditions

$$(\eta_1+\mu_1)^1(0)=0, \tag{42b}$$

$$(\eta_1+\mu_1)^2(T)=0. \tag{42c}$$

Now the system (41b), (41c), (42) represents a continuous linear mapping from (η_1, η_2, η_3) to (μ_1, μ_2, μ_3), since $\det L(t)\neq 0$.

To check that this map has a bounded linear inverse, it is convenient for us to set

$$\delta_i=\eta_i+\mu_i, \qquad i=1,2,3.$$

With this change of variables, (41b), (41c), (42) become

$$\dot{\delta}_1=f_x(t)\delta_1+f_y(t)\delta_2+f_u(t)\delta_3-[f_x(t)\mu_1+f_y(t)\mu_2+f_u(t)\mu_3], \tag{43a}$$

$$0=g_x(t)\delta_1+g_y(t)\delta_2+g_u(t)\delta_3-g_x(t)\mu_1, \tag{43b}$$

$$0=H_{ux}(t)\delta_1+H_{uy}(t)\delta_2+H_{uu}(t)\delta_3-H_{ux}(t)\mu_1, \tag{43c}$$

$$\delta_1^1(0)=0, \tag{43d}$$

$$\delta_1^2(T)=0. \tag{43e}$$

But our assumption of the formal solvability of the outer system is now precisely what is required to ensure the existence of $(\delta_1, \delta_2, \delta_3)\in\mathcal{B}$ satisfying (43) for any choice of $(\mu_1, \mu_2, \mu_3)\in\mathcal{B}$. Moreover, the linear map so defined from $(\mu_1, \mu_2, \mu_3)\to(\delta_1, \delta_2, \delta_3)$ is clearly continuous. Finally, since $\eta_i=\delta_i-\mu_i$, $i=1,2,3$, the linear map from $(\mu_1, \mu_2, \mu_3)\to(\eta_1, \eta_2, \eta_3)$ must be continuous. Of course, this is Φ^{-1}.

We have therefore shown that $D_{(r,s,v)}\Psi|_{(x_1,y_1,u_1,0)}$ is a topological linear isomorphism. We conclude that there exist $\alpha(t,\varepsilon)$, $\beta(t,\varepsilon)$, $\gamma(t,\varepsilon)$ continuous in $0\leqq t\leqq T$ and $0\leqq\varepsilon<\varepsilon_0$, satisfying system (31). As previously remarked,

$$(x^*(t,\varepsilon), y^*(t,\varepsilon), u^*(t,\varepsilon))=(x_0(t), y_0(t), u_0(t))+(\alpha(t,\varepsilon), \beta(t,\varepsilon), \gamma(t,\varepsilon))\varepsilon$$

constitutes the outer solution whose existence is asserted in the statement of Theorem 4.1.

We now remark that the general case in which $L(t)$ is not block diagonalized can be reduced to the previous case. Let $P(t)$ be the matrix given in Remark 3.1. The change of variables $y=P(t)z$ will now transform the outer system into the system

$$\begin{aligned}\dot{x}&=f(t, x, Pz, u, \varepsilon),\\ \varepsilon\dot{z}&=[P(t)]^{-1}g(t, x, Pz, u, \varepsilon)-[P(t)]^{-1}\dot{P}(t)z,\\ 0&=H_u(t, x, P(t)z, u, \varepsilon).\end{aligned}$$

This system is of the same general form as the outer system, but the appropriate "$L(t)$" for this system is block diagonalized, thus reducing the general case to the one already treated. We omit the details.

The proof for the case $K>0$ proceeds inductively.

Suppose that we have already established the existence of an outer solution of order $k-1$, $1\leqq k\leqq K$, which satisfies the other requirements of the theorem. We

define

$$\alpha(t,\varepsilon)=\frac{x^*(t,\varepsilon)-\sum_{j=0}^{k-1}x_j(t)\varepsilon^j}{\varepsilon^k},$$

$$\beta(t,\varepsilon)=\frac{y^*(t,\varepsilon)-\sum_{j=0}^{k-1}y_j(t)\varepsilon^j}{\varepsilon^k},$$

$$\gamma(t,\varepsilon)=\frac{u^*(t,\varepsilon)-\sum_{j=0}^{k-1}u_j(t)\varepsilon^j}{\varepsilon^k}.$$

Then α, β, γ will satisfy a system of the same form as (31), with appropriately modified functions $\mathscr{F}$, $\mathscr{G}$, $\mathscr{H}_u$, A^*, B^*, and $\dot{y}_0$ replaced with $\dot{y}_{k-1}$.

By using the analysis developed previously, we may establish the existence of a continuous, bounded solution $\alpha(t,\varepsilon)$, $\beta(t,\varepsilon)$, $\gamma(t,\varepsilon)$ of the resulting system for $0\leqq t\leqq T$, $0\leqq\varepsilon<\varepsilon_0$. We omit the details.

5. The existence of left and right boundary layer solutions. In this section, we shall return again to the consideration of the left boundary layer equations (5) and the right boundary layer equations (7). For suitable choices of initial conditions, (5d) and (7d), respectively, we will establish the existence of solutions which satisfy our requirement of exponential decay. We will begin with the left boundary layer solution.

THEOREM 5.1. *Let* (9), (10) *and* (20) *hold, and suppose that the outer system is formally solvable. Let the matrix* $P(t)$ *of Remark* 3.1 *satisfy* $\det P_{11}(0)$. *Then for each* $\hat{c}(\varepsilon)\in E^{n_2}$ *sufficiently small, which is* $K+1$ *times continuously differentiable, there exists a solution* $X^L(\tau,\varepsilon)$, $Y^L(\tau,\varepsilon)$, $U^L(\tau,\varepsilon)$ *of* (5) *on* $0\leqq\tau\leqq T/\varepsilon$, *satisfying* $\rho^L(0,\varepsilon)=c(\varepsilon)$. *Moreover,* $X^L(\tau,\varepsilon)$, $Y^L(\tau,\varepsilon)$, $U^L(\tau,\varepsilon)$ *are* $K+1$ *times continuously differentiable with respect to* ε, *and satisfy decay condition* (6) *for some* $C>0$, $\delta>0$ *and for all* $0\leqq\tau\leqq T/\varepsilon$.

In addition, if $X_0^L(\tau)$, $Y_0^L(\tau)$, $U_0^L(\tau)$ *represents the solution of* (17), *while for* $1\leqq k\leqq K$, $X_k^L(\tau)$, $Y_k^L(\tau)$, $U_k^L(\tau)$ *represents the solution of* (18), *then*

$$\begin{aligned}
X^L(\tau,\varepsilon)-\sum_{k=0}^{K}X_k^L(\tau)\varepsilon^k&=O(\varepsilon^{K+1}),\\
Y^L(\tau,\varepsilon)-\sum_{k=0}^{K}Y_k^L(\tau)\varepsilon^k&=O(\varepsilon^{K+1}),\\
U^L(\tau,\varepsilon)-\sum_{k=0}^{K}U_k^L(\tau)\varepsilon^k&=O(\varepsilon^{K+1}),
\end{aligned}\tag{44}$$

uniformly on $0\leqq\tau\leqq T/\varepsilon$, *as* $\varepsilon\to 0+$.

Our proof of Theorem 5.1 will follow by an appropriate interpretation of Hoppensteadt's Lemma 2 [8]. For easy reference, we restate that result as our Lemma 5.2, using Hoppensteadt's notation.

Consider the systems

$$\frac{dX}{d\tau} = \varepsilon f(\varepsilon\tau, X, Y, \varepsilon), \quad X \in E^m, \quad Y \in E^n, \tag{45a}$$

$$\frac{dY}{d\tau} = g(\varepsilon\tau, X, Y, \varepsilon), \tag{45b}$$

$$X(0) = \hat{\xi}(\varepsilon), \tag{45c}$$

$$Y(0) = \hat{\eta}(\varepsilon), \tag{45d}$$

$$\frac{dX_0}{d\tau} = 0, \tag{46a}$$

$$\frac{dY_0}{d\tau} = \hat{g}(0, X_0, Y_0, 0), \tag{46b}$$

$$X_0(0) = \hat{\xi}(0), \tag{46c}$$

$$Y_0(0) = \hat{\eta}(0), \tag{46d}$$

and, for $1 \leqq k \leqq K$,

$$\frac{dX_k}{d\tau} = p_k(\tau), \tag{47a}$$

$$\frac{dY_k}{d\tau} = \hat{g}_x(\tau) X_k + g_y(\tau) Y_k + q_k(\tau), \tag{47b}$$

$$X_k(0) = \hat{\xi}_k, \tag{47c}$$

$$Y_k(0) = \hat{\eta}_k, \tag{47d}$$

where p_k, q_k are polynomials in $X_1, Y_1, \cdots, X_{k-1}, Y_{k-1}$ with coefficients depending on τ, $X_0(\tau)$, $Y_0(\tau)$. In the above, f, g have arisen in a manner similar to that of our paper from functions f and g, respectively.

LEMMA 5.2 (Hoppensteadt [8]). *Let hypotheses analogous to* (9) *and* (10) *hold for system* (45). *Suppose, moreover, that the analogue of* (20) *holds, in which we require that* $g_y(t)$ *have* k *eigenvalues with negative real part* $\operatorname{Re}\lambda(t) \leqq -\mu < 0$, *and* $n-k$ *with* $\operatorname{Re}\lambda(t) \geqq \mu > 0$. *Then, for each small* $\varepsilon > 0$, *there exists a* k*-dimensional manifold* $S(\varepsilon) \in E^{m+n}$ *such that* (45) *has a unique solution* $X(t, \varepsilon)$, $Y(t, \varepsilon)$ *on* $0 \leqq t \leqq T$, *provided* $(\hat{\xi}(\varepsilon), \hat{\eta}(\varepsilon)) \in S(\varepsilon)$. *Moreover, if* $(\hat{\xi}, \hat{\eta}) \in S(\varepsilon)$, *the problems* (46) *and* (47) *have unique solutions existing on* $0 \leqq \tau < \infty$, *and*

$$(X(t, \varepsilon), Y(t, \varepsilon)) - \sum_{k=0}^{K} (X_k(\tau), Y_k(\tau))\varepsilon^k = O(\varepsilon^{K+1}),$$

where $O(\varepsilon^{K+1})$ *holds uniformly for* $0 \leqq t \leqq T$ *as* $\varepsilon \to 0+$. *In addition, there are positive constants* K_1, δ_1, ε_0'' *such that*

$$|X(t, \varepsilon)| + |Y(t, \varepsilon)| \leqq K_1 e - \delta_1 t/\varepsilon$$

for $0 \leqq t \leqq T$, $0 < \varepsilon \leqq \varepsilon_0''$.

Thus the differences between Hoppensteadt's Lemma 2 (Lemma 5.2 above) and our own Theorem 5.1 are as follows. First, we must also contend with the additional function $U(\tau, \varepsilon)$ in (5). Second, we impose the additional assumption

that $\det P_{11}(0) \neq 0$. As we shall see, this enables us to show that the projection of $E^{2n_1} \times E^{2n_2} \to E^{n_2}$ given by $(\xi, \eta) \to \eta^1$ maps $S(\varepsilon)$ 1–1 onto some neighborhood of the origin in E^{n_2}. Here we have again used η^1 to denote the first n_2 components of $\eta = \begin{pmatrix} \eta^1 \\ \eta^2 \end{pmatrix}$.

Proof of Theorem 5.1. As was pointed out, our proof will rely heavily upon Hoppensteadt's proof of Lemma 5.2.

First, let us attempt to solve (5c) for U in terms of X and Y. By a simple continuity argument, $\hat{H}_{uu}(\varepsilon\tau, X, Y, U, \varepsilon)$ will be nonsingular for X, Y, U and ε in a sufficiently small neighborhood of 0 and $0 \leqq \tau \leqq T/\varepsilon$. This follows, since

$$\hat{H}_{uu}(\varepsilon\tau, X, Y, U, \varepsilon) = H_{uu}(t, x^*(t, \varepsilon) + X(\tau, \varepsilon), y^*(t, \varepsilon) + Y(\tau, \varepsilon), u^*(t, \varepsilon) + U(\tau, \varepsilon), \varepsilon),$$

and by hypothesis (10c), $H_{uu}(t, x_0(t), y_0(t), u_0(t), 0)$ is nonsingular for $0 \leqq t \leqq T$.

Thus, by the implicit function theorem in Euclidean space, there exists a continuously differentiable map ϕ with $U = \phi(\varepsilon\tau; X, Y, \varepsilon)$ satisfying

$$0 = \phi(\varepsilon\tau, 0, 0, \varepsilon),$$

which is defined for X, Y in a sufficiently small neighborhood of 0 and $0 < \varepsilon \leqq \varepsilon_1 < \varepsilon_0$, $0 \leqq \tau \leqq T/\varepsilon$. Moreover,

$$0 = \hat{H}_u(\varepsilon, X, Y, \phi(\varepsilon\tau, X, Y, \varepsilon), \varepsilon)$$

and

$$\begin{aligned} \Phi_x(0) &= -[H_{uu}(0)]^{-1} H_{ux}(0), \\ \Phi_y(0) &= -[H_{uu}(0)]^{-1} H_{uy}(0). \end{aligned} \tag{48}$$

If we make use of these relations in system (5), we obtain the new system

$$\frac{dX}{d\tau} = \varepsilon \hat{f}(\varepsilon\tau, X, Y, \Phi(\varepsilon\tau, X, Y, \varepsilon), \varepsilon), \tag{49a}$$

$$\frac{dY}{d\tau} = \hat{g}(\varepsilon\tau, X, Y, \Phi(\varepsilon\tau, X, Y, \varepsilon), \varepsilon), \tag{49b}$$

which is of the general form treated by Lemma 5.2. We may verify, moreover, that the hypotheses of Lemma 5.2 will be satisfied, provided

$$L(0) = g_y(0) - g_u(0)[H_{uu}(0)]^{-1} H_{uy}(0)$$

has n_2 eigenvalues with negative real part, and n_2 with positive real part, and the system

$$\begin{aligned} \dot{\tilde{x}} &= M(0)\tilde{x}, \\ \tilde{\xi}(0) &= 0, \\ \tilde{\chi}(T) &= 0, \end{aligned}$$

has only the trivial solution. Here we have again used the notation $\tilde{x} = \begin{pmatrix} \tilde{\xi} \\ \tilde{\chi} \end{pmatrix}$, and $M(t)$ was defined in (15). Thus Hoppensteadt's Lemma 2 is applicable here.

A careful examination of the details of the proof of Hoppensteadt's Lemma 2 to our system shows that there exist smooth mappings

$$W_1(\tau, \varepsilon, \eta(\varepsilon)) \in E^{2n_1}, \quad W_2(\tau, \varepsilon, \eta(\varepsilon)) \in E^{n_2}, \quad W_3(\tau, \varepsilon, \eta(\varepsilon)) \in E^{n_3},$$

defined for $0 \leqq \tau \leqq T/\varepsilon$, ε sufficiently small, and $\eta(\varepsilon)$ in some neighborhood of 0 in E^{2n_2}, with the following properties:

$$W_2(0, \varepsilon, \eta(\varepsilon)) = \eta(\varepsilon),$$

and the solution of system (49) is given explicitly by

$$X(\tau, \varepsilon) = W_1(\tau, \varepsilon, \eta(\varepsilon)),$$

$$Y(\tau, \varepsilon) = P(\varepsilon\tau)\begin{pmatrix} W_2(\tau, \varepsilon, \eta(\varepsilon)) \\ W_3(\tau, \varepsilon, \eta(\varepsilon)) \end{pmatrix} - (L^{-1}(\varepsilon\tau))[g_x(\varepsilon\tau) - g_u(\varepsilon\tau)][H_{uu}(\varepsilon\tau)]^{-1} H_{ux}(\varepsilon\tau) W_1(\tau, \varepsilon, \eta(\varepsilon)).$$

It follows that

$$X(0, \varepsilon) = W_1(0, \varepsilon, \eta(\varepsilon)), \tag{50a}$$

$$Y(0, \varepsilon) = P(0)\begin{pmatrix} \eta(\varepsilon) \\ W_3(0, \varepsilon, \eta(\varepsilon)) \end{pmatrix} - [L(0)]^{-1}[g_x(0) - g_u(0)][H_{uu}(0))]^{-1} H_{ux}(0) W_1(0, \varepsilon, \eta(\varepsilon)). \tag{50b}$$

Our proof will be completed if we can show that the mapping $\Psi : (\eta(\varepsilon), \varepsilon) \to (Y(0, \varepsilon))^1$ is a local homeomorphism for ε sufficiently small and $\eta(\varepsilon)$ in some neighborhood of the origin in E^{2n_2}.

By the inverse function theorem, it will suffice to show that the linear mapping

$$D_\eta \Psi|_{(0,0)} = (\eta(\varepsilon), \varepsilon)$$

is a linear isomorphism.

Now, when $\varepsilon = 0$, our consideration of the mapping Ψ reduces to an examination of the solution $X_0^L(\tau)$, $Y_0^L(\tau)$, satisfying (17). (The function $U_0^L(\tau)$ is extraneous in this computation.)

We have already seen in Theorem 3.2(i) that $X_0^L(\tau) \equiv 0$. It follows that $X_0^L(0) = 0$, so that

$$D_\eta W_1(0, \varepsilon, \eta(0))|_{(0,0)} = 0.$$

Also, it follows from Lemma 3.2 (see [15] for details) that

$$D_\eta W_3(0, \varepsilon, \eta(0))|_{(0,0)} = 0.$$

Combining the above Fréchet derivatives with (50b), we obtain

$$D_\eta (Y(0, \varepsilon))^1 = P_{11}(0). \tag{51}$$

Our assumption that $\det P_{11}(0) \neq 0$ now guarantees that for each sufficiently small $\hat{c}(\varepsilon) \in E^{2n_2}$, equations (49) have a solution $X(\tau, \varepsilon)$, $Y(\tau, \varepsilon)$ on $0 \leqq \tau \leqq T/\varepsilon$, for ε

sufficiently small, which satisfies

$$Y^1(0, \varepsilon) = \hat{c}(\varepsilon),$$

as well as

$$|X(\tau, \varepsilon)| + |Y(\tau, \varepsilon)| \cdot \leqq C e^{-\delta\tau}$$

for some C, $\delta > 0$. Recalling that $U = \phi(\varepsilon\tau, X, Y, \varepsilon)$ now establishes that $|U(\tau, \varepsilon)| \leqq C e^{-\delta\tau}$, and $X(\tau, \varepsilon)$, $Y(\tau, \varepsilon)$, $U(\tau, \varepsilon)$ satisfy (5), as well as (44).

We note that, as a consequence of Theorem 5.1, we may consider $X(0, \varepsilon)$ as a smooth function of the initial condition $\hat{c}(\varepsilon)$. This may be indicated explicitly by $X(0, \varepsilon, \hat{c}(\varepsilon))$. We further note that when $\varepsilon = 0$, $X(0, \varepsilon, \hat{c}(\varepsilon)) = X_0^L(0) = 0$, so that $D_{\hat{c}(\varepsilon)} X(0, \varepsilon, \hat{c}(\varepsilon))|_{\varepsilon=0} = 0$.

We now turn our attention to the right boundary layer solution. By direct analogy with Theorem 5.1, we have the following theorem.

THEOREM 5.3. *Let* (9), (10) *and* (20) *hold, and suppose that the outer system is formally solvable. Let the matrix* $P(t)$ *satisfy* $\det P_{22}(T) \neq 0$. *Then for each* $\hat{d}(\varepsilon) \in E^{n_2}$ *sufficiently small, which is* $K+1$ *times continuously differentiable, there exists a solution* $X^R(\sigma, \varepsilon)$, $Y^R(\sigma, \varepsilon)$, $U^R(\sigma, \varepsilon)$ *of* (7) *on* $0 \leqq \sigma \leqq T/\varepsilon$, *satisfying* $(Y^2)^R(0, \varepsilon) = \hat{d}(\varepsilon)$. *Moreover,* $X^R(\sigma, \varepsilon)$, $Y^R(\sigma, \varepsilon)$, $U^R(\sigma, \varepsilon)$ *are* $K+1$ *times continuously differentiable with respect to* ε *and satisfy* (8) *for some* C, $\delta > 0$, $0 \leqq \sigma \leqq T/\varepsilon$.

In addition, if $X_0^R(\sigma)$, $Y_0^R(\sigma)$, $U_0^R(\sigma)$ *represents the solution of* (28), *while for* $1 \leqq k \leqq K$, $X_k^R(\sigma)$, $Y_k^R(\sigma)$, $U_k^R(\sigma)$ *represents the solution of* (29), *then*

$$\begin{aligned} X^R(\sigma, \varepsilon) - \sum_{k=0}^{K} X_k^R(\sigma)\varepsilon^k &= O(\varepsilon^{K+1}), \\ Y^R(\sigma, \varepsilon) - \sum_{k=0}^{K} Y_k^R(\sigma)\varepsilon^k &= O(\varepsilon^{K+1}), \\ U^R(\sigma, \varepsilon) - \sum_{k=0}^{K} U_k^R(\sigma)\varepsilon^k &= O(\varepsilon^{K+1}), \end{aligned} \tag{52}$$

uniformly on $0 \leqq \sigma \leqq T/\varepsilon$, *as* $\varepsilon \to 0+$.

6. The main result. Our only remaining task is to show that we can synthesize a solution of the full system as the sum of an appropriate outer solution, a left boundary layer solution and a right boundary layer solution.

THEOREM 6.1. *Let* (9), (10) *and* (20) *hold, and suppose that the outer system is formally solvable. Let the matrix* $P(t)$ *of Remark* 3.1 *satisfy* $\det P_{11}(0) \neq 0$, $\det P_{22}(T) \neq 0$. *Then there exists* $\eta > 0$ *such that if* $|c_0 - \rho_0(0)\| < \eta$ *and* $|d_0 - \nu_0(T)| < \eta$, *the full system* (1) *has a unique solution* $x(t, \varepsilon)$, $y(t, \varepsilon)$, $u(t, \varepsilon)$ *for* ε *sufficiently small. In fact, there exists a choice of* $a^*(\varepsilon)$, $b^*(\varepsilon)$, $\hat{c}(\varepsilon)$, $\hat{d}(\varepsilon)$ *such that there exists an outer solution* $x^*(t, \varepsilon)$, $y^*(t, \varepsilon)$, $u^*(t, \varepsilon)$, *a left boundary layer solution of* (5), (6), $X^L(\tau, \varepsilon)$, $Y^L(\tau, \varepsilon)$, $U^L(\tau, \varepsilon)$, *and a right boundary layer solution* $X^R(\sigma, \varepsilon)$,

$Y^R(\sigma, \varepsilon)$, $U^R(\sigma, \varepsilon)$ *of* (7), (8), *with*

$$(53)\qquad \begin{aligned}(x(t,\varepsilon), y(t,\varepsilon), u(t,\varepsilon)) &= (x^*(t,\varepsilon), y^*(t,\varepsilon), u^*(t,\varepsilon))\\ &+(X^L(t/\varepsilon,\varepsilon), Y^L(t/\varepsilon,\varepsilon), U^L(t/\varepsilon,\varepsilon))\\ &+\left(X^R\left(\frac{T-t}{\varepsilon},\varepsilon\right), Y^R\left(\frac{T-t}{\varepsilon},\varepsilon\right), U^R\left(\frac{T-t}{\varepsilon},\varepsilon\right)\right).\end{aligned}$$

We remark that the outer solution, the left boundary layer solution and the right boundary layer solution satisfy (30), (44) and (52), respectively.

Proof. Consider the mapping π defined on a subset of $E^{n_1}\times E^{n_1}\times E^{n_2}\times E^{n_2}$ into $E^{n_1}\times E^{n_1}\times E^{n_2}\times E^{n_2}$ given as follows:

$$(a^*(\varepsilon), b^*(\varepsilon), \hat{c}(\varepsilon), \hat{d}(\varepsilon)) = (\xi(0,\varepsilon), \chi(T,\varepsilon), \rho(0,\varepsilon), \nu(T,\varepsilon)),$$

where

$$x(t,\varepsilon) = \begin{pmatrix}\xi(t,\varepsilon)\\ \chi(t,\varepsilon)\end{pmatrix} \quad \text{and} \quad y(t,\varepsilon) = \begin{pmatrix}\rho(t,\varepsilon)\\ \nu(t,\varepsilon)\end{pmatrix}$$

are given in (53). The domain of definition of the mapping π is chosen so that for an $(a^*(\varepsilon), b^*(\varepsilon), \hat{c}(\varepsilon), \hat{d}(\varepsilon))$ in its domain, our previous existence theorems ensure the existence of the corresponding outer solution and the left and right boundary layer solutions. Now, using the asymptotic decay of the boundary layer solutions, we see that for $\varepsilon = 0$ we have

$$\begin{aligned}\pi(a^*(0), b^*(0), \hat{c}(0), \hat{d}(0)) &= \pi(a_0, b_0, c(0), d(0))\\ &= (a_0, b_0, \rho_0(0)+\hat{c}(0), \nu_0(T)+\hat{d}(0)).\end{aligned}$$

If we now choose $\hat{c}(0) = c_0 - \rho_0(0)$, $\hat{d}(0) = d_0 - \nu_0(T)$, then, provided these are sufficiently small, our existence theorems, Theorems 5.1 and 5.3, will guarantee boundary layer solutions for which

$$(a^*(0), b^*(0), \hat{c}(0), \hat{d}(0)) = (a_0, b_0, c_0, d_0),$$

that is, the desired boundary conditions are satisfied at $\varepsilon = 0$.

We may now compute the Jacobian matrix of the mapping π. (The terms in which we are interested have, in fact, been computed previously at various points in the paper.) This matrix, when written in block form, will be a 4×4 matrix, with the blocks along the principal diagonal having dimensions $n_1\times n_1$, $n_1\times n_1$, $n_2\times n_2$ and $n_2\times n_2$, respectively. Let us examine one of the typical terms of this matrix. For example, the block in the first row, second column is $\partial\xi(0,\varepsilon)/\partial b^*(\varepsilon)$. But $\xi(t,\varepsilon)$ depends only upon the choice of $a^*(\varepsilon)$, so that $(\partial\xi/\partial b^*(\varepsilon))(0,\varepsilon) = 0$. In fact, this same reasoning applies to all the superdiagonal elements. Thus the Jacobian matrix is lower triangular. It is given by

$$\begin{pmatrix} I_{n_1\times n_1} & & 0 & \\ \cdot & I_{n_1\times n_1} & & \\ \cdot & \cdot & P_{11}(0) & \\ \cdot & \cdot & \cdot & P_{22}(T)\end{pmatrix},$$

where we have omitted the specific form of the subdiagonal elements. It is readily seen that our hypotheses imply that this matrix is nonsingular. It follows from what are by now standard implicit function theorem arguments that for $\varepsilon > 0$, ε sufficiently small, we may solve for $(a^*(\varepsilon), b^*(\varepsilon), \hat{c}(\varepsilon), \hat{d}(\varepsilon))$ as a function of $a(\varepsilon)$, $b(\varepsilon)$, $c(\varepsilon)$, $d(\varepsilon)$.

REFERENCES

[1] K. W. CHANG, *Remarks on a certain hypothesis in singular perturbations*, Proc. Amer. Math. Soc., 23 (1969), pp. 41–45.

[2] K. W. CHANG AND W. A. COPPEL, *Singular perturbations of initial value problems over a finite interval*, Arch. Rational Mech. Anal., 32 (1969) pp. 268–280.

[3] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[4] L. FLATTO AND N. LEVINSON, *Periodic solutions of singularly perturbed systems*, J. Rational Mech. Anal., 4 (1955), pp. 943–950.

[5] M. I. FREEDMAN AND B. GRANOFF, *The formal asymptotic solution of a singularly perturbed nonlinear optimal control problem*, (1974) to appear.

[6] C. R. HADLOCK, *Existence and dependence on a parameter of solutions of a nonlinear two point boundary value problem*, J. Differential Equations, 14 (1973), no. 3, pp. 498–517.

[7] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[8] F. HOPPENSTEADT, *Properties of solutions of ordinary differential equations with small parameters*, Comm. Pure Appl. Math., 24 (1971), pp. 807–840.

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

[10] J. LEVIN, *Singular perturbations of nonlinear systems of differential equations related to conditional stability*, Duke Math. J., 24 (1956), no. 4, pp. 609–620.

[11] R. O'MALLEY, JR., *The singularly perturbed linear state regulator problem*, this Journal, 10 (1972), pp. 399–413.

[12] ———, *Boundary layer methods for certain nonlinear singularly perturbed optimal control problems*, J. Math. Anal. Appl., 45 (1974), pp. 468–484.

[13] P. SANNUTI, *Asymptotic series solution of singularly perturbed optimal control problems*, Automatica, 10 (1974) pp. 183–194.

[14] R. R. WILDE AND P. V. KOKOTOVIC, *Optimal open- and closed-loop control of singularly perturbed linear systems*, IEEE Trans. Automatic Control, AC-18 (1973) pp. 616–625.

[15] M. I. FREEDMAN AND J. L. KAPLAN, *Singular perturbations of two point boundary value problems arising in optimal control*, Boston Univ. Research Rep. 1974–4, Boston, Mass.

[16] P. C. FIFE, *Transition layers in singular perturbation problems*, J. Differential Equations, 15 (1974), pp. 77–105.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

ON PENALTY AND MULTIPLIER METHODS FOR CONSTRAINED MINIMIZATION*

DIMITRI P. BERTSEKAS†

Abstract. In this paper we consider a generalized class of quadratic penalty function methods for the solution of nonconvex nonlinear programming problems. This class contains as special cases both the usual quadratic penalty function method and the recently proposed multiplier method. We obtain convergence and rate of convergence results for the sequences of primal and dual variables generated. The convergence results for the multiplier method are global in nature and constitute a substantial improvement over existing local convergence results. The rate of convergence results show that the multiplier method should be expected to converge considerably faster than the pure penalty method. At the same time, we construct a global duality framework for nonconvex optimization problems. The dual functional is concave, everywhere finite, and has strong differentiability properties. Furthermore, its value, gradient and Hessian matrix within an arbitrary bounded set can be obtained by unconstrained minimization of a certain augmented Lagrangian.

1. Introduction. One of the most effective methods for solving the constrained optimization problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } h_i(x)=0, \qquad i=1,\cdots,m, \end{aligned} \tag{1}$$

is the quadratic penalty function method (see, e.g., [6], [12], [13]). This method consists of sequential unconstrained minimization of the function

$$f(x)+\frac{c_k}{2}\sum_{i=1}^{m}[h_i(x)]^2 \tag{2}$$

for an increasing unbounded scalar sequence $\{c_k\}$. The properties of the method are well known, and we refer to [6] for an extensive discussion.

Recently a method, often referred to as the multiplier method, has been proposed and investigated by a number of authors [2]–[5], [7]–[11], [15]–[18] (see [2] and the survey papers [20], [21] for a more detailed account). In this method, the function

$$f(x)+\sum_{i=1}^{m}y_k^i h_i(x)+\frac{c_k}{2}\sum_{i=1}^{m}[h_i(x)]^2 \tag{3}$$

is minimized over x for a sequence of vectors $y_k=(y_k^1,\cdots,y_k^m)'$, and scalars c_k. The function above can be interpreted as a Lagrangian function to which a penalty term has been added. A number of ways of updating of the scalar c_k have been proposed. One possibility is to let c_k increase to infinity in a predetermined fashion. It is also possible to keep c_k fixed after a certain index. The distinctive feature of the method is that after each unconstrained minimization, yielding a

* Received by the editors September 14, 1973, and in revised form February 2, 1975.

† Department of Electrical Engineering and Coordinated Science Laboratory of the University of Illinois, Urbana, Illinois 61801. This research was carried out partly at the Department of Engineering-Economic Systems, Stanford University, Stanford, California, and supported by the National Science Foundation under Grant GK 32870, and partly at University of Illinois, Coordinated Science Laboratory and supported by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract DAAB-07-72-C-0259.

minimizing point x_k, the vector y_k is updated by means of the iteration

$$y_{k+1} = y_k + c_k h(x_k), \qquad i = 1, \cdots, m, \tag{4}$$

where $h(x_k)$ denotes the column vector $(h_1(x_k), \cdots, h_m(x_k))'$, (prime throughout this paper denotes transposition).

The convergence of iteration (4) to a Lagrange multiplier $\bar{y}$ of the problem has been shown under various assumptions. Global convergence results (i.e., results where the starting point y_0 is not required to be sufficiently close to $\bar{y}$) have been given for convex programming problems in [17], and in [3], [10], [11]. For nonconvex problems, the results available [2], [5], assume boundedness of the penalty parameter sequence $\{c_k\}$ and are local in nature; i.e., convergence has been shown under the assumption that the initial point y_0 is within a sufficiently small neighborhood of $\bar{y}$. Existing rate of convergence results [2] also assume boundedness of the sequence $\{c_k\}$.

All the results mentioned above have been obtained by interpreting the multiplier method as a primal-dual method. In this paper we adopt instead a penalty function viewpoint. Both the quadratic penalty method and the multiplier method are imbedded in a more general penalty function algorithm. In this algorithm, the augmented Lagrangian (3) is minimized for sequences of scalars $\{c_k\}$ and vectors $\{y_k\}$. The only requirement imposed on the sequence $\{y_k\}$ is that it remains within an arbitrary given bounded set S. Thus the quadratic penalty method is obtained as a special case by taking

$$c_k \to \infty \quad \text{and} \quad y_k = 0, \qquad \forall k.$$

The multiplier method is obtained by updating y_k via iteration (4), whenever $y_k + c_k h(x_k) \in S$.

Under assumptions which are specified in the next section, we show that for the general penalty method described above there exist nonnegative scalars c^* and M such that for all $c_k > c^*$ and $y_k \in S$, we have

$$\|x_k - \bar{x}\| \leqq M\|y_k - \bar{y}\|/c_k \tag{5}$$

and

$$\|y_{k+1} - \bar{y}\| \leqq M\|y_k - \bar{y}\|/c_k \tag{6}$$

where $\bar{x}$, $\bar{y}$ are the optimal solution and Lagrange multiplier vector for problem (1), x_k is a point locally minimizing the augmented Lagrangian (3) in a neighborhood of $\bar{x}$, and y_{k+1} is given in terms of c_k, y_k, and x_k by (4). The result mentioned above can be used to establish global convergence of the multiplier method, when S is, for example, an open sphere centered at $\bar{y}$, under the assumption that $c_k > M$, $c_k > c^*$ for all k greater than some index. Furthermore, the result shows that in the multiplier method, the sequence $\{\|y_k - \bar{y}\|\}$ converges at least linearly if c_k is bounded above and superlinearly if $c_k \to \infty$, while in the quadratic penalty method ($y_k = 0$), the convergence rate is much less favorable. A similar (but sharper) rate of convergence result has been shown in [2] under the assumption that c_k is bounded above.

From the computational point of view, it appears advantageous to carry out the minimization of the augmented Lagrangian only approximately while increasing the accuracy of the approximation with each minimization. We consider this case as well, and we obtain estimates similar to (5), (6), for two different gradient-based termination criteria. The estimates obtained are used in turn to establish global convergence and rate of convergence results for the corresponding algorithms.

In § 4 we use the results obtained to construct a global duality theory much in the spirit of the one recently proposed by Rockafellar [18]. However, our dual functional is continuously differentiable, and its value and gradient can be calculated by unconstrained minimization of the augmented Lagrangian (3) in a manner similar to that for convex programming problems. In this way we are able to interpret multiplier methods as primal-dual methods in a global sense.

For simplicity of presentation, we consider equality constraints only. The analysis, however, applies in its entirety to inequality constraints as well, since such constraints can be converted to equality constraints by using (squared) slack variables. This device, due to Rockafellar [16], results in no loss of computational efficiency and is discussed in § 5.

2. A generalized penalty function algorithm. Consider the nonlinear programming problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } h_i(x) = 0, \qquad i = 1, \cdots, m. \end{aligned} \tag{7}$$

The functions f and h_i for all i are real-valued functions on R^n (n-dimensional Euclidean space). Let $\bar{x}$ be an optimal solution of problem (7). We make the following assumptions concerning the nature of f and h_i in an open ball $B(\bar{x}, \varepsilon)$ of radius $\varepsilon > 0$ centered at $\bar{x}$.

A. The point $\bar{x}$ together with a unique Lagrange multiplier vector $\bar{y}$ satisfies the standard second order sufficiency conditions for $\bar{x}$ to be a local minimum [12, p. 226], i.e.,

A.1. The functions f, h_i, $i = 1, \cdots, m$, are twice continuously differentiable within the open ball $B(\bar{x}, \varepsilon)$.

A.2. The gradients $\nabla h_i(\bar{x})$, $i = 1, \cdots, m$, are linearly independent, and there exists a unique Lagrange multiplier vector $\bar{y} = (\bar{y}^1, \cdots, \bar{y}^m)'$ such that

$$\nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{y}^i \nabla h_i(\bar{x}) = 0.$$

A.3. The Hessian matrix of the Lagrangian $L_0(x, y) = f(x) + \sum_{i=1}^{m} y^i h_i(x)$,

$$\nabla^2 L_0(\bar{x}, \bar{y}) = \nabla^2 f(\bar{x}) + \sum_{i=1}^{m} \bar{y}^i \nabla^2 h_i(\bar{x}),$$

is positive definite on the tangent plane corresponding to the constraints, i.e.,

$$w' \nabla^2 L_0(\bar{x}, \bar{y}) w > 0$$

for all $w \in R^n$ such that

$$w \neq 0, \qquad w'\nabla h_i(\bar{x}) = 0, \qquad i = 1, \cdots, m.$$

B. The Hessian matrices $\nabla^2 f, \nabla^2 h_i$ are Lipschitz continuous within the open ball $B(\bar{x}, \varepsilon)$, i.e., for some $K > 0$, we have for all $x, x' \in B(\bar{x}, \varepsilon)$

$$\|\nabla^2 f(x) - \nabla^2 f(x')\| \leqq K\|x - x'\|$$

and

$$\|\nabla^2 h_i(x) - \nabla^2 h_i(x')\| \leqq K\|x - x'\|, \qquad i = 1, \cdots, m,$$

where $\|\cdot\|$ corresponds to the usual Euclidean norm.

Now let S be an arbitrary bounded subset of R^m. Consider also for any scalar $c > 0$ and any vector $y \in S$ the augmented Lagrangian function

$$L(x, y, c) = f(x) + y'h(x) + \frac{c}{2}\|h(x)\|^2. \tag{8}$$

We shall be interested in algorithms of the following general (and imprecise) form:

Step 1. Given $c_k > 0$, $y_k \in S$, find a (perhaps approximate) minimizing point x_k of the function $L(x, y_k, c_k)$ defined by (8).

Step 2. Determine $c_{k+1} > 0$, $y_{k+1} \in S$ on the basis of x_k, y_k, c_k according to some procedure and return to Step 1.

It is easy to verify that for every $x \in R^n$ we have

$$L(x, y_k, c_k) \geqq f(x) + \frac{c_k}{4}\|h(x)\|^2 - \frac{1}{c}\|y_k\|^2.$$

Hence, as $c_k \to \infty$, we have $L(x, y_k, c_k) \to \infty$ for all sequences $\{y_k\} \in S$, and all infeasible vectors x. It is, thus, evident that one may devise a penalty function method based on sequential unconstrained minimization of $L(x, y_k, c_k)$ for any sequences $\{c_k\} \to \infty, \{y_k\} \subset S$. This method exhibits the same convergence properties as the usual quadratic penalty function method [6]. Thus there is no difficulty in showing convergence of some sort for the general alogrithm described earlier whenever $c_k \to \infty$. The question which is most interesting, however, is to determine methods of updating y_k which result in desirable behavior such as accelerated convergence. Before proceeding to a detailed analysis, let us consider a heuristic geometric argument which shows that *it is advantageous to select* y_k *as close as possible to the Lagrange multiplier* $\bar{y}$.

Let p be the primal functional or perturbation function [19] corresponding to problem (7)

$$p(u) = \min_{h(x) = u} f(x)$$

In the above equation, the minimization is understood to be local within an appropriate neighborhood of $\bar{x}$. Also p is defined locally on a neighborhood of

$u = 0$. It is known that

$$p(0) = f(\bar{x}) = \text{optimal value of problem (7)},$$

$$\nabla p(0) = -\bar{y}.$$

Now we can write

$$\min_x L(x, y, c) = \min_u \min_{h(x)=u} \left\{ f(x) + y'h(x) + \frac{c}{2}\|h(x)\|^2 \right\}$$

or

$$\min_x L(x, y, c) = \min_u \left\{ p(u) + y'u + \frac{c}{2}\|u\|^2 \right\}.$$

The above equation can be interpreted geometrically as shown in Fig. 1. Notice that the addition of $(c/2)\|u\|^2$ to $p(u)$ has an important convexification effect. It can be seen from Fig. 1 that as y is closer to the Lagrange multiplier $\bar{y}$, the corresponding value $\min_x L(x, y, c)$ is closer to the optimal value of the problem. This fact will also be brought out by the analysis that follows.

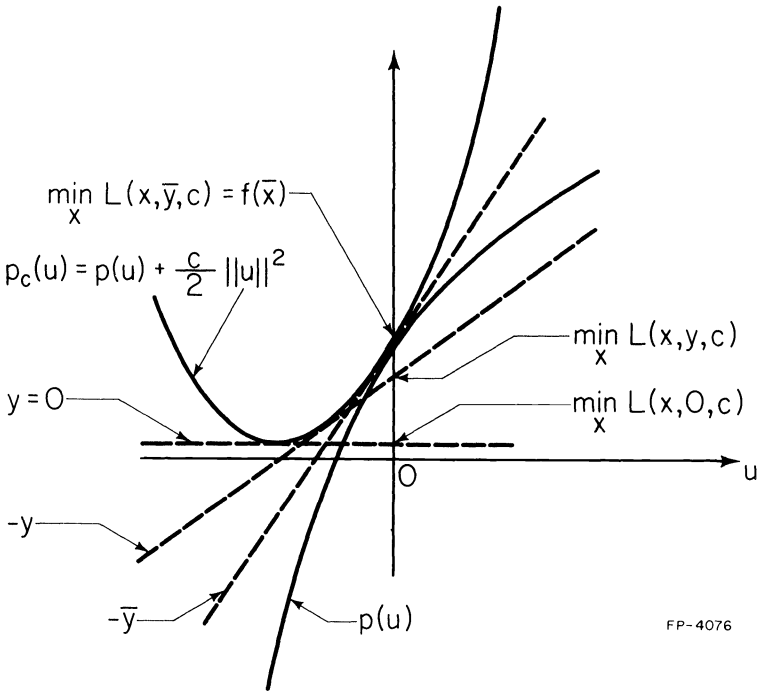


FIG. 1

The preceding argument leads us to the conclusion that the convergence of the generalized penalty algorithm could be accelerated if, at the kth minimization, a vector y_k, where $y_k \to \bar{y}$, were to be used in the augmented Lagrangian (8) in place of $y = 0$ which is used in the ordinary penalty function method. Such vectors are

readily available. It is easily shown that if $x(y, c)$ minimizes $L(x, y, c)$, then the vector

$$\tilde{y} = y + ch[x(y, c)]$$

is an approximation to the Lagrange multiplier $\bar{y}$ in the sense that $\lim_{c\to\infty} \tilde{y} = \bar{y}$. Thus we are led to a particular scheme whereby at the end of each minimization, the vector y is updated by means of the equation above. This iteration is identical to the one used in the method of multipliers. In the analysis that follows, it is shown that this iteration leads to a much faster convergence rate than the one of the ordinary penalty method. Furthermore, in order for the iteration to converge to $\bar{y}$, it is not necessary to increase c_k to infinity.

PROPOSITION 1. *There exists a scalar $c_1^* \geqq 0$ such that for every $c > c_1^*$, and $y \in S$, the augmented Lagrangian $L(x, y, c)$ of (8) has a unique minimizing point $x(y, c)$ with respect to x within some open ball centered at $\bar{x}$. Furthermore, for some scalar $M_1 > 0$ we have*

$$\|x(y, c) - \bar{x}\| \leqq \frac{M_1\|y - \bar{y}\|}{c} \quad \forall c > c_1^* \text{ and } y \in S \tag{9}$$

and

$$\|\tilde{y}(y, c) - \bar{y}\| \leqq \frac{M_1\|y - \bar{y}\|}{c} \quad \forall c > c_1^* \text{ and } y \in S, \tag{10}$$

where the vector $\tilde{y}(y, c) \in R^m$ is given by

$$\tilde{y}(y, c) = y + ch[x(y, c)]. \tag{11}$$

The proof of the above proposition is given in the next section. The result of the proposition has been proved for the case of the pure quadratic penalty method ($y = 0$) by Polyak [14] under the additional assumption that the Hessian matrix $\nabla^2 L_0$ in assumption A.3 is positive definite over the whole space, i.e., local strong convexity holds. Our proof is based in part on Polyak's analysis.

Some important conclusions can now be obtained from the result of Proposition 1. Assuming that $0 \in S$, we have that in the quadratic penalty method ($y_k = 0$), we obtain convergence if $c_k \to \infty$ and, furthermore, the sequences $\{x(0, c_k)\}, \{\tilde{y}(0, c_k)\}$ converge to $\bar{x}$, $\bar{y}$, respectively, at least as fast as $M_1\|\bar{y}\|/c_k$. It is evident, however, from the proposition that a great deal can be gained if the vector y_k is not held fixed but rather is updated by means of the iteration of the multiplier method

$$y_{k+1} = \tilde{y}(y_k, c_k) = y_k + c_k h[x(y_k, c_k)]. \tag{12}$$

In order to guarantee that the sequence $\{y_k\}$ remains bounded, we require that the updating takes place provided the resulting vector y_{k+1} belongs to the set S. Otherwise $y_{k+1} = y_k$, i.e., y_k is left unchanged. Of course, the choice of S is arbitrary, and in particular, we can assume that S contains $\bar{y}$ as an interior point. Under these circumstances, we have that if $c_k \to \infty$, then

$$\lim_{k\to\infty} \frac{\|y_{k+1} - \bar{y}\|}{\|y_k - \bar{y}\|} = 0,$$

i.e., the sequence $\{y_k\}$ converges to $\bar{y}$ superlinearly. If $c_k \to c < \infty$, where c is sufficiently large (large enough to ensure that $c > M_1$, $c > c_1^*$ and that $y_k + ch(x_k)$ belongs to an open sphere centered at $\bar{y}$ and contained in S), then

$$\limsup_{k\to\infty} \frac{\|y_{k+1} - \bar{y}\|}{\|y_k - \bar{y}\|} \leqq \frac{M_1}{c},$$

i.e., $\{y_k\}$ converges to $\bar{y}$ at least linearly with a convergence ratio inversely proportional to c

In conclusion, the method of multipliers defined by (12) converges from an arbitrary starting point within the bounded set S provided c_k is sufficiently large after some index $\bar{k}$, $\bar{y}$ is an interior point of S, and the unconstrained minimizations yield the points $x(y_k, c_k)$ for all $k \geqq \bar{k}$. In addition, the multiplier method offers distinct advantages over the quadratic penalty method in that it avoids the necessity of increasing c_k to infinity, and furthermore, the estimate of its convergence rate is much more favorable. For example, if $c_k = s^k$, $s > 1$, then for the penalty method, we have

$$\|x(0, c_k) - \bar{x}\| \leqq M_1 \|\bar{y}\| s^{-k},$$

while in the multiplier method with $y_0 = 0$,

$$\|x(y_k, c_k) - \bar{x}\| \leqq M_1^{k+1} \|\bar{y}\| s^{-(1+2+\cdots+k)}.$$

The ratio of the two bounds in the above inequalities is

$$\prod_{i=0}^{k-1} \frac{s^i}{M_1}$$

and tends to infinity as $k \to \infty$.

In order to avoid creating false impressions, it is perhaps worthwhile to emphasize the fact that the global convergence property of the method of multipliers concluded above is contingent upon the generation of the points $x(y_k, c_k)$, $k \geqq \bar{k}$, by the unconstrained minimization method employed. These points are, by Proposition 1, well-defined as local minimizing points of $L(x, y_k, c_k)$ which are closest to $\bar{x}$. Naturally $L(x, y_k, c_k)$ may have other local minimizing points to which the unconstrained minimization method may be attracted, and unless after some index the unconstrained minimization method stays in the neighborhood of the same local minimizing point of problem (7), our convergence analysis is invalid and there is no reason to believe that the method of multipliers should do better (or worse) than the penalty method. On the other hand, it should be noted that the usual practice of using the last point x_k of the kth minimization as the starting point of the $(k+1)$-st minimization is helpful in producing sequences $\{x_k\}$ which are close to one and the same local minimizing point of problem (7).

We now turn our attention to a generalized penalty method where, given c_k and y_k the augmented Lagrangian $L(x, y_k, c_k)$ of (8) is not minimized exactly, but rather the minimization process is terminated when a certain stopping criterion is satisfied. We consider two different stopping criteria. Similar criteria have been considered in the past in the context of penalty [14] and multiplier methods [2], [3], [5], [10]. According to the first criterion, minimization of $L(x, y_k, c_k)$ is

terminated at a point x_k satisfying

$$\|\nabla L(x_k, y_k, c_k)\| \leqq \gamma_k / c_k, \tag{13}$$

where $\{\gamma_k\}$ is a bounded sequence with $\gamma_k \geqq 0$. According to the second criterion minimization is terminated at a point x_k satisfying

$$\|\nabla L(x_k, y_k, c_k)\| \leqq \min \{\gamma_k / c_k, \gamma_k' \|h(x_k)\|\}, \tag{14}$$

where $\{\gamma_k\}, \{\gamma_k'\}$ are bounded with $\gamma_k \geqq 0$, $\gamma_k' \geqq 0$.

We have the following proposition, the proof of which is given in the next section.

PROPOSITION 2. *There exists a scalar $c_2^* \geqq 0$ such that for every $c > c_2^*$ and $y \in S$ and every vector $a \in R^n$ with $\|a\| \leqq \gamma_k / c$, there exists a unique point $x_a(y, c)$ within some open ball centered at $\bar{x}$ satisfying*

$$\nabla L[x_a(y, c), y, c] = a. \tag{15}$$

Furthermore, for some scalar $M_2 > 0$ we have

$$\|x_a(y, c) - \bar{x}\| \leqq \frac{M_2(\|y - \bar{y}\|^2 + \gamma_k^2)^{1/2}}{c} \quad \forall c > c_2^*,\ y \in S \text{ and } \|a\| \leqq \frac{\gamma_k}{c} \tag{16}$$

and

$$\|\tilde{y}_a(y, c) - \bar{y}\| \leqq \frac{M_2(\|y - \bar{y}\|^2 + \gamma_k^2)^{1/2}}{c} \quad \forall c > c_2^*,\ y \in S \text{ and } \|a\| \leqq \frac{\gamma_k}{c}, \tag{17}$$

where $\tilde{y}_a$ is given by

$$\tilde{y}_a(y, c) = y + ch[x_a(y, c)]. \tag{18}$$

If, in addition, a and $x_a(y, c)$ satisfy

$$\|a\| \leqq \gamma_k' \|h[x_a(y, c)]\| \tag{19}$$

then we have

$$\|x_a(y, c) - \bar{x}\| \leqq \frac{M_2(4(\gamma_k')^2 + 1)^{1/2} \|y - \bar{y}\|}{c} \quad \forall c > c_2^* \text{ and } y \in S, \tag{20}$$

and

$$\|\tilde{y}_a(y, c) - \bar{y}\| \leqq \frac{M_2(4(\gamma_k')^2 + 1)^{1/2} \|y - \bar{y}\|}{c} \quad \forall c > c_2^* \text{ and } y \in S. \tag{21}$$

The proposition above may now be used to establish convergence and rate of convergence results for the iteration

$$y_{k+1} = y_k + c_k h(x_k). \tag{22}$$

This iteration takes place if $y_k + c_k h(x_k) \in S$. Otherwise $y_{k+1} = y_k$, i.e., y_k is left unchanged. The point x_k satisfies either the criterion

$$\|\nabla L(x_k, y_k, c_k)\| \leqq \gamma_k / c_k \tag{23}$$

or the criterion

$$\|\nabla L(x_k, y_k, c_k)\| \leqq \min \{\gamma_k / c_k, \gamma_k' \|h(x_k)\|\}. \tag{24}$$

Furthermore, x_k is the unique point $x_a(y_k, c_k)$ corresponding to $a = \nabla L(x_k, y_k, c_k)$ and closest to $\bar{x}$ in accordance with Proposition 2. It is assumed that the unconstrained minimization algorithm yields such points after a certain index.

It is clear from Proposition 2 that any sequence $\{x_k, y_k\}$ generated by the iteration (22) and the termination criterion (24) converges to $(\bar{x}, \bar{y})$, provided c_k is sufficiently large after a certain index and $\bar{y}$ is an interior point of S. Furthermore, $\{y_k\}$ converges to $\bar{y}$ at least linearly when $c_k \to c < \infty$, and superlinearly when $c_k \to \infty$. However, for the termination criterion (23), linear convergence cannot be guaranteed, and in fact an example given in [2] shows that convergence may not be linear. In addition, for this termination criterion it is necessary to increase c_k to infinity in order to achieve global convergence unless $\{\gamma_k\}$ is a sequence converging to zero.

3. Proofs of Propositions 1 and 2.

Proof of Proposition 1. The proof proceeds in two parts. We first prove the proposition under the following condition:

C. The Hessian matrix of the ordinary Lagrangian function

$$\nabla^2 L_0(\bar{x}, \bar{y}) = \nabla^2 f(\bar{x}) + \sum_{i=1}^{m} \bar{y}^i \nabla^2 h_i(\bar{x})$$

is a positive definite matrix; i.e., local strong convexity holds.

Subsequently, we extend the result to the general case.

Let C hold. For all $x \in B(\bar{x}, \varepsilon)$ and any fixed $y \in S$, $c > 0$, consider the auxiliary variables

$$p = x - \bar{x}, \qquad q = y + ch(x) - \bar{y}, \tag{25}$$

where $h(x)$ is the m-vector with coordinates $h_i(x)$, $i = 1, \cdots, m$. For every $x \in B(\bar{x}, \varepsilon)$ we have

$$\nabla f(x) = \nabla f(\bar{x}) + \nabla^2 f(\bar{x}) p + r_1(p) \tag{26}$$

$$\nabla h_i(x) = \nabla h_i(\bar{x}) + \nabla^2 h_i(\bar{x}) p + r_2^i(p), \qquad i = 1, \cdots, m, \tag{27}$$

where r_1 and r_2^i are n-vector valued functions of p satisfying

$$\begin{aligned} &r_1(0) = r_2^i(0) = 0, \qquad i = 1, \cdots, m, \\ &\nabla r_1(p) = \nabla^2 f(x) - \nabla^2 f(\bar{x}), \\ &\nabla r_2^i(p) = \nabla^2 h_i(x) - \nabla^2 h_i(\bar{x}), \qquad i = 1, \cdots, m. \end{aligned} \tag{28}$$

By the Lipschitz condition assumption B, we have for all $\|p\| < \varepsilon$,

$$\|\nabla r_1(p)\| \leqq K\|p\| \tag{29}$$

and

$$\|\nabla r_2^i(p)\| \leqq K\|p\| \quad \forall i = 1, \cdots, m. \tag{30}$$

Consider now the augmented Lagrangian $L(x, y, c)$ of (8). We have, by (25), (26)

and (27),

$$\nabla L(x, y, c) = \nabla f(x) + \nabla h(x)[y + ch(x)] = \nabla f(\bar{x}) + \nabla^2 f(\bar{x})p + r_1(p)$$
$$+ \sum_{i=1}^{m} (q^i + \bar{y}^i)[\nabla h_i(\bar{x}) + \nabla^2 h_i(\bar{x})p + r_2^i(p)],$$

or equivalently,

$$\nabla L(x, y, c) = \nabla^2 L_0(\bar{x}, \bar{y})p + \nabla h(\bar{x})q + r_3(p, q), \tag{31}$$

where $\nabla h(\bar{x})$ is the $n \times m$ matrix with columns $\nabla h_i(\bar{x})$, and $r_3(p, q) \in R^n$ is the vector defined by

$$r_3(p, q) = r_1(p) + \sum_{i=1}^{m} (q^i + \bar{y}^i) r_2^i(p) + \sum_{i=1}^{m} q^i \nabla^2 h_i(\bar{x})p. \tag{32}$$

We also have from (25),

$$\frac{q + \bar{y} - y}{c} = h(x) = h(\bar{x}) + \nabla h(\bar{x})'p + r_4(p),$$

or equivalently,

$$\nabla h(\bar{x})'p - \frac{1}{c}q = \frac{1}{c}(\bar{y} - y) - r_4(p), \tag{33}$$

where the function $r_4: B(0, \varepsilon) \to R^m$ satisfies

$$r_4(0) = 0, \qquad \nabla r_4^i(p) = \nabla h_i(x) - \nabla h_i(\bar{x}), \qquad i = 1, \cdots, m, \tag{34}$$

and using the Lipschitz condition assumption B,

$$\|\nabla r_4^i(p)\| \leqq (K\|p\| + \|\nabla^2 h_i(\bar{x})\|)\|p\| \leqq (K\varepsilon + \|\nabla^2 h_i(\bar{x})\|)\|p\|. \tag{35}$$

Combining now (31) and (33), we have that in order for a point $x \in B(\bar{x}, \varepsilon)$ to satisfy $\nabla L(x, y, c) = 0$ it is necessary and sufficient that the corresponding point $s = [p', q']'$, as given by (25), solves the equation

$$As = t + r(s), \tag{36}$$

where we use the notation

$$A = \begin{bmatrix} \nabla^2 L_0(\bar{x}, \bar{y}) & \nabla h(\bar{x}) \\ \nabla h(\bar{x})' & -\frac{I}{c} \end{bmatrix}, \quad s = \begin{bmatrix} p \\ q \end{bmatrix}, \quad t = \begin{bmatrix} 0 \\ \frac{\bar{y} - y}{c} \end{bmatrix}, \quad r(s) = \begin{bmatrix} -r_3(p, q) \\ -r_4(p) \end{bmatrix} \tag{37}$$

and I is the $m \times m$ identity matrix. Concerning $r(s)$, we have, from (28), (32), (34) and (35),

$$r(0) = 0. \tag{38}$$

Furthermore, for any s corresponding to an $x \in B(\bar{x}, \varepsilon)$ we have, by straightforward calculation from (29), (30), (32), (35) and (37),

$$\|\nabla r(s)\| \leqq \alpha\|s\|, \tag{39}$$

where $\alpha > 0$ is a constant depending only on ε.

The proof now follows the pattern of [14] by showing that (36) has a unique solution within the domain of definition of s for any $y \in S$ and $c > c_1^*$, where c_1^* is a sufficiently large constant. We make use of the following two lemmas due to Polyak [14].

LEMMA 1. *The matrix A of* (37) *has an inverse for every $c > 0$. Furthermore, the inverse is uniformly bounded, i.e., for some $M_1 > 0$ and all $c > 0$,*

$$\|A^{-1}\| \leqq 2M_1. \tag{40}$$

LEMMA 2. *The equation* (36), *$As = t + r(s)$, has a unique solution s^* within the open ball $B(0, 8M_1\|t\|) \subset B(0, \varepsilon)$ for every $y \in S$ and every c sufficiently large to guarantee that*

$$\|t\| \leqq \min\left\{\frac{1}{16M_1\alpha}, \frac{\varepsilon}{8M_1}\right\},$$

where α, M_1 are as in (39), (40). *The solution s^* satisfies $\|s^*\| \leqq M_1\|t\|$.*

Now from Lemma 2 and the definition (25), (36), (37), it follows immediately that for every $y \in S$ and $c > c_1^*$, where c_1^* is a sufficiently large constant, the equation

$$\nabla L(x, y, c) = 0 \tag{41}$$

has a unique solution $x(y, c)$ within an open ball centered at $\bar{x}$ satisfying

$$\|x(y, c) - \bar{x}\| \leqq \frac{M_1\|y - \bar{y}\|}{c}, \tag{42}$$

$$\|y + ch[x(y, c)] - \bar{y}\| \leqq \frac{M_1\|y - \bar{y}\|}{c}. \tag{43}$$

Hence, in order to complete the proof of Proposition 1, we only need to show that for c sufficiently large the point $x(y, c)$ is a local minimum of $L(x, y, c)$. To this end, it is sufficient to show that $\nabla^2 L[x(y, c), y, c]$ is positive definite for all $y \in S$ and c sufficiently large. Indeed, we have

$$\nabla^2 L[x(y, c); y, c] = \nabla^2 f[x(y, c)] + \sum_{i=1}^{m} (y^i + ch_i[x(y, c)])\nabla^2 h_i[x(y, c)] + c\nabla h[x(y, c)]\nabla h[x(y, c)]'.$$

Now the third term in the above expression is a positive semidefinite matrix. The sum of the first two terms, in view of (42), (43), is arbitrarily close to the positive matrix $\nabla^2 L_0(\bar{x}, \bar{y})$ for sufficiently large c. Hence, for all c greater than some c_1^*, $\nabla^2 L[x(y, c), y, c]$ is positive definite and $x(y, c)$ is a local minimum of $L(x, y, c)$. Thus Proposition 1 has been proved under condition C.

In order to extend the proof of Proposition 1 to the general case where condition C is not satisfied, we convert the general nonlinear programming problem (7) to an equivalent locally convex problem for which condition C is satisfied. We achieve local convexity by adding a sufficiently high penalty term to the objective function as first indicated by Arrow and Solow [1].

It is evident that problem (7) is equivalent for every $\mu \geqq 0$ to the following problem

$$\begin{aligned} &\text{minimize } f(x)+\frac{\mu}{2}\|h(x)\|^2 \\ &\text{subject to } h_i(x)=0, \qquad i=1,\cdots,m. \end{aligned} \tag{44}$$

Problem (44) has $\bar{x}$ as an optimal solution and $\bar{y}$ as Lagrange multiplier vector. Now consider the Hessian with respect to x of the ordinary Lagrangian of problem (44). We have

$$\nabla^2 L_\mu(\bar{x}, \bar{y}) = \nabla^2 L_0(\bar{x}, \bar{y}) + \mu[\nabla h(\bar{x}) \nabla h(\bar{x})'], \tag{45}$$

where $\nabla^2 L_0(\bar{x}, \bar{y})$ is the Hessian of the ordinary Lagrangian of problem (7). Using assumption A.3, we have the following easily proved lemma.

LEMMA 3. *There exists a scalar $\mu^* > 0$ such that for every $\mu \geqq \mu^*$, the matrix $\nabla^2 L_\mu(\bar{x}, \bar{y})$ of (45) is positive definite.*

The immediate consequence of the above lemma is that problem (44) satisfies the local convexity condition C for all $\mu \geqq \mu^*$. We apply now the result of Proposition 1 as proved under C and with c replaced by $c - \mu^*$, to problem (44) with $\mu = \mu^*$. We have the following:

There exists $c^* \geqq 0$ such that for all $c - \mu^* > c^*$, $y \in S$ the augmented Lagrangian

$$L(x, y, c) = f(x) + y'h(x) + \frac{c}{2}\|h(x)\|^2 \tag{46}$$

has a unique unconstrained minimum $x(y, c)$ within some open ball centered at $\bar{x}$. From the estimates (9), (10), we obtain that for some constant $M > 0$,

$$\|x(y, c) - \bar{x}\| \leqq \frac{M\|y - \bar{y}\|}{c - \mu^*}, \tag{47}$$

$$\|\tilde{y}(y, c) - \bar{y} - \delta(y, c)\| \leqq \frac{M\|y - \bar{y}\|}{c - \mu^*}, \tag{48}$$

where $\tilde{y}$ is given by (11) and the vector $\delta(y, c) \in R^m$ is given by

$$\delta(y, c) = \mu^* h[x(y, c)]. \tag{49}$$

From the equation above and (47), it follows that for some constant $B > 0$

$$\|\delta(y, c)\| \leqq \mu^* B\|x(y, c) - \bar{x}\| \leqq \frac{\mu^* BM\|y - \bar{y}\|}{c - \mu^*}. \tag{50}$$

Combining the inequalities (47), (48) and (50), we have

$$\|x(y, c) - \bar{x}\| \leqq \frac{M\|y - \bar{y}\|}{c - \mu^*} \tag{51}$$

and

$$\|\tilde{y}(y,c)-\bar{y}\| \leqq \frac{M\|y-\bar{y}\|}{c-\mu^*} + \|\delta(y,c)\| \leqq \frac{(M+\mu^* BM)\|y-\bar{y}\|}{c-\mu^*}. \tag{52}$$

Let $M_1 > 0$ and $c_1^* \geqq c^* + \mu^* > 0$ be any constants such that

$$\frac{M}{c-\mu^*} \leqq \frac{M+\mu^* BM}{c-\mu^*} \leqq \frac{M_1}{c} \quad \forall c \geqq c_1^*. \tag{53}$$

Then the desired estimates (9) and (10) follow immediately from (51), (52) and (53), and the proposition is completely proved. Q.E.D.

Proof of Proposition 2. The proof of Proposition 2 follows similar lines as the proof of Proposition 1. Again we assume first that C holds. For an y $a \in R^n$ with $\|a\| \leqq \gamma_k/c$, we have in place of (36) the equation

$$As = t_a + r(s), \tag{54}$$

where A, s, $r(s)$ are as in (37) and t_a is given by

$$t_a = \begin{bmatrix} a \\ \dfrac{\bar{y}-y}{c} \end{bmatrix}. \tag{55}$$

Now from Lemma 2 with t_a in place of t and using the fact that $\{\gamma_k\}$ is bounded and $\|a\| \leqq \gamma_k/c$, we obtain for some $M_2 > 0$ and all $y \in S$, $c \geqq c_2^*$, where c_2^* is a sufficiently large positive scalar (cf. (42) and (43)),

$$\|x_a(y,c)-\bar{x}\| \leqq M_2\left(\frac{\|y-\bar{y}\|^2}{c^2} + \|a\|^2\right)^{1/2} \leqq \frac{M_2(\|y-\bar{y}\|^2+\gamma_k^2)^{1/2}}{c} \tag{56}$$

and

$$\|\tilde{y}_a(y,c)-\bar{y}\| \leqq M_2\left(\frac{\|y-\bar{y}\|^2}{c^2} + \|a\|^2\right)^{1/2} \leqq \frac{M_2(\|y-\bar{y}\|^2+\gamma_k^2)^{1/2}}{c}, \tag{57}$$

which are the relations to be proved.

Now assume, in addition, that a and $x_a(y,c)$ satisfy

$$\|a\| \leqq \gamma_k' \|h[x_a(y,c)]\|. \tag{58}$$

Then we have, from (56), (57) and (58),

$$\|x_a(y,c)-\bar{x}\| \leqq M_2\left(\frac{\|y-\bar{y}\|^2}{c^2} + (\gamma_k')^2\|h[x_a(y,c)]\|^2\right)^{1/2} \tag{59}$$

and

$$\|\tilde{y}_a(y,c)-\bar{y}\| \leqq M_2\left(\frac{\|y-\bar{y}\|^2}{c^2} + (\gamma_k')^2\|h[x_a(y,c)]\|^2\right)^{1/2}. \tag{60}$$

Using (18), the last relation is written as

$$\|y-\bar{y}+ch[x_a(y,c)]\| \leqq M_2\left(\frac{\|y-\bar{y}\|^2}{c^2}+(\gamma_k')^2\|h[x_a(y,c)]\|^2\right)^{1/2},$$

from which

$$c\|h[x_a(y,c)]\| \leqq M_2\left(\frac{\|y-\bar{y}\|^2}{c^2}+(\gamma_k')^2\|h[x_a(y,c)]\|^2\right)^{1/2}+\|y-\bar{y}\|$$

$$\leqq \left(\frac{M_2}{c}+1\right)\|y-\bar{y}\|+M_2\gamma_k'\|h[x_a(y,c)]\|.$$

Thus, finally, we have

$$\|h[x_a(y,c)]\| \leqq \frac{c+M_2}{c(c-M_2\gamma_k')}\|y-\bar{y}\|.$$

For $c \geqq (1+2\gamma_k')M_2$, the inequality above yields

$$\|h[x_a(y,c)]\| \leqq \frac{2\|y-\bar{y}\|}{c}. \tag{61}$$

Substitution of (61) in (59) and (60) yields

$$\|x_a(y,c)-\bar{x}\| \leqq \frac{M_2(1+4(\gamma_k')^2)^{1/2}\|y-\bar{y}\|}{c}$$

and

$$\|\bar{y}_a(y,c)-\bar{y}\| \leqq \frac{M_2(1+4(\gamma_k')^2)^{1/2}\|y-\bar{y}\|}{c},$$

which are the desired estimates. Thus Proposition 2 is proved under condition C. The extension to the general case is entirely similar to the corresponding extension in Proposition 1 and is omitted. Q.E.D.

4. A global duality framework for the method of multipliers. In this section we utilize the results of § 2 to construct a duality framework for problem (7). In contrast with past formulations for nonconvex problems (see, e.g., [5], [12]), the framework is *global* in nature (at least in as much as the dual variables are concerned). By this we mean that the dual functional is an everywhere defined real-valued concave function. The theory is similar in spirit with the one recently proposed by Rockafellar [18] under weaker assumptions, and the one of Buys [5] which is local in nature. Our construction, however, is more suitable to the analysis of algorithms since in our case *the dual functional has strong differentiability properties.* Furthermore, its value and derivatives within an arbitrary open bounded set may be computed by unconstrained local minimization of the augmented Lagrangian. In this way the iteration of the multiplier method can be interpreted as a gradient iteration in a global sense.

For any vector $u \in R^m$, consider the minimization problem

$$\min_{h(x)=u} f(x). \tag{62}$$

Now by applying the implicit function theorem to the system of equations

$$\nabla f(x)+\sum_{i=1}^{m} y_i \nabla h_i(x)=0, \qquad h_i(x)=u_i, \quad i=1,\cdots,m,$$

and using assumption A, we have the following lemma.

LEMMA 4. *Under assumption* A, *there exist positive scalars* β *and* δ *such that for every* u *with* $\|u\|<\beta$, *problem* (62) *has a unique solution* $x(u)$ *within the open ball* $B(\bar{x}, \delta)$ *with a Lagrange multiplier* $y(u)$ *satisfying* $\|y(u)-\bar{y}\|<\delta$. *Furthermore, the functions* $x(u)$, $y(u)$ *are continuously differentiable within the open ball* $B(0, \beta)$ *and satisfy* $x(0)=\bar{x}$, $y(0)=\bar{y}$.

We define now the *primal functional* $p: B(0, \beta) \to R$ by means of

$$p(u)=\min_{\substack{h(x)=u \\ x\in B(\bar{x},\delta)}} f(x)=f[x(u)]. \tag{63}$$

It follows from the implicit function theorem that

$$\nabla p(u)=-y(u), \qquad u\in B(0,\beta), \tag{64}$$

and, since $y(u)$ is continuously differentiable, we have that p is twice continuously differentiable on $B(0, \beta)$. Without loss of generality, we assume that the Hessian matrix of p is uniformly bounded on $B(0, \beta)$.

Now for any $c \geqq 0$, consider the function

$$p_c(u)=p(u)+\frac{c}{2}\|u\|^2.$$

It is clear that for c sufficiently large, the Hessian matrix of p_c is positive definite on $B(0, \beta)$, and hence p_c is strictly convex on $B(0, \beta)$. We define for such c the *dual functional* $d_c: R^m \to R$ by means of

$$d_c(y)=\inf_{u\in B(0,\beta)}\left\{p(u)+\frac{c}{2}\|u\|^2+y'u\right\}=\inf_{u\in B(0,\beta)}\{p_c(u)+y'u\}. \tag{65}$$

We note that this way of defining the dual functional is not unusual, since it corresponds to a perturbation function taking the value $p_c(u)$ on $B(0, \beta)$ and $+\infty$ outside $B(0, \beta)$.

The function d_c of (65) has the following properties which we state as a proposition.

PROPOSITION 3. *Under assumption* A, *for every* c *for which the Hessian matrix of* p_c *is positive definite on* $B(0, \beta)$, *we have the following*:

(a) *The function* d_c *is a real-valued, everywhere continuously differentiable concave function. Furthermore, it is twice continuously differentiable on the open set* $A=\{y \mid y=-\nabla p_c(u),\ u\in B(0,\beta)\}$.

(b) *For any* $y\in A$, *the infimum in* (65) *is attained at a unique point* $u_y \in B(0, \beta)$, *and we have* $\nabla d_c(y)=u_y$, $\nabla^2 d_c(y)=-[\nabla^2 p_c(u_y)]^{-1}$.

(c) *The function* d_c *has a unique maximizing point, the Lagrange multiplier* $\bar{y}$.

Proof. (a) Consider the closure [19, § 7] of the convex function taking the value $p_c(u)$ on $B(0, \beta)$ and $+\infty$ outside $B(0, \beta)$. The closure is essentially strictly

convex [19, § 26], and its effective domain is a compact set. Hence its conjugate convex function is real-valued and continuously differentiable [19, Thm. 26.3]. Since this conjugate is simply $-d_c(-y)$, we have that d_c is everywhere finite and continuously differentiable. Also by the conjugacy relation between p_c and d_c, we have

$$\begin{aligned} &\nabla d_c[-\nabla p_c(u)] = u \quad \forall u \in B(0, \beta), \\ &\nabla p_c[\nabla d_c(y)] = -y \quad \forall y \in A, \end{aligned} \tag{66}$$

where

$$A = \{y | y = -\nabla p_c(u),\ u \in B(0, \beta)\} = \{y | \nabla d_c(y) = u,\ u \in B(0, \beta)\}.$$

The set on the right above is open by the continuity of ∇d_c, thus implying that the set A is open. Now let $\tilde{y}$ be any point in A and let $\tilde{u} = \nabla d_c(\tilde{y})$. We have that $\tilde{u} \in B(0, \beta)$ and

$$\nabla p_c(\tilde{u}) = -\tilde{y}.$$

Applying the implicit function theorem in the equation above, we have that there exists an open ball $B(\tilde{y}, \lambda) \subset A$ and a continuously differentiable function $u(\cdot) : B(\tilde{y}, \lambda) \to B(0, \beta)$ such that $u(\tilde{y}) = \tilde{u}$ and

$$\nabla p_c[u(y)] = -y.$$

It follows from (66) that

$$\nabla d_c(y) = u(y) \quad \forall y \in B(\tilde{y}, \lambda).$$

Since $u(y)$ is continuously differentiable on $B(\tilde{y}, \lambda)$, so is $\nabla d_c(y)$. Hence d_c is twice continuously differentiable at $\tilde{y}$. Since $\tilde{y}$ is an arbitrary point in A, we have that d_c is twice continuously differentiable on A, which was to be proved.

(b) The fact that for $y \in A$ the infimum in (65) is attained at a unique point is evident from the argument above. The formula $\nabla^2 d_c(y) = -[\nabla^2 p_c(u_y)]^{-1}$ follows from (66).

(c) We have, by (66) and the fact that $\nabla p_c(0) = -\bar{y}$,

$$\nabla d_c(\bar{y}) = 0,$$

and hence $\bar{y}$ is a maximizing point of d_c. It is a unique maximizing point by the differentiability of p_c. Q.E.D.

We now proceed to show that the value and the derivatives of the dual functional d_c can be obtained by local minimization of the augmented Lagrangian $L(x, y, c)$ of (8) provided c is sufficiently large. Let S be any open bounded subset of R^m. Then for any $y \in S$, by Proposition 1, we have that for c sufficiently large,

$$\|x(y, c) - \bar{x}\| = \frac{M_1 \|y - \bar{y}\|}{c} < \delta,$$

$$\|\tilde{y}(y, c) - \bar{y}\| \leqq \frac{M_1 \|y - \bar{y}\|}{c} < \delta, \qquad \|\tilde{u}\| < \beta,$$

where

$$\tilde{y}(y, c) = y + ch[x(y, c)], \qquad \tilde{u} = h[x(y, c)].$$

Furthermore, we have

$$\nabla f[x(y, c)] + \sum_{i=1}^{m} \tilde{y}^i(y, c)\nabla h_i[x(y, c)] = 0.$$

It follows from the implicit function theorem and Lemma 4 that $x(y, c)$ is the unique minimizing point in problem (62) when $u = \tilde{u}$. This implies

$$p(\tilde{u}) = f[x(y, c)], \qquad \nabla p(\tilde{u}) = -\tilde{y}(y, c) = -y - c\tilde{u},$$

and therefore

$$\nabla p_c(\tilde{u}) + y = 0.$$

Hence $y \in A$, $\tilde{u}$ attains the infimum in the right-hand side of (65), and by part (b) of Proposition 3,

$$\nabla d_c(y) = \tilde{u} = h[x(y, c)], \qquad \nabla^2 d_c(y) = -[\nabla^2 p_c(\tilde{u})]^{-1}.$$

Furthermore,

$$d_c(y) = p(\tilde{u}) + y'\tilde{u} + \frac{c}{2}\|\tilde{u}\|^2$$

$$= f[x(y, c)] + y'h[x(y, c)] + \frac{c}{2}\|h[x(y, c)]\|^2 = \min_x L(x, y, c),$$

where the minimization above is understood to be local in the sense of Proposition 1. In addition, a straightforward calculation [5], [12] yields

$$D_c(y) = \nabla^2 d_c(y) = -\nabla h[x(y, c)]'\{\nabla^2 L[x(y, c), y, c]\}^{-1}\nabla h[x(y, c)], \tag{67}$$

where $\nabla h[x(y, c)]$ is the $n \times m$ matrix having as columns the gradients $\nabla h_i[x(y, c)]$, $i = 1, \cdots, m$, and $\nabla^2 L$ denotes the Hessian matrix of the augmented Lagrangian L with respect to x. Thus we have proved the following proposition.

PROPOSITION 4. *Let S be any open bounded subset of R^m, and let assumptions* A *and* B *hold. Then there exists a scalar $c^* \geq 0$ such that for every $y \in S$ and every $c > c^*$, the dual functional d_c satisfies*

$$d_c(y) = f[x(y, c)] + y'h[x(y, c)] + \frac{c}{2}\|h[x(y, c)]\|^2 = \min_x L(x, y, c),$$

$$\nabla d_c(y) = h[x(y, c)],$$

where $x(y, c)$ is as in Proposition 1. Furthermore, d_c is twice continuously differentiable on S and $\nabla^2 d_c(y)$ is given by (67).

It is now clear that the iteration of the method of the multipliers can be written, for c sufficiently large,

$$y_{k+1} = y_k + c\nabla d_c(y_k),$$

and hence can be viewed as a *fixed step size gradient iteration* for maximizing the

dual functional d_c. Thus one may obtain a tight rate of convergence result by utilizing a known result on gradient methods. This result, however, is rather uniformative since it involves the eigenvalues of the matrix D_c of (67) which strongly depend on c. A modified version of this result which is more amenable to proper interpretation is given in [2], together with an analysis of the convergence rate aspects of the method of multipliers in the presence of inexact minimization.

The primal-dual interpretation of the multiplier method suggests also several possibilities for modification of the basic iteration. One such modification was suggested in [2], [3]. Another interesting possibility rests on the fact that when second derivatives are calculated during the unconstrained minimization cycle, then one obtains the Hessian matrix D_c of (67) in addition to the gradient ∇d_c. Thus it is possible to carry out a Newton iteration aimed at maximizing d_c in place of the gradient iteration corresponding to the method of multipliers. It is also possible to use a variable metric method for maximization of d_c. Such possibilities have already been suggested by Buys [5], who in addition provided some local convergence results. It is to be noted, however, that for large scale problems arising, for example, in optimal control, where the number of primal and dual variables may easily reach several hundreds or even thousands, such modifications do not seem to be attractive. This is particularly so since the simple gradient iteration already has excellent convergence rate.

5. Treatment of inequality constraints. As pointed out in the Introduction, inequality constraints may be treated in a simple way by introducing slack variables. Indeed, the problem

(68)
$$\min_{g_j(x)\leqq 0,\quad j=1,\cdots,r} f(x)$$

is equivalent to the equality constrained problem

(69)
$$\min_{g_j(x)+z_j^2=0,\quad j=1,\cdots,r} f(x),$$

where $z_1, \cdots, z_r$ represent additional variables.

Now assume that $(\bar{x}, \bar{y})$ is an optimal solution–Lagrange multiplier pair for problem (68) satisfying the following second order sufficiency conditions for optimality (which include strict complementarity).

A′. The functions $f, g_j, j=1, \cdots, r$, are twice continuously differentiable within an open ball $B(\bar{x}, \varepsilon)$. The gradients $\nabla g_j(\bar{x}), j \in J(\bar{x})$, with $J(\bar{x}) = \{j | g_j(x) = 0\}$, are linearly independent. We have $\nabla f(x) + \sum_{j=1}^{r} \bar{y}^j \nabla g_j(\bar{x}) = 0$ and $\bar{y}^j \geqq 0$ with $\bar{y}^j > 0$ if and only if $j \in J(\bar{x})$. Furthermore,

$$w'\left[\nabla^2 f(\bar{x}) + \sum_{j=1}^{r} \bar{y}^j \nabla^2 g_j(\bar{x})\right] w > 0$$

for all $w \neq 0$ such that $w' \nabla g_j(\bar{x}) = 0$ for all $j \in J(\bar{x})$.

Then it is easy to show that $(\bar{x}, |g_1(\bar{x})|^{1/2}, \cdots, |g_r(\bar{x})|^{1/2})$ is an optimal solution of problem (69) satisfying (together with $\bar{y}$) assumption A and hence it is covered by the theory of §§ 2 and 3 provided the Lipschitz assumption B is also satisfied. Thus one may use the multiplier method for solving problem (69) instead of

problem (68). On the other hand, slack variables need not be present explicitly in the computations, since the minimization of the augmented Lagrangian,

$$L(x, z, y, c) = f(x) + \sum_{j=1}^{r} y^j[g_j(x) + z_j^2] + \frac{c}{2}\sum_{j=1}^{r} [g_j(x) + z_j^2]^2,$$

can be carried out first with respect to $z_1, \cdots, z_r$, yielding

$$\tilde{L}(x, y, c) = \min_z L(x, z, y, c)$$

$$= f(x) + \frac{1}{2c}\left\{\sum_{j=1}^{r} [\max(0, y^j + cg_j(x))]^2 - (y^j)^2\right\}.$$

The optimal values of z_j are given in terms of x, y, c by

$$z_j^2(x, y, c) = \max[0, -y^j/c - g_j(x)], \qquad j = 1, \cdots, r. \tag{70}$$

Now minimization of $\tilde{L}(x, y, c)$ with respect to x yields a vector $x(y, c)$, and the multiplier method iteration in view of (70) takes the form

$$\begin{aligned} y_{k+1}^j &= y_k^j + c[g_j[x(y, c)] + z_j^2[x(y, c), y, c]] \\ &= \max[0, y_k^j + cg_j[x(y, c)]], \qquad j = 1, \cdots, r. \end{aligned} \tag{71}$$

Also in view of (70), the stopping criterion (14) takes the form

$$\|\nabla \tilde{L}(x_k, y_k, c_k)\| \leqq \min\left\{\frac{\gamma_k}{\mathcal{S}_k}, \gamma_k'\left(\sum_{j=1}^{r}\left[\max\left\{-\frac{y_k^j}{c_k}, g_j(x_k)\right\}\right]^2\right)^{1/2}\right\}.$$

Thus there is no difference in treating equality or inequality constraints, at least within the second order sufficiency assumption framework of this paper.

6. Conclusions. In this paper we provided an analysis of multiplier methods by imbedding them within a general class of penalty function methods. The viewpoint adopted yields strong global convergence results. Furthermore, it provides a fair basis for comparison of multiplier methods with pure penalty function methods. This comparison conclusively demonstrates the superiority of multiplier over penalty methods. The global duality theory obtained has many similarities with the duality theory associated with multiplier methods for convex programming. In addition, it provides a framework within which multiplier methods can be viewed as primal-dual methods in a global sense.

Notes added in proof. The results of this paper have been presented at the 1973 IEEE Decision and Control Conference, San Diego, Calif., Dec. 1973 and at the SIGMAP Symposium on Nonlinear Programming, Madison, Wis., April 1974. They were reported without proofs in [4] and in *Nonlinear Programming 2*, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York, 1975, pp. 165–191.

While this paper was under review, results similar as those of Propositions 1 and 2 appeared in B. T. Polyak and N. V. Tret'yakov, *The Method of Penalty Estimates for Conditional Extremum Problems*, U.S.S.R. Comput. Math. and Mathematical Phys., 13(1974), pp. 42–58.

Generalized versions of Propositions 1 and 2, involving augmented Lagrangians with nonquadratic penalty functions and adjusting essentially the same proof as the one given here, are provided in D. P. Bertsekas, *Multiplier Methods: A Survey*, Preprints of IFAC 6th Triennial World Congress, Part IB, Boston, Mass., Aug. 1975.

REFERENCES

[1] 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 University Press, Stanford, Calif., 1958.

[2] D. P. BERTSEKAS, *Combined primal-dual and penalty methods for constrained minimization*, this Journal, 13 (1975), pp. 521–544.

[3] ———, *On the method of multipliers for convex programming*, EES Dept. Working Paper, Stanford Univ., 1973; IEEE Trans. Automatic Control, AC-20 (1975), pp. 385–388.

[4] ———, *Convergence rate of penalty and multiplier methods*, Proc. 1973 IEEE Conf. on Decision and Control, San Diego, Calif., Dec. 1973, pp. 260–264.

[5] J. D. BUYS, *Dual algorithms for constrained optimization*, Ph.D. thesis, Rijksuniversiteit de Leiden, 1972.

[6] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[7] 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.

[8] M. R. HESTENES, *Multiplier and gradient methods*, J. Optimization Theory Appl., 4 (1969), pp. 303–320.

[9] B. W. KORT AND D. P. BERTSEKAS, *A new penalty function method for constrained minimization*, Proc. 1972 IEEE Conf. on Decision and Control, New Orleans, Dec. 1972.

[10] ———, *Combined primal–dual and penalty methods for convex programming*, EES Dept. Working Paper, Stanford Univ., 1973; this Journal, 14 (1976), pp. 268–294.

[11] ———, *Multiplier methods for convex programming*, Proc. 1973 IEEE Conf. on Decision and Control, San Diego, Calif., Dec. 1973, pp. 428–432.

[12] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass., 1973.

[13] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York, 1971.

[14] B. T. POLYAK, *The convergence rate of the penalty function method*, Ž. Vyčisl. Mat. i Mat. Fiz., 11 (1971), pp. 3–11.

[15] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, New York, pp. 283–298.

[16] R. T. ROCKAFELLAR, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Programming, 5 (1973), pp. 354–373.

[17] ———, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optimization Theory Appl., 12 (1973), pp. 555–562.

[18] ———, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.

[19] ———, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[20] ———, *Penalty methods and augmented Lagrangians in nonlinear programming*, Proc. 5th IFIP Conf. on Optimization Techniques, Rome, 1973; Springer-Verlag, 1974.

[21] ———, *Solving a nonlinear programming problem by way of a dual problem*, Symposia Matematica, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

CANONICAL FORMS OF LINEAR MULTIVARIABLE SYSTEMS*

S. H. WANG† AND E. J. DAVISON‡

Abstract. This paper considers the problem of finding a complete set of invariants and canonical forms for a linear, time-invariant, multivariable system (A, B, C) under a group of transformations. The group consists of input coordinate transformations, state coordinate transformations and state feedback transformations. A set of canonical forms is derived. A set of invariants called the transmission zeros is given, which has application to the servomechanism problem.

Introduction. The problem of finding invariants and canonical forms of linear, time-invariant, multivariable systems under different groups of transformations has been an interesting topic in recent years. In [1]–[3], some invariants for feedback transformations have been utilized in solving different control problems. For a controllable pair (A, B), Brunovský [4] and other authors [5], [6] have found a complete set of feedback invariants and a set of canonical forms. Rosenbrock [6] and Kalman [7] have related feedback invariants to the Kronecker invariants associated with a singular pencil of matrix. Morse [8] has investigated the invariants of a system (A, B, C) under a large group of transformations, which includes output-coordinate transformations and output injection transformations. Popov [9] has recently found a complete set of invariants and canonical forms for a controllable pair (A, B) under state coordinate transformations with or without feedback transformations.

In this paper, we derive a complete set of invariants and a set of canonical forms of a system (A, B, C) under input coordinate transformations, state coordinate transformations and state feedback transformations. We also derive a set of feedback invariants, which is not complete, called the transmission zeros of a linear controllable system (A, B, C). This set of transmission zeros plays an important role in linear multivariable system theory; as an application, we have shown that the necessary and sufficient conditions for a linear control system to exist to regulate a linear multivariable system with arbitrary unmeasurable disturbances present can be stated in terms of the transmission zeros of the system.

In the sequel, the following notation is used. Let k be a positive integer. Then $\mathbf{k}$ denotes $\mathbf{k} \equiv \{1, \cdots, k\}$. Let $\mathscr{V}$ and $\mathscr{W}$ be linear subspaces in R^n. Then $\mathscr{V} + \mathscr{W} = \{v + w | v \in \mathscr{V}, w \in \mathscr{W}\}$.

1. Definitions and problem statements.

Consider the following definitions.

DEFINITION 1.1 (MacLane and Birkhoff [10]). Let X be a set, and let E be an equivalence relation on X. If Γ is another set, a function $f: X \to \Gamma$ is said to be an

* Received by the editors May 11, 1972, and in revised form November 25, 1974. This research was supported by the National Research Council of Canada under Grant A4396.

† Department of Electrical Engineering, University of Toronto, Toronto, Canada. Now at the College of Engineering and Applied Science, University of Colorado, Colorado Springs, Colorado 80907.

‡ Department of Electrical Engineering, University of Toronto, Toronto, Canada.

invariant for E if $xEy \Rightarrow f(x) = f(y)$; it is said to be a *complete invariant* for E if $xEy \Leftrightarrow f(x) = f(y)$.

DEFINITION 1.2. Let X be a set, and let E be an equivalence relation on X. A map $\phi: X \to X$ is said to be a *canonical map* for E on X if

(i) $xE\phi(x), \forall x \in X$,

(ii) $xEy \Leftrightarrow \phi(x) = \phi(y), \forall x, y \in X$.

The image of ϕ, denoted by Im ϕ, is said to be a set of *canonical forms* for E on X. It is also said to be a set of *E-canonical forms* on X.

Remark 1.1. The above definition of canonical forms is clearly equivalent to the definition given by MacLane and Birkhoff [10].

The following lemma states that the set of fixed points of ϕ coincides with the set of canonical forms given by ϕ. Its proof is omitted.

LEMMA 1.1. *Let E be an equivalence relation on a set X, and let ϕ be a canonical map for E on X. Then*

$$x \in \text{Im } \phi \Leftrightarrow \phi(x) = x.$$

Let $X = \{(A, B, C) | A \in R^{n \times n}, B \in R^{n \times m}$ and $C \in R^{p \times n}\}$, and let

$\mathfrak{S} = \{(T, F, G) | T \in R^{n \times n}$ and $G \in R^{m \times m}$ are nonsingular matrices, and $F \in R^{m \times n}\}$.

We define an equivalence relation on X as follows.

DEFINITION 1.3. Two triples $(A_1, B_1, C_1), (A_2, B_2, C_2) \in X$ are said to be $\mathfrak{S}$-*equivalent* if and only if there exists a $(T, F, G) \in \mathfrak{S}$ such that

$$A_1 = T(A_2 + B_2F)T^{-1}, \tag{1.1}$$

$$B_1 = TB_2G, \tag{1.2}$$

$$C_1 = C_2T^{-1}. \tag{1.3}$$

With respect to different groups of transformations, say,

$$\bar{\mathfrak{S}} = \{(T, 0, I_m) | T \in R^{n \times n} \text{ is a nonsingular matrix, 0 is an } m \times n \text{ zero matrix and } I_m \text{ is an } m \times m \text{ identity matrix}\}, \tag{1.4}$$

$$\mathfrak{S}^* = \{(T, 0, G) | T \in R^{n \times n} \text{ and } G \in R^{m \times m} \text{ are nonsingular matrices, and 0 is an } m \times n \text{ zero matrix}\}. \tag{1.5}$$

One can define equivalence relations $\bar{\mathfrak{S}}$-*equivalence* and $\mathfrak{S}^*$-*equivalence* on X in a similar manner.

In the next section, we will derive canonical forms for $\mathfrak{S}$-equivalence.

2. Canonical forms for $\mathfrak{S}$-equivalence. In the sequel, we will impose some conditions on the triples (A, B, C) and consider the set

$X_1 = \{(A, B, C) | A \in R^{n \times n}, B \in R^{n \times m}, C \in R^{p \times n}, (A, B)$ is a controllable pair and rank $B = m\}$.

We first state the following lemma.

LEMMA 2.1 (Brunovský [4] and other authors [5], [6]. *Consider the set* $X_0 = \{(A, B) | A \in R^{n \times n}, B \in R^{n \times m}, (A, B)$ *is a controllable pair and* rank $B = m\}$. *Two*

pairs $(A_1, B_1), (A_2, B_2) \in X_0$ *are said to be* $\mathfrak{S}$*-equivalent if and only if there exists a* $(T, F, G) \in \mathfrak{S}$, *where* $\mathfrak{S}$ *is defined in* § 1, *such that*

$$A_1 = T(A_2 + B_2F)T^{-1}, \tag{2.1}$$

$$B_1 = TB_2G. \tag{2.2}$$

Any pair $(A, B) \in X_0$ *is* $\mathfrak{S}$*-equivalent to a pair* $(A_c, B_c) \in X_0$ *shown below:*

$$\begin{aligned} A_c &= \text{block diag}\,[A_1, \cdots, A_m], \\ B_c &= \text{block diag}\,[b_1, \cdots, b_m], \end{aligned} \tag{2.3}$$

where A_i, b_i *are of size* $n_i \times n_i$, $n_i \times 1$ *respectively, and*

$$A_i = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 1 \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 \end{bmatrix}, \qquad b_i = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix}.$$

The pair (A_c, B_c) *is called the Brunovský canonical form. The set of integers,* $n_1, n_2, \cdots, n_m$, *which satisfies*

$$n_1 \geqq n_2 \geqq \cdots \geqq n_m \geqq 1; \qquad n_1 + n_2 + \cdots + n_m = n,$$

is called the set of controllability indices of the controllable pair (A, B), *and is uniquely determined by* (A, B).

For a proof of this lemma and a detailed discussion on this subject, see [4]–[7].

In the construction of canonical forms for $\mathfrak{S}$-equivalence, the following preliminary results are required.

DEFINITION 2.1. For a given set of integers $n_1, n_2, \cdots, n_m$ satisfying $n_1 \geqq n_2 \geqq \cdots \geqq n_m \geqq 1$; $n_1 + n_2 + \cdots + n_m = n$, define a set of $n \times n$ real matrices T as follows:

$$T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1m} \\ T_{21} & T_{22} & \cdots & T_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ T_{m1} & T_{m2} & \cdots & T_{mm} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n_2 \\ \\ \}n_m \end{matrix} \tag{2.4}$$

$$\underbrace{\phantom{T_{11}}}_{n_1} \quad \underbrace{\phantom{T_{12}}}_{n_2} \qquad \underbrace{\phantom{T_{1m}}}_{n_m}$$

where T_{ij} is an $n_i \times n_j$ real matrix of the following form:

$$T_{ij} = \begin{bmatrix} t_{ij}^1 & t_{ij}^2 & \cdot & \cdot & t_{ij}^{(n_j - n_i + 1)} & & 0 \\ & t_{ij}^1 & t_{ij}^2 & \cdot & \cdot & t_{ij}^{(n_j - n_i + 1)} & \\ & & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & & t_{ij}^1 & t_{ij}^2 & \cdot & & t_{ij}^{(n_j - n_i + 1)} \end{bmatrix} \quad \text{for } n_j \geqq n_i,$$

and

$$T_{ij} = 0 \quad \text{for } n_j < n_i.$$

For convenience, any matrix T of the form (2.4) is said to be a *block stripe matrix.*

LEMMA 2.2. *For a given set of integers* $n_1, n_2, \cdots, n_m$ *satisfying* $n_1 \geqq n_2 \geqq \cdots \geqq n_m \geqq 1$ *and* $\sum_{i=1}^{m} n_i = n$, *the set of nonsingular block stripe matrices* T *in* (2.4) *forms a group under the usual matrix muitiplications. Any block stripe matrix* T *in* (2.4) *is nonsingular if and only if the* $m \times m$ *real matrix*

$$V = [t_{ij}^{(n_j - n_i + 1)}]$$

is nonsingular, where $t_{ij}^{(n_j - n_i + 1)}$ *is the* (i,j)*-th element of* V.

Proof. By straightforward matrix calculation, (though tedious), one can prove this lemma. The details are omitted.

The relationship between the Brunovský canonical form and the group of block stripe matrices is stated in the following proposition.

PROPOSITION 2.1. *For a given set of controllability indices* $n_1, \cdots, n_m$, *let* $\mathscr{T}_{n_i}$ *be the corresponding group of nonsingular block stripe matrices, and let* (A_c, B_c) *be a pair of matrices in the Brunovský canonical form specified by* $n_1, \cdots, n_m$. *Then*

$$\mathscr{T}_{n_i} = \left\{ T \,\middle|\, \begin{array}{l} 1.\ T \text{ is an } n \times n \text{ real nonsingular matrix} \\ 2.\ A_c = T(A_c + B_cF)T^{-1} \text{ for some } m \times n \text{ real matrix } F \\ 3.\ B_c = TB_cG \text{ for some } m \times m \text{ real nonsingular matrix } G \end{array} \right\}.$$

Proof. Let $\hat{T}$ be an element of $\mathscr{T}_{n_i}$. We want to show that there exist an $m \times n$ matrix $\hat{F}$ and an $m \times m$ nonsingular matrix $\hat{G}$ such that

$$A_c = \hat{T}(A_c + B_c\hat{F})\hat{T}^{-1} \tag{2.5}$$

and

$$B_c = \hat{T}B_c\hat{G}. \tag{2.6}$$

From the elements of $\hat{T}$, (see (2.4)), define an $m \times m$ matrix

$$V = [t_{ij}^{(n_j - n_i + 1)}],$$

where $t_{ij}^{(n_j - n_i + 1)}$ is the (i,j)th element of V. From Lemma 2.2 and the assumption that $\hat{T}$ is nonsingular, V is clearly a nonsingular matrix. Let $\hat{G} \triangleq V^{-1}$. It can be easily verified that the pair of matrices $(\hat{T}, \hat{G})$ satisfies (2.6). Next consider (2.5). From (2.6) one can rewrite (2.5) as follows:

$$\begin{aligned} \hat{T}A_c - A_c\hat{T} &= -\hat{T}B_c\hat{F} \\ &= B_c\tilde{F}, \end{aligned} \tag{2.7}$$

where $\tilde{F} \triangleq -\hat{G}^{-1}\hat{F}$. It is easy to show that there exists a unique $m \times n$ matrix

$\tilde{F}$ satisfying (2.7) which is as follows:

$$\tilde{F} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1m} \\ f_{21} & f_{22} & \cdots & f_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ f_{m1} & f_{m2} & \cdots & f_{mm} \end{bmatrix},$$

where f_{ij} is a $1 \times n_j$ row vector defined by

$$f_{ij} = [\underbrace{0 \cdots 0}_{n_i}, \underbrace{t_{ij}^1, t_{ij}^2, \cdots, t_{ij}^{(n_j - n_i)}}_{(n_j - n_i)}].$$

Hence $\hat{F} \triangleq -\hat{G}\tilde{F}$ satisfies (2.5).

It remains to show that any real nonsingular matrix $\hat{T}$ satisfying (2.5) and (2.6) is a block stripe matrix. It is again easier to use (2.6) and (2.7) to show this. The details are omitted. Q.E.D.

In the construction of $\mathfrak{S}$-canonical form, we first transform the triple $(A, B, C) \in X_1$ into a new triple $(A_c, B_c, \tilde{C}) \in X_1$ by applying appropriate transformation $(T, F, G) \in \mathfrak{S}$ constructed in Lemma 2.1, where A_c and B_c are specified in (2.3) and $\tilde{C} = CT^{-1}$. Note that for a given triple $(A, B, C) \in X_1$, the set of matrices $(T, F, G) \in \mathfrak{S}$ constructed in Lemma 2.1 is not unique in general. Therefore, the procedure of transforming a given triple $(A, B, C) \in X_1$ into a new triple $(A_c, B_c, \tilde{C}) \in X_1$ is not a function in the usual sense. In other words, the matrix $\tilde{C}$ is not uniquely determined by (A, B, C). However, from Proposition 2.1, the following results can be immediately obtained.

PROPOSITION 2.2. *Consider two triples* (A_1, B_1, C_1), $(A_2, B_2, C_2) \in X_1$. *Assume that these two triples are transformed into* $(A_{c1}, B_{c1}, \tilde{C}_1), (A_{c2}, B_{c2}, \tilde{C}_2) \in X_1$ *according to Lemma* 2.1. *Then* (A_1, B_1, C_1) *and* (A_2, B_2, C_2) *are* $\mathfrak{S}$*-equivalent if and only if*

(i) $A_{c1} = A_{c2}$ *and* $B_{c1} = B_{c2}$, *i.e., the two sets of controllability indices of* (A_1, B_1) *and* (A_2, B_2) *are equal, which are denoted by* $n_1, \cdots, n_m$;

(ii) $\tilde{C}_1 T = \tilde{C}_2$, *for some nonsingular block stripe matrix* T *in* $\mathscr{T}_{n_i}$.

From Proposition 2.2, we can see that the problem of finding $\mathfrak{S}$-canonical forms on X_1 can be performed in two steps.

Step 1. Find an appropriate set of matrices $(T, F, G) \in \mathfrak{S}$ which transforms a given triple $(A, B, C) \in X_1$ into a new triple $(A_c, B_c, \tilde{C}) \in X_1$, where the pair (A_c, B_c) is in the Brunovský canonical form.

Step 2. Find a nonsingular block stripe matrix $\tilde{T}$ in $\mathscr{T}_{n_i}$, which transforms $\tilde{C}$ into its canonical form $\hat{C}$. More precisely, in Step 2, we want to solve the following problem.

Canonical form for block stripe equivalence.

DEFINITION 2.2. Let $X_2 = \{C | C \in R^{p \times n}\}$, and let n_j, $(j = 1, 2, \cdots, m)$, be a given set of positive integers satisfying

$$n_1 \geqq n_2 \geqq \cdots \geqq n_m \geqq 1 \quad \text{and} \quad \sum_{j=1}^{m} n_j = n.$$

Two matrices C_1 and $C_2 \in X_2$ are said to be *block stripe equivalent* if and only if there exists a nonsingular block stripe matrix T in $\mathscr{T}_{n_i}$, such that $C_1 = C_2 T$.

The problem now becomes: find a canonical map $\psi: X_2 \to X_2$ with the following two properties:

(i) $\psi(C)$ and C are block stripe equivalent, $\forall C \in X_2$,

(ii) $\psi(C_1) = \psi(C_2) \Leftrightarrow C_1$ and C_2 are block stripe equivalent, $\forall C_1, C_2 \in X_2$.

Then from the definition of canonical form, $\operatorname{Im} \psi$ is a set of canonical forms for block stripe equivalence on X_2.

Before we solve the above problem, we first consider a simple example.

Example 2.1. Let $n = 7$, $m = 3$, $n_1 = 3$, $n_2 = 2$ and $n_3 = 2$. Assume that two matrices C and D are block stripe equivalent and they are partitioned as follows:

$$\left[\underbrace{C_{11}}_{3} \mid \underbrace{C_{21}}_{2} \mid \underbrace{C_{22}}_{2}\right] \left[\begin{array}{ccc|ccc|cc} t_{11}^1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & t_{11}^1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & t_{11}^1 & 0 & 0 & 0 & 0 \\ \hline t_{21}^1 & t_{21}^2 & 0 & t_{22}^1 & 0 & t_{23}^1 & 0 \\ 0 & t_{21}^1 & t_{21}^2 & 0 & t_{22}^1 & 0 & t_{23}^1 \\ \hline t_{31}^1 & t_{31}^2 & 0 & t_{32}^1 & 0 & t_{33}^1 & 0 \\ 0 & t_{31}^1 & t_{31}^2 & 0 & t_{32}^1 & 0 & t_{33}^1 \end{array}\right] = \left[\underbrace{D_{11}}_{3} \mid \underbrace{D_{21}}_{2} \mid \underbrace{D_{22}}_{2}\right] \tag{2.8}$$

or, equivalently,

$$\begin{aligned} D_{11} = {} & t_{11}^1 C_{11} + t_{21}^1 [C_{21} \vdots 0] + t_{21}^2 [0 \vdots C_{21}] \\ & + t_{31}^1 [C_{22} \vdots 0] + t_{31}^2 [0 \vdots C_{22}], \end{aligned} \tag{2.9}$$

$$D_{21} = t_{22}^1 C_{21} + t_{32}^1 C_{22}, \tag{2.10}$$

$$D_{22} = t_{23}^1 C_{21} + t_{33}^1 C_{22}. \tag{2.11}$$

Writing $D_{11} = [d_{11}^1 \vdots d_{11}^2 \vdots d_{11}^3]$, $D_{21} = [d_{21}^1 \vdots d_{21}^2]$, $D_{22} = [d_{22}^1 \vdots d_{22}^2]$, and similarly for C_{ij}'s, then from (2.10), (2.11), we have

$$\left[\begin{array}{c|c} d_{21}^1 & d_{22}^1 \\ d_{21}^2 & d_{22}^2 \end{array}\right] = \left[\begin{array}{c|c} c_{21}^1 & c_{22}^1 \\ c_{21}^2 & c_{22}^2 \end{array}\right] \left[\begin{array}{c|c} t_{22}^1 & t_{23}^1 \\ t_{32}^1 & t_{33}^1 \end{array}\right]. \tag{2.12}$$

Similarly from (2.9), we have

$$\begin{bmatrix} d_{11}^1 \\ d_{11}^2 \\ d_{11}^3 \end{bmatrix} = t_{11}^1 \begin{bmatrix} c_{11}^1 \\ c_{11}^2 \\ c_{11}^3 \end{bmatrix} + t_{21}^1 \begin{bmatrix} c_{21}^1 \\ c_{21}^2 \\ 0 \end{bmatrix} + t_{21}^2 \begin{bmatrix} 0 \\ c_{21}^1 \\ c_{21}^2 \end{bmatrix} + t_{31}^1 \begin{bmatrix} c_{22}^1 \\ c_{22}^2 \\ 0 \end{bmatrix} + t_{31}^2 \begin{bmatrix} 0 \\ c_{22}^1 \\ c_{22}^2 \end{bmatrix}. \tag{2.13}$$

Let

$$\mathscr{C}_1 = \operatorname{span}\begin{bmatrix} c_{11}^1 \\ c_{11}^2 \\ c_{11}^3 \end{bmatrix}, \quad \mathscr{C}_2 = \operatorname{span}\begin{bmatrix} c_{21}^1 & c_{22}^1 \\ & \\ c_{21}^2 & c_{22}^2 \end{bmatrix}.$$

and

$$\mathscr{C}_2^1 = \operatorname{span}\begin{bmatrix} c_{21}^1 & c_{22}^1 & 0 & 0 \\ c_{21}^2 & c_{22}^2 & c_{21}^1 & c_{22}^1 \\ 0 & 0 & c_{21}^2 & c_{22}^2 \end{bmatrix}.$$

Similar subspaces are constructed from matrix D. Then from (2.12) and (2.13), the following relations hold:

$$\mathscr{D}_2 = \mathscr{C}_2 \tag{2.14}$$

and

$$\mathscr{D}_1 \subset \mathscr{C}_1 + \mathscr{C}_2^1. \tag{2.15}$$

From (2.14) and the definitions of $\mathscr{C}_2^1, \mathscr{D}_2^1$, there follows

$$\mathscr{D}_2^1 = \mathscr{C}_2^1. \tag{2.16}$$

Since block stripe equivalence is symmetric, if we exchange the role of matrix C and matrix D, we have

$$\mathscr{C}_1 \subset \mathscr{D}_1 + \mathscr{D}_2^1. \tag{2.17}$$

From (2.15)–(2.17), it is easy to see that

$$\mathscr{D}_1 + \mathscr{D}_2^1 = \mathscr{C}_1 + \mathscr{C}_2^1. \tag{2.18}$$

The above derivation shows that, if two matrices C and D in (2.8) are block stripe equivalent, then (2.14) and (2.18) hold. In fact, one can show that the converse statement is also true.

In order to deal with the general case, we need the following notations.

Consider a given matrix $C \in X_2$ and a given set of integers $n_1, n_2, \cdots, n_m$ satisfying

$$n_1 \geqq n_2 \geqq \cdots \geqq n_m \geqq 1,$$

$$n_1 + n_2 + \cdots + n_m = n,$$

and

$$n_1 = \cdots = n_{q_1} = \gamma_1,$$

$$n_{q_1+1} = \cdots = n_{q_1+q_2} = \gamma_2,$$

$$n_{m-q_l+1} = \cdots = n_m = \gamma_l,$$

where

$$\gamma_1 > \gamma_2 > \cdots > \gamma_l \geqq 1.$$

In other words, γ_1 is the largest number among n_i's, and γ_2 is the second largest number among n_i's, etc.

Any matrix $C \in X_2$ can be partitioned as follows:

$$C = [\underbrace{C_{11}}_{\gamma_1} \vdots \cdots \vdots \underbrace{C_{1q_1}}_{\gamma_1} \vdots \underbrace{C_{21}}_{\gamma_2} \vdots \cdots \vdots \underbrace{C_{2q_2}}_{\gamma_2} \vdots \underbrace{C_{31}}_{\gamma_3} \vdots \cdots \vdots \underbrace{C_{lq_l}}_{\gamma_l}]\}p,$$

where each submatrix C_{ij} can be written as

$$C_{ij} = [c_{ij}^1 \vdots c_{ij}^2 \vdots \cdots \vdots c_{ij}^{\gamma_i}], \qquad i \varepsilon \mathbf{l}, \quad j \in \mathbf{q}_i.$$

From C_{ij}, one can define a $p \times \gamma_i$ column vector consisting of the columns of C_{ij}, i.e.,

$$c_{ij} = \begin{bmatrix} c_{ij}^1 \\ \hline c_{ij}^2 \\ \hline \\ \vdots \\ \\ \hline c_{ij}^{\gamma_i} \end{bmatrix} \begin{matrix} \}p \\ \}p \\ \\ \\ \\ \\ \}p \end{matrix}, \qquad i \in \mathbf{l}, \quad j \in \mathbf{q}_i.$$

Then we can define a subspace $\mathscr{C}_i \subset R^{p \times \gamma_i}$, $i \in \mathbf{l}$, as

$$\mathscr{C}_i = \text{span}\,\{c_{ij}, j \in \mathbf{q}_i\}, \tag{2.19}$$

and a subspace $\mathscr{C}_i^k$ in $R^{p \times \gamma_k}$, $i = 2, \cdots, l$; $k \in \mathbf{i} - \mathbf{1}$, as

$$\mathscr{C}_i^k = \{x | x = [\underbrace{0, \cdots, 0}_{p \times \mu} \vdots \underbrace{c^t}_{p \times \gamma_i} \vdots \underbrace{0, \cdots, 0}_{p \times (\gamma_k - \gamma_i - \mu)}]^t, c \in \mathscr{C}_i, \mu = 0, 1, \cdots, \gamma_k - \gamma_i\}, \tag{2.20}$$

where $\mathscr{C}_i^k$ is constructed by adding zeros to the tops and bottoms of the vectors in $\mathscr{C}_i$ in order to make its length equal to $p \times \gamma_k$.

With the above notations, we are ready to state the following lemma motivated by Example 2.1.

LEMMA 2.3. *If two matrices C and D are block stripe equivalent, then using the above notations, we have*

$$\begin{aligned} &\mathscr{C}_l = \mathscr{D}_l, \\ &\mathscr{C}_{l-1} + \mathscr{C}_l^{l-1} = \mathscr{D}_{l-1} + \mathscr{D}_l^{l-1}, \\ &\qquad \vdots \\ &\mathscr{C}_1 + \mathscr{C}_2^1 + \cdots + \mathscr{C}_l^1 = \mathscr{D}_1 + \mathscr{D}_2^1 + \cdots \mathscr{D}_l^1. \end{aligned}$$

The proof of Lemma 2.3 follows closely the derivations in Example 2.1 and is omitted.

Continuing our effort in the construction of canonical form for block stripe equivalence, we return to Example 2.1.

From equations (2.14), (2.18), we construct two subspaces

$$\hat{\mathscr{C}}_2 = \mathscr{C}_2, \tag{2.21}$$

$$\hat{\mathscr{C}}_1 = (\mathscr{C}_1 + \mathscr{C}_2^1) \cap (\mathscr{C}_2^1)^\perp, \tag{2.22}$$

where $(\mathscr{C}_2^1)^\perp \subset R^{p\times 3}$ denotes the orthogonal complement of $\mathscr{C}_2^1 \subset R^{p\times 3}$. Subspaces $\hat{\mathscr{D}}_1$ and $\hat{\mathscr{D}}_2$ are constructed in a similar way. Clearly $\hat{\mathscr{C}}_i = \hat{\mathscr{D}}_i$, $i = 1, 2$. From (2.21), (2.22), we construct two matrices $\hat{C}_1$ and $\hat{C}_2$ as follows:

$$\hat{C}_1 = \begin{bmatrix} \hat{c}_{11}^1 \\ \hat{c}_{11}^2 \\ \hat{c}_{11}^3 \end{bmatrix}, \qquad \hat{C}_2 = \left[\begin{array}{c|c} \hat{c}_{21}^1 & \hat{c}_{22}^1 \\ \hat{c}_{21}^2 & \hat{c}_{22}^2 \end{array}\right] \tag{2.23}$$

such that $\hat{\mathscr{C}}_i = \operatorname{span} \hat{C}_i$, $i = 1, 2$, and both $\hat{C}_1$ and $\hat{C}_2$ are in the *column-reduced echelon* form (see Loomis and Sternberg [11, p. 104]).

From (2.23), we construct $\hat{C}$ as follows:

$$\hat{C} = [\underbrace{\hat{c}_{11}^1 \,\vdots\, \hat{c}_{11}^2 \,\vdots\, \hat{c}_{11}^3}_{3} \,\vdots\, \underbrace{\hat{c}_{21}^1 \,\vdots\, \hat{c}_{21}^2}_{2} \,\vdots\, \underbrace{\hat{c}_{22}^1 \,\vdots\, \hat{c}_{22}^2}_{2}]. \tag{2.24}$$

We can construct $\hat{D}$ from subspaces $\hat{\mathscr{D}}_1$ and $\hat{\mathscr{D}}_2$ in a similar fashion. From the property of the column-reduced echelon form and the fact that $\hat{\mathscr{C}}_i = \hat{\mathscr{D}}_i$, $i = 1, 2$, we can see that $\hat{C} = \hat{D}$.

In order to show that $\hat{C}$ is block stripe equivalent to the original matrix C in (2.8), we need the following lemma.

LEMMA 2.4. *Let* $\mathscr{A} = \operatorname{span}\{a_1, \cdots, a_k\}$ *and* $\mathscr{B} = \operatorname{span}\{b_1, \cdots, b_l\}$, *where* a_i, $b_j \in R^n$, $i \in \mathbf{k}$, $j \in \mathbf{l}$. *Write each vector* a_i *as*

$$a_i = \bar{a}_i + \tilde{a}_i, \qquad i \in \mathbf{k},$$

where $\tilde{a}_i \in \mathscr{B}^\perp$ *and* $\bar{a}_i \in \mathscr{B}$. *Define a subspace* $\hat{\mathscr{A}}$ *as follows*:

$$\hat{\mathscr{A}} = (\mathscr{A} + \mathscr{B}) \cap \mathscr{B}^\perp.$$

Then

$$\hat{\mathscr{A}} = \operatorname{span}\{\tilde{a}_1, \cdots, \tilde{a}_k\}.$$

Proof. Since $a_i \in \mathscr{A}$ and $\bar{a}_i \in \mathscr{B}$, $\tilde{a}_i \in (\mathscr{A} + \mathscr{B})$. By construction, $\tilde{a}_i \in \mathscr{B}^\perp$, hence $\tilde{a}_i \in \hat{\mathscr{A}}$. This shows that

$$\hat{\mathscr{A}} \supset \operatorname{span}\{\tilde{a}_1, \cdots, \tilde{a}_k\}.$$

Conversely, any vector $x \in \hat{\mathscr{A}}$ can be written as

$$x = a + b = \left(\sum_{i=1}^{k} \alpha_i a_i\right) + b = \left(\sum_{i=1}^{k} \alpha_i \bar{a}_i\right) + \left(\sum_{i=1}^{k} \alpha_i \tilde{a}_i\right) + b,$$

where $a \in \mathscr{A}, b \in \mathscr{B}$. Since $\bar{a}_i, b \in \mathscr{B}$ and $\tilde{a}_i, x \in \mathscr{B}^{\perp}$, there follows

$$\left(\sum_{i=1}^{k} \alpha_i \bar{a}_i \right) + b = 0,$$

and

$$x = \sum_{i=1}^{k} \alpha_i \tilde{a}_i.$$

This proves that $\hat{\mathscr{A}} \subset \text{span}\{\tilde{a}_1, \cdots, \tilde{a}_k\}$. Q.E.D.

From (2.22), (2.23) and the above lemma, we can write

$$\begin{bmatrix} c_{11}^1 \\ c_{11}^2 \\ c_{11}^3 \end{bmatrix} = \begin{bmatrix} \bar{c}_{11}^1 \\ \bar{c}_{11}^2 \\ \bar{c}_{11}^3 \end{bmatrix} + \begin{bmatrix} \tilde{c}_{11}^1 \\ \tilde{c}_{11}^2 \\ \tilde{c}_{11}^3 \end{bmatrix}, \tag{2.25}$$

where

$$\begin{bmatrix} \bar{c}_{11}^1 \\ \bar{c}_{11}^2 \\ \bar{c}_{11}^3 \end{bmatrix} \in \mathscr{C}_2^1 \quad \text{and} \quad \begin{bmatrix} \tilde{c}_{11}^1 \\ \tilde{c}_{11}^2 \\ \tilde{c}_{11}^3 \end{bmatrix} \in (\mathscr{C}_2^1)^{\perp}. \tag{2.26}$$

From (2.22), (2.26) and Lemma 2.4, we have

$$\hat{\mathscr{C}}_1 = \text{span} \begin{bmatrix} \tilde{c}_{11}^1 \\ \tilde{c}_{11}^2 \\ \tilde{c}_{11}^3 \end{bmatrix}. \tag{2.27}$$

Hence from (2.23), there exists $t_{11}^1 \neq 0$ such that

$$t_{11}^1 \cdot \begin{bmatrix} \tilde{c}_{11}^1 \\ \tilde{c}_{11}^2 \\ \tilde{c}_{11}^3 \end{bmatrix} = \begin{bmatrix} \hat{c}_{11}^1 \\ \hat{c}_{11}^2 \\ \hat{c}_{11}^3 \end{bmatrix}. \tag{2.28}$$

From (2.26) and the definition of $\mathscr{C}_2^1$, there exist constants $t_{21}^1, t_{31}^1, t_{21}^2$ and t_{31}^2 such that

$$t_{11}^1 \cdot \begin{bmatrix} \bar{c}_{11}^1 \\ \bar{c}_{11}^2 \\ \bar{c}_{11}^3 \end{bmatrix} = t_{21}^1 \cdot \begin{bmatrix} c_{21}^1 \\ c_{21}^2 \\ 0 \end{bmatrix} + t_{31}^1 \cdot \begin{bmatrix} c_{22}^1 \\ c_{22}^2 \\ 0 \end{bmatrix} + t_{21}^2 \cdot \begin{bmatrix} 0 \\ c_{21}^1 \\ c_{21}^2 \end{bmatrix} + t_{31}^2 \cdot \begin{bmatrix} 0 \\ c_{22}^1 \\ c_{22}^2 \end{bmatrix}. \tag{2.29}$$

From (2.21), (2.23) and the definition of $\mathscr{C}_2$, there exist constants $t_{22}^1, t_{23}^1, t_{32}^1, t_{33}^1$ such that

$$\left[\begin{array}{c|c} c_{21}^1 & c_{22}^1 \\ \hline c_{21}^2 & c_{22}^2 \end{array}\right] \left[\begin{array}{c|c} t_{22}^1 & t_{23}^1 \\ \hline t_{32}^1 & t_{33}^1 \end{array}\right] = \left[\begin{array}{c|c} \hat{c}_{21}^1 & \hat{c}_{22}^1 \\ \hline \hat{c}_{21}^2 & \hat{c}_{22}^2 \end{array}\right]. \tag{2.30}$$

Furthermore, in the above equation, $t_{22}^1 \cdot t_{33}^1 - t_{23}^1 \cdot t_{32}^1 \neq 0$. From (2.25), we

can write the C matrix as follows:

$$
\begin{aligned}
C &= [C_{11} \,\vdots\, C_{21} \,\vdots\, C_{22}] \\
&= [\bar{c}_{11}^1 + \tilde{c}_{11}^1 \,\vdots\, \bar{c}_{11}^2 + \tilde{c}_{11}^2 \,\vdots\, \bar{c}_{11}^3 + \tilde{c}_{11}^3 \,\vdots\, c_{21}^1 \,\vdots\, c_{21}^2 \,\vdots\, c_{22}^1 \,\vdots\, c_{22}^2].
\end{aligned} \tag{2.31}
$$

From (2.25), (2.28) – (2.31), we have

$$
C \cdot \left[\begin{array}{ccc:ccc:cc}
t_{11}^1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & t_{11}^1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & t_{11}^1 & 0 & 0 & 0 & 0 \\
\hdashline
-t_{21}^1 & -t_{21}^2 & 0 & t_{22}^1 & 0 & t_{23}^1 & 0 \\
0 & -t_{21}^1 & -t_{21}^2 & 0 & t_{22}^1 & 0 & t_{23}^1 \\
\hdashline
-t_{31}^1 & -t_{31}^2 & 0 & t_{32}^1 & 0 & t_{33}^1 & 0 \\
0 & -t_{31}^1 & -t_{31}^2 & 0 & t_{32}^1 & 0 & t_{33}^1
\end{array}\right] = \hat{C}.
$$

This shows that $\hat{C}$ is block stripe equivalent to C. Hence the above procedure of constructing $\hat{C}$ is a canonical map and $\hat{C}$ is in a canonical form for block stripe equivalence.

The following algorithm summarizes the procedures of constructing canonical forms for block stripe equivalence.

ALGORITHM 2.1.

Step 1. For any given real matrix $C \in X_2$ and a set of controllability indices $n_1, \cdots, n_m$, construct a set of subspaces $\mathscr{C}_1, i \in \mathbf{l}$, and $\mathscr{C}_i^k, i = 2, \cdots, l; k \in \mathbf{i} - \mathbf{1}$, as in (2.19), (2.20).

Step 2. Construct a set of subspaces as follows:

$$
\begin{aligned}
\hat{\mathscr{C}}_l &= \mathscr{C}_l, \\
\hat{\mathscr{C}}_{l-1} &= (\mathscr{C}_{l-1} + \mathscr{C}_l^{l-1}) \cap (\mathscr{C}_l^{l-1})^{\perp}, \\
&\vdots \\
\hat{\mathscr{C}}_1 &= (\mathscr{C}_1 + \mathscr{C}_2^1 + \cdots + \mathscr{C}_l^1) \cap (\mathscr{C}_2^1 + \cdots + \mathscr{C}_l^1)^{\perp},
\end{aligned}
$$

where $\mathscr{C}^{\perp} \subset R^{\alpha}$ denotes the orthogonal complement of $\mathscr{C} \subset R^{\alpha}$.

Step 3. Construct a $(p \times \gamma_i) \times q_i$ matrix $\hat{C}_i$ in the column-reduced echelon form (see [11, p. 104]),

$$
\hat{C}_i = [\hat{c}_{i1} \,\vdots\, \hat{c}_{i2} \,\vdots\, \cdots \,\vdots\, \hat{c}_{iq_i}],
$$

such that $\hat{\mathscr{C}}_i = \operatorname{span} \hat{C}_i, i \in \mathbf{l}$.

Step 4. Write each of the above column vectors of $\hat{C}_i$ as follows:

$$\hat{c}_{ij} = \begin{bmatrix} \hat{c}_{ij}^1 \\ \hline \hat{c}_{ij}^2 \\ \hline \vdots \\ \hline \hat{c}_{ij}^{\gamma_i} \end{bmatrix} \begin{matrix} \}p \\ \}p \\ \\ \}p \end{matrix}$$

and define a $p \times \gamma_i$ matrix $\hat{C}_{ij}$, $i \in \mathbf{l}$, $j \in \mathbf{q}_i$, as

$$\hat{C}_{ij} = [\hat{c}_{ij}^1 \mid \hat{c}_{ij}^2 \mid \cdots \mid \hat{c}_{ij}^{\gamma_i}].$$

The canonical form $\hat{C}$ of the matrix C is

$$\hat{C} = [\hat{C}_{11} \mid \cdots \mid \hat{C}_{1q_1} \mid \hat{C}_{21} \mid \cdots \mid \hat{C}_{2q_2} \mid \cdots \mid \hat{C}_{lq_l}].$$

END OF THE ALGORITHM.

Comment. From the definition of $\hat{\mathscr{C}}_i$ and Lemma 2.4, one can see that $\dim(\hat{\mathscr{C}}_i) \leqq q_i$. Hence the construction of $\hat{C}_i$ in Step 3 of the above algorithm is always possible.

3. Transmission zeros of linear multivariable systems. In this section, we will define a set of zeros, called *transmission zeros*, for any linear multivariable system which is completely controllable. This set of zeros is an invariant for $\mathfrak{S}$-equivalence, hence it can be defined in terms of the $\mathfrak{S}$-canonical form derived in § 2.

Consider a given triple $(A, B, C) \in X_1$. Using the algorithm in § 2, we first transform (A, B, C) into its canonical form $(A_c, B_c, \hat{C})$, where (A_c, B_c) is in the Brunovský canonical form. Let $n_1, \cdots, n_m$ be the set of controllability indices of (A_c, B_c). We define an $n \times m$ polynomial matrix as follows:

$$M(s) = \text{block diag}\,[m_i(s)],$$

$$m_i(s) = \underbrace{\begin{bmatrix} 1 \\ s \\ \vdots \\ s^{n_i - 1} \end{bmatrix}}_{1} \Big\} n_i, \qquad i \in \mathbf{m}.$$

Then we can calculate the product of $\hat{C}$ and $M(s)$,

$$N(s) \equiv \hat{C}M(s). \tag{3.1}$$

Note that $N(s)$ is a $p \times m$ polynomial matrix which can be considered as the *numerator* of the system (A, B, C) (see [6], [12]–[16]).

Then we transform $N(s)$ into its *Smith canonical form* as follows, where $U_1(s)$ and $U_2(s)$ are polynomial matrices with nonzero constant determinants,

$$
\begin{aligned}
U_1(s)\,N(s)\,U_2(s) &= \Lambda(s) \\
&= [\operatorname{diag}(e_i(s)) \,\vdots\, 0_{p,m-p}] \quad (m > p) \\
&= \operatorname{diag}(e_i(s)) \quad (m = p) \\
&= \begin{bmatrix} \operatorname{diag}(e_i(s)) \\ \hline 0_{p-m,m} \end{bmatrix} \quad (p > m).
\end{aligned} \tag{3.2}
$$

DEFINITION 3.1. The polynomial $\prod_{i=1}^{l} e_i(s)$, $l = \min(p, m)$, derived from the Smith canonical form of $N(s)$, is said to be the *transmission polynomial* of the given system $(A, B, C) \in X_1$. The set of zeros of the transmission polynomial is said to be the set of *transmission zeros* of $(A, B, C) \in X_1$.

The set of transmission zeros defined above has an interesting application in the control of linear multivariable systems with disturbances.

We first state the following lemma.

LEMMA 3.1. *Consider a given triple* $(A, B, C) \in X_1$ *and its numerator* $N(s)$ *defined in* (6.1). *Let* λ *be any complex number. Then the following two conditions are equivalent*:

(i) $$\operatorname{rank} \begin{bmatrix} \lambda I - A & B \\ \hline C & 0 \end{bmatrix} = n + p,$$

(ii) $$\operatorname{rank} N(\lambda) = p.$$

Proof. By transforming (A, B, C) into its $\mathfrak{S}$-canonical form $(A_c, B_c, \hat{C})$, one can easily prove this lemma. The details are omitted.

LEMMA 3.2. *Consider a controllable linear time-invariant system specified by*

$$\dot{x}(t) = Ax(t) + Bu(t) + w(t), \tag{3.3}$$

$$y(t) = Cx(t), \tag{3.4}$$

where $u(t) \in R^m$ *is the input,* $x(t) \in R^n$ *is the state,* $y(t) \in R^p$ *is the output,* $w(t) \in R^n$ *is the disturbance whose components* $w_i(t)$, $i \in \mathbf{n}$, *satisfy a differential equation with unstable characteristic roots* $\lambda_1, \cdots, \lambda_q$. *Then a necessary and sufficient condition that there exists a linear controller (either feedback or feedforward) so that* $y(t) \to 0$ *as* $t \to \infty$ *and such that the controlled system is controllable is that*

$$\operatorname{rank} \begin{bmatrix} \lambda_j I - A & B \\ \hline C & 0 \end{bmatrix} = n + p, \qquad j \in \mathbf{q}. \tag{3.5}$$

The proof of Lemma 3.2 and a detailed discussion on this subject can be found in [17], [18].

PROPOSITION 3.1. *Consider the controllable system specified by* (3.3), (3.4) *and its transmission zeros defined in Definition* 3.1. *Then condition* (3.5) *is equivalent to the following*:

(i) *The set of the transmission zeros of* (A, B, C) *are disjoint from* $\{\lambda_1, \cdots, \lambda_q\}$, *and*

(ii) $m \geqq p$.

Proof. From Lemma 3.1, condition (3.5) is equivalent to the following:

$$\text{rank } N(\lambda_j) = p, \qquad j \in \mathbf{q}. \tag{3.6}$$

In (3.2), since both $U_1(s)$ and $U_2(s)$ are matrices with nonzero constant determinants, the condition (3.6) is equivalent to the condition that $m \geqq p$ and

$$\left(\prod_{i=1}^{l} e_i(\lambda_j) \right) \neq 0, \qquad l = \min(p, m), \quad j \in \mathbf{q},$$

where $\prod_{i=1}^{l} e_i(\lambda)$ is the transmission polynomial of (A, B, C) in (3.3), (3.4). Q.E.D.

Remark 3.1. If a given system (A, B, C) is both completely controllable and completely observable, then one can define the transmission polynomial of (A, B, C) to be the product of the numerator polynomials in the Smith–McMillan form of the transfer function $H(s) = C(sI - A)^{-1}B$ (see [6]). But if a given system (A, B, C) is not completely observable, the Smith–McMillan form only gives transmission zeros corresponding to the controllable and observable part.

Remark 3.2. In the case that $p = m$, Kwakernaak and Sivan [3] give a definition of zeros of a system. Their definition can be shown to be equivalent to Definition 5.1 when $p = m$. Recently, Morse [8] gave a definition of transmission polynomials of a system (A, B, C). With the assumption of complete controllability, one can see that Morse's definition of transmission polynomials is closely related to ours via Lemma 3.1.

4. Conclusions. The problem of finding canonical forms of linear multivariable systems under state coordinate transformations has been investigated by many authors [19]–[24]. It is worthwhile to investigate which of these forms is canonical in the sense of Definition 1.2. In this paper, a precise definition of canonical forms is given, and according to this definition, a set of canonical forms under input coordinate transformations, state coordinate transformations and state feedback transformations ($\mathfrak{S}$-canonical forms) is derived in § 2. Under different groups of transformations, $\bar{\mathfrak{S}}$-canonical forms and $\mathfrak{S}^*$-canonical forms have also been derived in [25].

An interesting set of invariants (which is not complete) for $\mathfrak{S}$-equivalence, called transmission zeros, is derived in § 3. This set of zeros is seen to play an important role in linear multivariable system theory, e.g., in the general servomechanism problem.

Acknowledgment. The authors are grateful to Professor W. M. Wonham for his many stimulating discussions.

REFERENCES

[1] V. M. POPOV, *Hyperstability and optimality of automatic systems with several control functions*, Rev. Roumaine Sci. Tech. Electrotech. et Engrg., 9 (1964), pp. 629–690.

[2] E. G. GILBERT, *The decoupling of multivariable systems by state feedback*, this Journal, 7 (1969), pp. 50–63.

[3] H. KWAKERNAAK AND R. SIVAN, *The maximally achievable accuracy of linear optimal regulators and linear optimal filters*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 79–85.

[4] P. BRUNOVSKÝ, *A classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–187.

[5] W. M. WONHAM AND A. S. MORSE, *Feedback invariants of linear multivariable systems*, Automatica, 8 (1972), pp. 93–100.

[6] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Wiley–Interscience, New York, 1970.

[7] R. E. KALMAN, *Kronecker invariants and feedback*, Proc. Conf. on Ordinary Differential Equations, Math. Research Center, Naval Research Lab., Washington, D.C., 1971.

[8] A. S. MORSE, *Structural invariants of linear multivariable systems*, this Journal, 11 (1973), pp. 446–465.

[9] V. M. POPOV, *Invariants description of linear, time-invariant controllable systems*, this Journal, 10 (1972), pp. 252–264.

[10] S. MACLANE AND G. BIRKHOFF, *Algebra*, Macmillan, New York, 1967.

[11] L. H. LOOMIS AND S. STERNBERG, *Advanced Calculus*, Addison-Wesley, Reading, Mass., 1968.

[12] V. M. POPOV, *Some properties of the control systems with irreducible matrix transfer functions* (Lecture Notes in Mathematics, 144), Seminar on Differential Equations and Dynamical Systems, Springer, New York, 1970, pp. 250–261.

[13] W. A. WOLOVICH, *The application of state feedback invariants to exact model matching*, 5th Annual Princeton Conf. Information Sciences and Systems, Princeton, N.J., 1971.

[14] S. H. WANG, *Design of linear multivariable systems*, Electron. Res. Lab., Univ. of California, Berkeley, Memo. ERL-M309, 1971; also Ph.D. dissertation, Department of Electrical Engineering Comput. Sci., Univ. of California, Berkeley, 1971.

[15] S. H. WANG AND C. A. DESOER, *The exact model matching of linear multivariable systems*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 347–349.

[16] S. H. WANG AND E. J. DAVISON, *A minimization algorithm for the design of linear multivariable systems*, Ibid., AC-18 (1973), pp. 220–225.

[17] E. J. DAVISON, *The ouptut control of linear time-invariant multivariable systems with unmeasurable arbitrary disturbances*, Ibid., AC-17 (1972), pp. 621–630.

[18] ———, *The feedforward control of linear multivariable time-invariant systems*, Automatica, 9 (1973), pp. 561–573.

[19] D. G. LUENBERGER, *Canonical forms for linear multivariable systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 290–293.

[20] W. M. WONHAM AND C. D. JOHNSON, *Optimal bang-bang control with quadratic performance index*, Trans. A.S.M.E., J. Basic Engrg., 86 (1964), pp. 107–115.

[21] R. W. BASS AND I. GURA, *Canonical forms for controllable systems with application to optimal nonlinear feedback*, Proc. 1966 Congress of the Internat. Federation of Automatic Control, London, England.

[22] R. S. BUCY, *Canonical forms for multivariable systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 567–569.

[23] M. HEYMANN, *A unique canonical form for multivariable systems*, Internat. J. Control, 12 (1970), pp. 913–927.

[24] L. M. SILVERMAN, *Transformation of time-variable systems to canonical (phase-variable) form*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 300–303.

[25] S. H. WANG AND E. J. DAVISON, *Canonical forms of linear multivariable systems*, Control System Rep. 7203, Dept. of Electrical Engineering, University of Toronto, Canada, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

A FUNCTIONAL DIFFERENTIAL EQUATION APPROACH TO SOLVING INFINITE GAMES*

R. G. UNDERWOOD†

Abstract. This paper develops a theoretical foundation for the numerical solution of two classes of infinite zero sum games, namely continuous and L^∞ games. Our approach is to introduce a dynamical model (a functional differential equation) for nondynamic L^∞ games and then to show that approximate solutions to a symmetric L^∞ game can be obtained by examining the limiting behavior of the game's dynamical model. By viewing a continuous game as an L^∞ game, it is shown that exact solutions to a symmetric continuous game can be found by examining the limiting behavior of the corresponding L^∞ game dynamical model. Since the dynamical model is nonlinear, a proof of the existence and uniqueness of its solutions is included. Finally a symmetrization is described for continuous and L^∞ games, and thus the theory provides a general method for solving games of these classes.

1.1. Introduction. In this paper, two classes of infinite games, L^∞ and continuous games, will be considered. The terms L^∞ and continuous conventionally refer to the game kernel. Our objective is to develop a dynamic model for these two classes of games and to study its stability. The dynamical model and associated theory provide a theoretical foundation for the numerical computation of solutions to these games.

It is well known that approximate solutions to continuous games can be found by first approximating the continuous kernel with a matrix and then solving the matrix game. However, if the continuous kernel is irregular, the dimension of the matrix game may need to be very large in order to obtain sufficient accuracy. The techniques for solving matrix games (viz., the simplex method and fictitious play) are generally not quick enough for handling games of large size [6]. The method of fictitious play can also be applied directly to continuous games. This method resembles a multistage learning process. At each stage, it is assumed that the players choose a strategy that would yield the optimum result if employed against the empirical distribution of all past choices of their opponents. For instance, consider the game of matching pennies. Assume that player P_r wins if there is a match and P_c wins if there is not a match. Now suppose in the first three plays P_c has chosen "heads" once and "tails" twice, then the "fictitious" play for P_r is to choose "tails".

The advantages of our procedure over the methods using a matrix game approximation is that any continuous game can be solved numerically *without* being removed from its original function space setting. By analyzing the problem from this viewpoint, it is possible to exploit numerical methods which are directly applicable to the function space setting. Whether our procedure compares more favorably to that of fictitious play with respect to the relative rates of convergence etc. is a question for further study. We offer it as an alternative.

Let us now briefly outline our approach. We will first analyze symmetric games. For symmetric continuous games, we will show that a solution to the game can be obtained by examining the limiting behavior of the game's dynamic model. In the case of L^∞ games, the stability is not as decisive. However, we will show

* Received by the editors July 12, 1974, and in revised form March 18, 1975.

† Department of Mathematics and Computer Science, University of South Carolina, Columbia, South Carolina 29208.

that by changing the game's kernel on a set of arbitrarily small measure, we can obtain ε-optimal solutions to the game. Having analyzed the symmetric case, we will then describe a symmetrization for L^∞ and continuous games; and thus the theory will provide a general method for solving games of these classes.

1.2. Solving basic definitions. A two-player zero sum game can be thought of as a triple $\{X, Y, \mathscr{K}\}$, where X and Y are the strategy spaces of the two players respectively and where $\mathscr{K}$ is the payoff functional defined on $X \times Y$. In this paper, $\mathscr{K}$ will always be a bilinear functional which can be identified with a kernel $K(\cdot, \cdot)$. Therefore, in order to define an L^∞ or continuous games, we shall first need to define the game kernel and strategy spaces. Throughout this paper we shall denote the interval $[0, 1]$ by I.

Since the L^∞ kernel will be defined on $I \times I$, we need to specify an appropriate measure space. Let M denote the collection of Lebesgue measurable sets on I and μ Lebesgue measure. We will consider the measure space $\{I \times I, M \times M, \mu \times \mu\}$. (In Lemma 2.3–1, the reason for choosing this measure space will become apparent.) $L^\infty(\mu \times \mu)$ denotes the set of all essentially bounded (with respect to $\mu \times \mu$) functions on $I \times I$.

DEFINITION 1.2–1. A function K is an $n \times m$ L^∞ *kernel* if it is an $n \times m$ matrix of real-valued function K_{ij} defined on $I \times I$, where each $K_{ij} \in L^\infty(\mu \times \mu)$.

DEFINITION 1.2–2. An R^n density function f is a *Lebesgue measurable function* mapping I into R^n, where

$$f_i(x) \geqq 0 \quad \text{a.e.}, \qquad i = 1, \cdots, n, \tag{1.2–1}$$

and

$$\sum_{i=1}^{n} \int_I f_i(x)\, d\mu(x) = 1. \tag{1.2–2}$$

Player 1's strategy space, $\mathscr{F}$, is the set of all R^n density functions f and player 2's strategy space, $\mathscr{G}$, is the set of all R^m density functions g. The triple $(\mathscr{F}, \mathscr{G}, K\}$ is called an L^∞ *game*, where the payoff functional is given by

$$\mathscr{K}[f, g] = \int_I d\mu(x) \int_I f(x) \cdot (K(x, y)g(y))\, d\mu(y).$$

We shall now define a continuous game in an analogous manner.

DEFINITION 1.2–3. A function K is an $n \times m$ *continuous kernel* if it is an $n \times m$ matrix of continuous real-valued function K_{ij} defined on $I \times I$.

DEFINITION 1.2–4. An R^n *policy function* is a function F mapping I into R^n, where each F_i is a nondecreasing function and right continuous on $(0, 1]$,

$$F_i(0) = 0 \quad \text{for} \quad i = 1, \cdots, n, \tag{1.2–3}$$

and

$$\sum_{i=1}^{n} F_i(1) = 1. \tag{1.2–4}$$

(Note: An R^n policy function can be given several conventional statistical interpretations. For instance, by properly juxtaposing the n components, an R^n policy

function can be viewed as a cumulative distribution defined on the interval $[0, n]$.) Player 1's strategy space, Δ, is defined to be the set of all R^n policy functions F, and player 2's strategy space, Γ, is the set of all R^m policy functions G. The triple $\{\Delta, \Gamma, K\}$ is called a *continuous* game, where K determines the payoff functional

$$\mathcal{K}[F, G] = \int_0^1 dF(x) \cdot \int_0^1 K(x, y)\, dG(y).$$

In these games, each player can be thought of as having n (or m) ordered copies of the unit interval. He must decide how to allocate a certain fractional part of his resources to each interval and how to distribute that fractional part over the interval.

A game is *symmetric* when the roles of the two players are interchangeable. In the case of the two classes of games described above, this property is equivalent to the skew-symmetry of the kernel K (i.e., $K(x, y) = -K^T(y, x)$). Hence in symmetric games, $m = n$, and the strategy spaces are identical. It can also be shown that the value, if it exists, is zero.

1.3. The dynamic model and the associated Lyapunov functional. First we shall discuss L^∞ games. Consider the symmetric L^∞ game $\{\mathcal{F}, \mathcal{G}, K\}$. A solution to its dynamical model described in this section will be a function $g(x, s)$ mapping $[0, 1] \times [0, \infty)$ into R^n, where $g(\cdot, s) \in \mathcal{G}$ for all $s \in [0, \infty)$ and $(\partial g/\partial s)(x, s)$ exists for all x and s. Before proceeding directly to a description of the dynamical system, some preliminary notation will be helpful. Let

$$(Wg)(x, s) = \int_I K(x, y) g(y, s)\, d\mu(y), \tag{1.3–1}$$

where K is the $n \times n$ L^∞ kernel for the game $\{\mathcal{F}, \mathcal{G}, K\}$ being modeled and g is as described above. We shall let $(Wg)_i$ denote the ith row of Wg, $\phi_i[(Wg)] = \max(0, (Wg)_i)$, and let ϕ represent the n-vector with components ϕ_i. Using this notation, the dynamical model for $\{\mathcal{F}, \mathcal{G}, K\}$ is the vector equation,

$$\frac{\partial g}{\partial s}(x, s) = \phi[(Wg)(x, s)] - \sum_{i=1}^{n} \int_I \phi_i[(Wg)(x, s)]\, d\mu(x) g(x, s), \tag{1.3–2}$$

with the initial condition $g(x, 0) = g_0(x)$ a.e., where $g_0 \in \mathcal{G}$. To investigate the stability of (1.3–2), we will introduce the following associated Lyapunov functional:

$$\Psi(g(\cdot, s)) = \sum_{i=1}^{n} \int_I \phi_i^2[(Wg)(x, s)]\, d\mu(x). \tag{1.3–3}$$

To understand the relationship (for symmetric games) between Ψ and the payoff functional $\mathcal{K}$, consider the following. Suppose for some s, $g(\cdot, s) \in \mathcal{G}$ and $\psi(g(\cdot, s)) = 0$. Then each $\phi_i[(Wg)(x, s)] = 0$ for a.e. x, i.e., $(Wg)_i(x, s) \leqq 0$ for a.e. x and $i = 1, \cdots, n$. But then from (1.3–1) we see that $\mathcal{K}(f, g(\cdot, s)) \leqq 0$ for all f in $\mathcal{F} = \mathcal{G}$. Since the value of a symmetric game is zero (if it exists), $g(\cdot, s)$ is a solution to the game. In Theorem 1.3–3, we will show that the dynamical model (1.3–2) steers $\Psi(g(\cdot, s))$ to zero, and thereby achieve a means of approximating solutions to the game.

We shall now state the main results obtained for symmetric L^∞ games and the dynamic model (1.3–2).

THEOREM 1.3–1. *Suppose* $\{\mathscr{F}, \mathscr{G}, K\}$ *is an* L^∞ *game and let* (1.3–2) *be its dynamic model. Then there exists a unique function* g *satisfying* (1.3–2) *and its initial condition on* $I \times [0, \infty)$, *where* $g(\cdot, s) \in \mathscr{G}$ *for each* $s \in [0, \infty)$.

Since the solution to L^∞ games, in general, will be only approximate, we will need the concept of an ε-optimal strategy.

DEFINITION 1.3–2. Suppose $\{\mathscr{F}, \mathscr{G}, K\}$ is an L^∞ game. A set of strategy pairs $\{(f_\varepsilon, g_\varepsilon)\}$ are ε-optimal if there exists a real number γ such that for each $\varepsilon > 0$ there exists a $(f_\varepsilon, g_\varepsilon)$ for which

$$\int_I d\mu(x) \int_I f(x) \cdot (K(x, y) g_\varepsilon(y))\, d\mu(y) < \gamma + \varepsilon \tag{1.3–4}$$

for all f in $\mathscr{F}$ and

$$\int_I d\mu(x) \int_I f_\varepsilon(x) \cdot (K(x, y) g(y))\, d\mu(x) > \gamma - \varepsilon \tag{1.3–5}$$

for all g in $\mathscr{G}$.

γ is called the value of the game. The theorem on the ε-optimal solution for symmetric L^∞ games can now be stated.

THEOREM 1.3–3. *Suppose* g *is a solution to the dynamical model for the symmetric* L^∞ *game* $(\mathscr{G}, \mathscr{G}, K\}$ *given by* (1.3–2) *and its initial condition. Let* $\{(g(\cdot, s_m))\}$ *be any sequence of strategies in* $\mathscr{G}$ *defined by the solution* g. *Then given any* $\delta > 0$, *there exists a set* $E_\delta \subset I$, *where* $\mu(E_\delta) < \delta$, *such that if we set* $K(x, y) = 0$ *for all* $(x, y) \in E_\delta \times E_\delta$ *(denote this new kernel by* K_δ*), then for any* $\varepsilon > 0$ *and* m *sufficiently large,* $\{g(\cdot, s_m)\}$ *are* ε*-optimal strategies for both players to the new symmetric game* $\{\mathscr{G}, \mathscr{G}, K_\delta\}$.

Now we shall state the results for symmetric continuous games. Consider the symmetric continuous game $\{\Gamma, \Gamma, K\}$ and let $\{\mathscr{G}, \mathscr{G}, K\}$ be the L^∞ game with the same game kernel. Let (1.3–2) be the model for the game $\{\mathscr{G}, \mathscr{G}, K\}$. Suppose g is the solution to (1.3–2) guaranteed by Theorem 1.3–1 and define

$$G(x, s) = \int_0^x g(t, s)\, d\mu(t). \tag{1.3–6}$$

Observe that $G(\cdot, s) \in \Gamma$ for all $s \in [0, \infty)$.

THEOREM 1.3–4. *Suppose* $\{\Gamma, \Gamma, K\}$ *is a symmetric continuous game and suppose* G *is given by* (1.3–6). *Then there exists a sequence* $\{G(\cdot, s_m)\}$ *and a function* G_∞ *in* Γ *such that for each* x *in* I *at which* G_∞ *is continuous,* $\lim_{s_m \to \infty} G(x, s_m) = G_\infty(x)$. *Furthermore,* G_∞ *is a solution for both players to the game* $\{\Gamma, \Gamma, K\}$.

Notice that here the solution is exact.

2.1. The existence of a unique a.e. solution to the dynamic model. We shall now examine the dynamic model (1.3–2) in detail. In that the system is nonlinear, the proof of existence and uniqueness is nontrivial. Our program will be the following. First we shall prove the existence and uniqueness of a solution to (1.3–2) on $I \times [\alpha, \alpha + \beta]$ for arbitrary α and sufficiently small β, and then we shall show how

the dynamical system properly constrains its solutions on this rectangle. We will conclude § 2 with the proof of Theorem 1.3–1.

2.2. $\mathscr{C}(I_\beta^\alpha, n)$ Banach spaces. The problems of existence and uniqueness will be based upon the "contraction mapping" fixed point theorem. Consequently, our first task will be to develop an appropriate metric space in which to analyze the problem. Let $I_\beta^\alpha = [\alpha, \alpha + \beta]$.

DEFINITION 2.2–1. $\mathscr{C}(I_\beta^\alpha, n)$ is the set of all real functions mapping $I \times I_\beta^\alpha$ into R^n such that

(a) for all $s \in I_\beta^\alpha$, $f(\cdot, s)$ is Lebesgue measurable;

(b) for a.e. $x \in I$, $f(x, \cdot)$ is continuous

and

$$\|f\| = \sum_{i=1}^{n} \int_I \sup_s |f_i(x, s)|\, d\mu(x) < \infty, \tag{2.2–1}$$

where f_i denotes the ith component of f and μ is the Lebesgue measure defined on the Lebesgue measurable sets of I.

For simplicity we shall study the case $n = 1$ and then show how the results can be extended for arbitrary n. Suppose $n = 1$. The measurability of $\sup |f(x, s)|$ needed in (2.2–1) is verified by the following lemma.

LEMMA 2.2–2. *Suppose f is a function mapping $I \times I_\beta^\alpha$ into R with properties* (a) *and* (b) *given in Definition* 2.2–1. *Then* $\sup_s |f(x, s)|$ *is Lebesgue measurable.*

The proof can be found in [7].

LEMMA 2.2–3. $\mathscr{C}(I_\beta^\alpha, 1)$ *is a real seminormed linear space.*

The proof of this lemma is a simple application of the preceding lemma and the triangle inequality.

The seminorm of Lemma 2.2–3 fails to be a norm since $\|f\| = 0$ does not imply that f is the zero vector of $\mathscr{C}$. Therefore, we regard two elements f and g as equivalent if $\|f - g\| = 0$. Let f denote the equivalent class containing $f \in \mathscr{C}$ and $\tilde{\mathscr{C}}(I_\beta^\alpha, 1)$ denote the collection of equivalent classes. With the vector space operation defined in the obvious manner and $\|\tilde{f}\| = \|f\|$, one can easily check that $\tilde{\mathscr{C}}$ becomes a normed linear space.

THEOREM 2.2–4. $\tilde{\mathscr{C}}(I_\beta^\alpha, 1)$ *is a Banach space.*

The proof follows the structure of the classical proof of completeness of L^p-spaces. The details are given in [7].

COROLLARY 2.2–5. $\tilde{\mathscr{C}}(I_\beta^\alpha, n)$ *is a Banach space for arbitrary n.*

2.3. The existence and uniqueness of a solution to (1.3–2) where the "s" variable is sufficiently bounded. In this section, K will be assumed to be the kernel of the L^∞ game $\{\mathscr{F}, \mathscr{G}, K\}$ being modeled. Let $k = \sup_{(i,j)} \operatorname{ess\,sup}_{(x,y)} |K_{ij}(x, y)|$, where the "ess sup" is with respect to the measure space $\{I \times I, M \times M, \mu \times \mu\}$ (see the paragraph preceding Definition 1.2–1). Define $\beta = (10\, nk)^{-1}$. Recall n is the dimension of the game's kernel. Throughout the section it will always be assumed that α is an arbitrary real number.

LEMMA 2.3–1. *Suppose K is an $n \times n$ L^∞ kernel whose elements K_{ij} are essentially bounded by k, and that $g \in \mathscr{C}(I_\beta^\alpha, n)$. Then $(Wg)(x, \cdot)$ is continuous on I_β^α for a.e. $x \in I$ and $(Wg)(\cdot, s)$ is measurable on I for all $s \in I_\beta^\alpha$. Furthermore, recalling the definition of $\tilde{\mathscr{C}}$, if $\tilde{h} = \tilde{g}$, then $\tilde{W}g = \tilde{W}h$. Also for a.e. x and all s,*

$$|(Wg)_i(x,s)| \leqq k\|g\|, \qquad i = 1, \cdots, n. \tag{2.3–1}$$

For the proof of this Lemma, see [7].

LEMMA 2.3–2. *Let $g, h \in R^1$; $u, v \in R^n$ and denote $\phi_i(u) = \max(0, u_i)$. Then*

$$|\phi_i(u) - \phi_i(v)| \leqq |u_i - v_i|, \qquad i = 1, \cdots, b, \tag{2.3–2}$$

and

$$\left| g \sum_{i=1}^{n} \phi_i(u) - h \sum_{i=1}^{n} \phi_i(v) \right| \leqq |g - h| \sum_{i=1}^{n} \phi_i(u) + |h| \sum_{i=1}^{n} |u_i - v_i|. \tag{2.3–3}$$

The proof follows from the observation that $\max(0, u_i) = (|u_i| + u_i)/2$. Let $B(2) = \{g \in \mathcal{G}; \|g\| \leqq 2\}$.

LEMMA 2.3–3. *Define $(Tg)(x,s)$ by*

$$(Tg)(x,s) = \phi[(Wg)(x,s)] - \sum_{i=1}^{n} \int_I \phi_i[(Wg)(t,s)]\, d\mu(t) g(x,s) \tag{2.3–4}$$

for $g \in B(2)$. $(Tg)(x, \cdot)$ is continuous on I_β^α for a.e. $x \in I$ and $(Tg)(\cdot, s)$ is Lebesgue measurable in x for all $s \in I_\beta^\alpha$. Furthermore, T is Lipschitz continuous in $B(2)$ with a Lipschitz constant $\gamma = 5nk$.

Proof. Assume $g \in B(2)$. From Lemma 2.3–1, we see that $\phi[(Wg)(x,s)]$ is measurable in x for each s in I_β^α. Hence $(Tg)(\cdot, s)$ is measurable for each $s \in I_\beta^\alpha$.

We also see from Lemma 2.3–1 that $\phi[(Wg)(x,s)]$ is continuous in s for a.e. x in I, and that each $\phi_i[(Wg)(x,s)] \leqq k\|g\| < \infty$ for a.e. x and all s. Therefore, by the Lebesgue dominated convergence theorem, for any sequence $s_m \to s$,

$$\lim_{s_m \to s} \int_I \phi_i[(Wg)(t, s_m)]\, d\mu(t) = \int_I \phi_i[(Wg)(t,s)]\, d\mu(t). \tag{2.3–5}$$

We conclude that for a.e. x, each term in (2.3–4) is continuous in s and thus $(Tg)(x, \cdot)$ is continuous on I_β^α for a.e. $x \in I$.

We shall now show the Lipschitz continuity of T. Suppose g and $h \in B(2)$. From the inequalities obtained in Lemma 2.3–2 and the linearity of W, we see that for $i = 1, \cdots, n$,

$$\begin{aligned} |(Tg)_i(x,s) - (Th)_i(x,s)| &= \Big| \phi_i[(Wg)(x,s)] - \phi_i[(Wh)(x,s)] \\ &\quad - \int_I \Big\{ g_i(x,s) \sum_{j=1}^{n} \phi_j[(Wg)(t,s)] \\ &\qquad - h_i(x,s) \sum_{j=1}^{n} \phi_j[(Wh)(t,s)] \Big\}\, d\mu(t) \Big| \\ &\leqq |(W(g-h))_i(x,s)| \\ &\quad + |g_i(x,s) - h_i(x,s)| \sum_{j=1}^{n} \int_I \phi_j[(Wg)(t,s)]\, d\mu(t) \\ &\quad + |h_i(x,s)| \sum_{j=1}^{n} \int_I |W(g-h)_j(t,s)|\, d\mu(t). \end{aligned} \tag{2.3–6}$$

Applying (2.3–1) to (2.3–6), we get for a.e. x and for any $s, i = 1, \cdots, n$,

$$\begin{aligned} |(Tg)_i(x,s) - (Th)_i(x,s)| &\leqq k\|g - h\| \\ &\quad + nk|g_i(x,s) - h_i(x,s)|\,\|g\| \\ &\quad + nk|h_i(x,s)|\,\|g - h\|. \end{aligned} \tag{2.3–7}$$

Taking the sup over s of the right-hand side of (2.3–7), then taking the sup over the left-hand side of the resulting inequality, integrating on x, and finally summing, we obtain

$$\|Tg - Th\| \leqq nk\{1 + \|g\| + \|h\|\}\|g - h\|. \tag{2.3–8}$$

Since $\|g\| \leqq 2$ and $\|h\| \leqq 2$, T is a Lipschitz continuous with a Lipschitz constant of $\gamma = 5nk$.

LEMMA 2.3–4. *Suppose f maps $I \times [a,b]$ into R, where for each s in $[a,b]$, $f(\cdot, s)$ is Lebesgue measurable for all s in $[a,b]$, and $f(x,\cdot)$ is continuous for a.e. x in I. Then $\int_a^b f(x,s)\,ds$ is measurable.*

The proof can be found in [7].

THEOREM 2.3–5. *Let g_0 be any function in $B(2)$ and let* (1.3–2) *be the dynamical model for the symmetric L^∞ game $\{\mathcal{G}, \mathcal{G}, K\}$ with initial function g_0. Then there exists a unique function $g \in B(2)$ satisfying* (1.3–2) *and $g(x,0) = g_0(x)$ a.e.*

Proof. Pick $g_0 \in \mathcal{G}$ and suppose $g \in B(2)$. Let E denote the set of all x, where $(Tg)(x,\cdot)$ is not continuous. Define

$$(\mathcal{T}g)(x,s) = \begin{cases} \displaystyle\int_\alpha^s (Tg)(x,\sigma)\,d\sigma + g_0(x) & \text{if } (x,s) \in (I - E) \times I_\beta^\alpha, \\ 0 & \text{if } (x,s) \in E \times I_\beta^\alpha. \end{cases} \tag{2.3–9}$$

We want to show that $\mathcal{T}$ is a contraction mapping $B(2)$ into $B(2)$.

By Lemma 2.3–4, we know for each s, $(\mathcal{T}g)(\cdot, s)$ is measurable. By the definition of E, $(\mathcal{T}g)(x,\cdot)$ is differentiable (and continuous).

Since $\mu(E) = 0$, from the Lipschitz condition on T shown in Lemma 2.3–3 we see that

$$\begin{aligned} \|\mathcal{T}g\| &= \sum_{i=1}^n \int_I \sup_s \left| \int_\alpha^s (Tg)_i(x,\sigma)\,d\sigma + g_0(x) \right| d\mu(x) \\ &\leqq \sum_{i=1}^n \int_I \int_\alpha^{\alpha+\beta} |(Tg)_i(x,\sigma)|\,d\sigma\,d\mu(x) + \|g_0\| \\ &\leqq \beta\|Tg\| + \|g_0\| \\ &\leqq \beta\gamma\|g\| + \|g_0\|. \end{aligned} \tag{2.3–10}$$

But $\|g_0\| \leqq 1$, $\beta = (10nk)^{-1}$, and $\sigma = 5nk$ and we conclude $\|\mathcal{T}g\| \leqq 2$. Hence $\mathcal{T}$ maps $B(2)$ into itself.

Suppose g and $h \in B(2)$. Following steps similar to (2.3–10), we have

$$\begin{aligned}\|\mathscr{T}g - \mathscr{T}h\| &= \sum_{i=1}^{n} \int_I \sup_s \left| \int_\alpha^s (Tg)_i(x,\sigma) - (Th)_i(x,\sigma)\, d\sigma \right| d\mu(x) \\ &\leqq \sum_{i=1}^{n} \int_I \int_\alpha^{\alpha+\beta} |(Tg)_i(x,\sigma) - (Th)_i(x,\sigma)|\, d\sigma\, d\mu(x) \\ &\leqq \beta \|Tg - Th\| \\ &\leqq \beta\gamma \|g - h\| \\ &\leqq \tfrac{1}{2}\|g - h\|.\end{aligned} \tag{2.3–11}$$

Therefore, $\mathscr{T}$ is a contraction.

The contraction mapping theorem requires a complete metric space. Hence we will have to define a map corresponding to $\mathscr{T}$ defined on $\tilde{B}(2)$, the set of all equivalence classes of $B(2)$. $\tilde{B}(2)$ is a closed subset of $\tilde{\mathscr{C}}(I_\beta^\alpha, n)$, and thus a complete metric space. Using $\tilde{g}$ to denote the equivalence class of g, define $\tilde{\mathscr{T}}\tilde{g} = \mathscr{T}g$ for any $g \in \tilde{g}$. It can easily be verified using Lemma 2.3–1 that $\tilde{\mathscr{T}}$ is well-defined and maps $\tilde{B}(2)$ into $\tilde{B}(2)$. Therefore, by the fixed point theorem for a contractive mapping, there exists a $\tilde{g}$ in $\tilde{B}(2)$ such that $\tilde{\mathscr{T}}\tilde{g} = \tilde{g}$; and hence, for any g in $\tilde{g}$ for a.e. x and all s, $(\mathscr{T}g)(x,s) = g(x,s)$. Let $\mathscr{E}$ denote the set where this equality fails. Let $g(x,s) = 0$ for $(x,s) \in \mathscr{E} \times I_\beta^\alpha$. Since $\mu(\mathscr{E}) = 0$, by the definition of $\mathscr{T}$ in (2.3–9),

$$\frac{\partial g}{\partial s}(x,s) = (Tg)(x,s). \tag{2.3–12}$$

If $(x,s) \in (I - \mathscr{E}) \times I_\beta^\alpha$, then $g(x,0) = g_0(x)$. Since $\mu(\mathscr{E}) = 0$, $g(x,0) = g_0(x)$ a.e.

2.4. The constraint condition on the solution to the dynamic model. We will maintain the same hypotheses set forth at the beginning of § 2.3.

THEOREM 2.4–1. *Suppose g is the function in $B(2)$ satisfying* (1.3–2) *with the initial condition $g(x,0) = g_0(x) \in \mathscr{G}$. Then for each s in I_β^α, $g(\cdot, s) \in \mathscr{G}$.*

Proof. Suppose g is the solution to (1.3–2) and its initial condition. Let $\Phi(s) = \sum_{i=1}^n \int_I \phi_i[(Wg)(x,s)]\, d\mu(x)$. Then for a.e. x and all s,

$$g(x,s) = \exp\left(\int_{\bar{\alpha}}^s \Phi(\sigma)\, d\sigma\right)\left\{ g_0(x) + \int_\alpha^s \exp\left(\int_\alpha^\sigma \Phi(\omega)\, d\omega\right) \phi[(Wg)(x,\sigma)]\, d\sigma \right\}. \tag{2.4–1}$$

(Note: $(Wg)_i(x, \cdot)$ need not be bounded on a set of measure zero, and hence equality (2.4–1) is for a.e. x and all s.) The validity of (2.4–1) can be checked by differentiating on s. From (2.4–1) and the definition of ϕ, it follows that if $g_0(x) \geqq 0$ a.e., then $g(x,s) \geqq 0$ for a.e. x in I and all s in I_β^α.

Summing and integrating the components of (1.3–2), we get

$$\sum_{i=1}^{n} \int_I \frac{\partial g_i}{\partial s}(x,s)\, d\mu(x) = \Phi(s) - \Phi(s) \sum_{i=1}^{n} \int_I g_i(x,s)\, d\mu(x). \tag{2.4–2}$$

Suppose

$$\frac{d}{ds}\sum_{i=1}^{n}\int_I g_i(x,s)\,d\mu(x) = \sum_{i=1}^{n}\int_I \frac{\partial g_i}{\partial s}(x,s)\,d\mu(x). \tag{2.4–3}$$

Then from (2.4–2) we obtain

$$\frac{d}{ds}\sum_{i=1}^{n}\int_I g_i(x,s)\,d\mu(x) = \Phi(s)\left\{1 - \sum_{i=1}^{n}\int_I g_i(x,s)\,d\mu(x)\right\}, \tag{2.4–4}$$

It follows from (2.4–4) that if initially

$$\sum_{i=1}^{n}\int_I g_i(x,\alpha)\,d\mu(x) = 1,$$

then

$$\sum_{i=1}^{n}\int_I g_i(x,s)\,d\mu(x) = 1$$

for all s in I_β^α.

All that remains to be shown is the validity of (2.4–3). From (1.3–2) and Lemma 2.3–1 we have, for a.e. x and all s,

$$\left|\sum_{i=1}^{n}\frac{\partial g_i}{\partial g}(x,s)\right| \leqq k\|g\|\left(n + \sum_{i=1}^{n}\sup_s |g_i(x,s)|\right). \tag{2.4–5}$$

Since $g \in B(2)$, the integral of the right-hand side of (2.4–5) is finite; (2.4–3) now follows from an application of the mean value theorem and the Lebesgue dominated convergence theorem [2, p. 60].

2.5. The proof of Theorem 1.3–1. This proof globalizes the local solution established in Theorem 2.4–1. Consider the sequence of intervals $\{I_\beta^n = [(n-1)\beta, n\beta]\}$ and $\{I_n = [0, n\beta]\}$, where $\beta = (10nk)^{-1}$ as defined in § 2.3. Let g_0 be the initial function given in the hypothesis of Theorem 1.3–2. Since α was arbitrary, from Theorem 2.3–5 we know there exists a function g_1 satisfying (1.3–2) and $g_1(x,0) = g_0(x)$ a.e. on $I \times I_1$. Furthermore, from Theorem 2.4–1, $g_1(\cdot, s) \in \mathcal{G}$ for all s in I_1. Applying the same theorems again, we see that there exists a unique solution $\bar{g}_2$ to (1.3–2) on $I \times I_\beta^2$, where we select the initial function $\bar{g}_2(x,\beta) = g_1(x,\beta)$ a.e.

Define

$$g_2(x,s) = \begin{cases} g_1(x,s) & \text{if } s \in I_1, \\ \bar{g}_2 & \text{if } s \in I_\beta^2. \end{cases} \tag{2.5–1}$$

It can be easily verified that g_2 satisfies (1.3–2) and $g_2(x,0) = g_0(x)$ a.e. on $I \times I_2$. Also $g_2(\cdot, s) \in \mathcal{G}$ for all $s \in I_{n-1}$.

Now suppose g_{n-1} satisfies (1.3–2) and $g_{n-1}(x,0) = g_0(x)$ a.e. on $I \times I_{n-1}$ and that $g_{n-1}(\cdot, s) \in \mathcal{G}$ for $s \in I_{n-1}$. Suppose $\bar{g}_n$ is the solution to (1.3–2) on $I \times I_\beta^n$ satisfying $\bar{g}_n(x,(n-1)\beta) = g_n(x,(n-1)\beta)$ a.e. Define

$$g_n(x,s) = \begin{cases} g_{n-1}(x,s) & \text{if } s \in I_{n-1}, \\ \bar{g}_n(x,s) & \text{if } s \in I_\beta^n. \end{cases} \tag{2.5–2}$$

Again it can be readily verified that g_n satisfies (1.3–2), $g_n(x, 0) = g_0(x)$ a.e., and $g_n(x, s) \in \mathcal{G}$ for all $s \in I_n$. It now follows by induction that there exists a function satisfying (1.3–2) and its initial condition on $I \times [0, \infty)$, where $g(\cdot, s) \in \mathcal{G}$ for all s in $[0, \infty)$.

3. A method for solving symmetric L^∞ and continuous games using their dynamic models. First we shall consider L^∞ games. Suppose that (1.3-2) is the model for the symmetric L^∞ game $\{\mathcal{G}, \mathcal{G}, K\}$ and that g is its solution. Using a differential inequality to be obtained in Lemma 3.3, we will show that $\Psi(g(\cdot, s))$ decreases monotonically to zero. Then we will complete the proof of Theorem 1.3–3.

Next we shall consider continuous games. Suppose $\{\Gamma, \Gamma, K\}$ is a symmetric continuous game and let $\{\mathcal{G}, \mathcal{G}, K\}$ be the L^∞ game corresponding to $\{\Gamma, \Gamma, K\}$ with the same kernel. Suppose that (1.3–2) is the model for $\{\mathcal{G}, \mathcal{G}, K\}$ and that g is its solution. Define $G(x, s)$ by (1.3–6) and denote

$$
\begin{aligned}
(WG)(x, s) &= \int_0^1 K(x, y)\, dG(y, s) \\
&= \int_0^1 K(x, y) g(y, s)\, d\mu(y) \\
&= (Wg)(x, s).
\end{aligned} \tag{3.1}
$$

Define

$$
\Psi(G(\cdot, s)) = \sum_{i=1}^{n} \int_0^1 \phi_i^2[(Wg)(x, s)]\, dx = \Psi(g(\cdot, s)), \tag{3.2}
$$

where ϕ_i are defined as before. Observe that each $\phi_i[(WG)(\cdot, s)]$ is continuous for any s in $[0, \infty)$. As in the case of L^∞ games, $\Psi(G(\cdot, s))$ decreases monotonically to zero, but here a weak convergence argument will prove the existence of a sequence $\{G(\cdot, s_m)\} \in \Gamma$ and a limit function $G_\infty \in \Gamma$ such that $\lim_{s_m \to \infty} \Psi((\cdot, s_m)) = \Psi(G_\infty) = 0$. It will follow that G_∞ is an optimal strategy to the game $\{\Gamma, \Gamma, K\}$ for both players, and thus Theorem 1.3–4 will be proven. This will conclude our analysis of symmetric games.

We will need the following lemmas for the proofs of Theorems 1.3–3 and 1.3–4.

LEMMA 3.1. *Suppose f is a differentiable function mapping $[0, \infty)$ into R. Let $\phi = \max(0, f)$. Then*

$$
d\frac{\phi^2(f(s))}{ds} = 2\phi(f(s))\frac{df(s)}{ds}. \tag{3.3}
$$

The differentiability of $\phi^2(f(s))$ is possible because the squaring operation sufficiently smoothes the composite function at points where $f(s) = 0$. The details are in [7].

LEMMA 3.2. *Suppose* $\{\mathcal{G}, \mathcal{G}, K\}$ *is an* L^∞ *game and its dynamical model is given by* (1.3–2). *Let* g *be the solution to the dynamical model. Then for a.e.* x *and all* s,

$$\begin{aligned}\frac{\partial}{\partial s}(Wg)(x,s) &= \int_I K(x,y)\phi[(Wg)(y,s)]\,d\mu(y) \\ &\quad - \sum_{i=1}^{n}\int_I \phi[(Wg)(y,s)]\,d\mu(y)(Wg)(x,s).\end{aligned} \tag{3.4}$$

Proof. It can be shown using the mean value theorem and the Lebesgue dominated convergence theorem [2, p. 60] that for a.e. x, all s, and $i = 1, \cdots, n$,

$$\begin{aligned}\frac{\partial}{\partial s}(Wg)_i(x,s) &= \frac{\partial}{\partial s}\sum_{j=1}^{n}\int_I K_{ij}(x,y)g_j(y,s)\,d\mu(y) \\ &= \sum_{j=1}^{n}\int_I K_{ij}(x,y)\frac{\partial}{\partial s}g_j(y,s)\,d\mu(y).\end{aligned} \tag{3.5}$$

Substituting (1.3–2) into (3.5), we obtain for a.e. x and s, (3.4).

LEMMA 3.3. *Suppose* g *is the solution to the dynamical model of the symmetric* L^∞ *game* $\{\mathcal{G}, \mathcal{G}, K\}$ *given by* (1.3–2). *Then if* $k = \sup_{(i,j)} \operatorname{ess\,sup}_{(x,y)}|K_{ij}(x,y)| \neq 0$,

$$\frac{d\Psi}{ds}(g(\cdot,s)) \leqq -\frac{2}{k}\Psi^2(g(\cdot,s)). \tag{3.6}$$

Proof. Suppose g is the solution to (1.3–2). First we will show Ψ is differentiable. From Lemmas 3.1 and 3.2, it follows for a.e. x and all s that

$$\frac{\partial}{\partial s}\phi^2[(Wg)(x,s)] = 2\phi[(Wg)(x,s)]\frac{\partial}{\partial s}(Wg)(x,s).$$

It can be shown that for all s, $\phi[(Wg)(\cdot,s)](\partial/\partial s)(Wg)(\cdot,s)$ is bounded uniformly by an integrable function. Furthermore, we know for any s that $\phi^2[(Wg)(\cdot,s)]$ is measurable. We can now conclude from the mean value theorem and the Lebesgue dominated convergence theorem [2, p. 60] that

$$\begin{aligned}\frac{d\Psi}{ds}(g(\cdot,s)) &= \frac{\partial}{\partial s}\sum_{i=1}^{n}\int_I \phi_i^2[(Wg)(x,s)]\,d\mu(x) \\ &= 2\int_I \phi[(Wg)(x,s)]\frac{\partial}{\partial s}(Wg)(x,s)\,d\mu(x).\end{aligned} \tag{3.7}$$

Substituting (3.4) into (3.7), we obtain

$$\begin{aligned}\frac{d\psi}{ds}(g(\cdot,s)) &= 2\int_I \{\phi[(Wg)(x,s)]\int_I K(x,y)\phi[(Wg)(y,s)]\,d\mu(y)\}\,d\mu(x) \\ &\quad - 2\sum_{i=1}^{n}\int_I \phi_i[(Wg)(x,s)]\,d\mu(x) \\ &\quad \cdot \int_I \phi[(Wg)(x,s)]\cdot(Wg)(x,s)\,d\mu(x).\end{aligned} \tag{3.8}$$

The skew-symmetry of K makes the first term in (3.8) zero, and we get

$$\frac{d\psi}{ds}(g(\cdot, s)) = -2 \sum_{i=1}^{n} \int_I \phi_i[(Wg)_i(x, s)]\, d\mu(x)\psi(g(\cdot, s)). \tag{3.9}$$

It can easily be verified that

$$\begin{aligned} \psi(g(\cdot, s)) &= \sum_{i=1}^{n} \int_I \phi_i^2[(Wg)(x, s)]\, d\mu(x) \\ &\leqq k \sum_{i=1}^{n} \int_I \phi_i[(Wg)(x, s)]\, d\mu(x). \end{aligned} \tag{3.10}$$

Equation (3.6) now follows from (3.9) and (3.10).

Now we will prove Theorem 1.3–3.

Proof of Theorem 1.3–3. Suppose g is a solution to the dynamical model for the symmetric L^∞ game $\{\mathcal{G}, \mathcal{G}, K\}$. If $k = \sup_{(i,j)} \operatorname{ess\,sup}_{(x,y)} |K_{ij}(x, y)| = 0$, then the theorem follows trivially.

From the preceding lemma, if $k \neq 0$, then

$$\frac{d\psi}{ds}(g(\cdot, s)) \leqq -\frac{1}{k}\psi^2(g(\cdot, s)). \tag{3.11}$$

Integrating both sides of this differential inequality gives us

$$\psi(g(\cdot, s)) \leqq \frac{k\psi(g(\cdot, 0))}{k + s\psi(g(\cdot, 0))}. \tag{3.12}$$

If $\psi(g(\cdot, s)) = 0$, then (3.11) implies it will remain zero and (3.13) is still valid. From the definition of ψ and the Schwarz inequality,

$$\left(\int_I \phi_j[(Wg)(x, s)]\, d\mu(x) \right)^2 \leqq \psi(g(\cdot, s)), \qquad j = 1, \cdots, n, \tag{3.13}$$

which along with (3.12) implies that each $\int_I \phi_j[(Wg)(x, s)]\, d\mu(x) \to 0$ as $s \to \infty$. Let $s_m \to \infty$. Then

$$\int_I |\phi_j[(Wg)(x, s_m)] - 0|\, d\mu(x) \to 0 \quad \text{as } m \to \infty, \tag{3.14}$$

$j = 1, \cdots, n$. In other words, each $\{\phi_j[(Wg)(\cdot, s_m)]\}$ converges in the mean to the zero function. Therefore, $\{\phi_j[(Wg)(\cdot, s_m)]\}$ converges in measure to zero [2, p. 49]. Since the sequence converges in measure to zero, there is a subsequence $\{\phi_j[(Wg) \cdot (\cdot, s_{m_k})]\}$ which converges almost uniformly to the zero function [2, p. 36]. There is a subsequence $\{s_{m_k}\}$ corresponding to each component; pick a subsequence $\{s_{m_k}\}$ which is included in all of the subsequences. Almost uniform convergence implies for any $\delta > 0$, there exists a measurable set $E_\delta \subset I$ such that $\mu(E_\delta) < \delta$ and for $j = 1, \cdots, n$, $\{\phi_j[(Wg)(\cdot, s_{m_k})]\}$ converges uniformly to the zero function on $I - E_\delta$. Hence given any $\varepsilon > 0$, for s_{m_k} sufficiently large,

$$0 \leqq \phi_j[(Wg)(x, s_{m_k})] < \varepsilon \tag{3.15}$$

for all $x \in I - E_\delta$, $j = 1, \cdots, n$. Therefore because of (3.15) and the definition of ϕ, for any $\varepsilon > 0$ and all s_{m_k} sufficiently large,

$$(Wg)_i(x, s_{m_k}) = \sum_{j=1}^{n} \int_I K_{ij}(x, y) g_i(y, s_{m_k})\, d\mu(y) < \varepsilon \tag{3.16}$$

for all $x \in I - E_\delta$, $j = 1, \cdots, n$. Suppose we set $K(x, y) = 0$ for $(x, y) \in E_\delta \times E_\delta$. Then (3.16) implies

$$\sum_{j=1}^{n} \int_I K_{\delta_{ij}}(x, y) g_j(y, s_{m_k})\, d\mu(y) < \varepsilon \tag{3.17}$$

for every $\varepsilon > 0$ and all correspondingly large s_{m_k}, $i = 1, \cdots, n$. For any $f \in \mathscr{F} = \mathscr{G}$, if we multiply both sides of (3.17) by f_i, integrate on x, and sum on i, we see that given $\varepsilon > 0$,

$$\int_I \int_I f(x) \cdot (K_\delta(x, y) g(y, s_{m_k}))\, d\mu(y)\, d\mu(x) \leqq \varepsilon \tag{3.18}$$

for s_{m_k} sufficiently large. Thus $g(y, s_{m_k})$ is an ε-optimal strategy for player 2 if s_{m_k} is sufficiently large. Since K_δ is skew-symmetric, by changing the order of integration in (3.18), and taking the transpose of the integrand, we know for any $\varepsilon > 0$,

$$\int_I \int_I g(x, s_{m_k}) \cdot K_\delta(x, y) f(y)\, d\mu(y)\, d\mu(x) > -\varepsilon \tag{3.19}$$

for all f in $\mathscr{F}$. But $\mathscr{F} = \mathscr{G}$, and thus, $g(x, s_{m_k})$ is ε-optimal for player 1 if s_{m_k} is sufficiently large. This completes the proof.

When the kernel for an L^∞ game is a constant matrix, then the R^n density functions f and g may be taken to be vectors x and y in the usual game simplex in R^n. Consequently, an L^∞ game with a constant matrix kernel reduces to a matrix game and the dynamic model (1.3–2) reduces to an ordinary differential equation. This differential equation is in agreement with Brown and von Neumann's results [1]. As was shown in their paper, in these games we have the weak convergence needed to obtain an optimal solution for the symmetric matrix game being modeled.

Proof of Theorem 1.3–4. Suppose $\{\Gamma, \Gamma, K\}$ is the given symmetric continuous game. Let $\{\mathscr{G}, \mathscr{G}, K\}$ be the L^∞ game with the same kernel and let (1.3–2) be the dynamic model for $\{\mathscr{G}\ \mathscr{G}\ K\}$. By Theorem 1.3–1, there exists a solution $g(x, s)$ to (1.3–2). Define $G(x, s) = \int_0^x g(t, s)\, d\mu(t)$. Pick any sequence $s_m \to \infty$. From Theorem 1.3–3 and the definition of G, we know that each member of the sequence $\{G(\cdot, s_m)\}$ is in Γ. Since each component of $G(x, s_m)$ is nondecreasing in x and right continuous on $(0, 1]$, it follows from Helly's theorem [4, p. 291] that there exists a subsequence $\{G(\cdot, s_{m_k})\}$ and $G_\infty \in \Gamma$ such that for each x in $[0, 1]$ at which G_∞ is continuous, $\lim_{s_{m_k} \to \infty} G(x, s_{m_k}) = (G_\infty(x))$. Since K is a matrix of continuous functions, we now have from the Helly–Bray theorem [4, p. 282], for all x in I and $i = 1, \cdots, n$,

$$\begin{aligned} \lim_{s_{m_k} \to \infty} (WG)_i(x, s_{m_k}) &= \lim_{s_{m_k}} \sum_{j=1}^{n} \int_0^1 K_{ij}(x, y)\, dG_j(y, s_{m_k}) \\ &= \sum_{j=1}^{n} \int_0^1 K_{ij}(x, y)\, d(G_\infty)_j(y) \\ &= (WG_\infty)_i(x). \end{aligned} \tag{3.20}$$

As in the proof of Theorem 1.3–3, we have for $i = 1, \cdots, n$,

$$\int_0^1 \phi_i[(WG)(x,s)]\,dx = \int_0^1 \phi_i[(Wg)(x,s)]\,dx \to 0 \tag{3.21}$$

as $s \to \infty$. Consequently, by the Lebesgue dominated convergence theorem and (3.20), for $i = 1, \cdots, n$,

$$\lim_{s_{m_k} \to \infty} \int_0^1 \phi_i[(WG)(x, s_{m_k})]\,dx = \int_0^1 \phi_i[(WG_\infty)(x)]\,dx = 0. \tag{3.22}$$

Observe that the uniform continuity of each K_{ij} implies $(WG_\infty)(x)$ is continuous, and hence we conclude from (3.22) that each $\phi_i[(WG_\infty)(x)]$ is identically zero. In other words, each $(WG_\infty)_i(x) \leqq 0$ for all x in I. It now follows for any F in $\Lambda = \Gamma$ that

$$\int_0^1 dF(x) \cdot \int_0^1 K(x,y)\,dG_\infty(y) \leqq 0. \tag{3.23}$$

Since $\{\Delta = \Gamma, \Gamma, K\}$ is symmetric, G_∞ is an optimal strategy for both players.

4. Symmetrization of L^∞ and continuous games. We will now describe a method which transforms an arbitrary L^∞ or continuous game into a symmetric L^∞ or continuous game. The method will show how to compute the solution of the original game from the solution of the symmetrized game, and hence the dynamic model for the symmetrized game can be used to solve the original game. We will only prove the results for continuous games. In the case of L^∞ games, the proof is very similar in format but because of the "weaker" ε-optimal solution it is technically involved.

Consider any arbitrary continuous game $\{\Delta, \Gamma, K\}$. Let $k = \sup_{(i,j)} \sup_{(x,y)} \cdot |K_{ij}(x,y)|$ and $\alpha = k + \delta$, where δ is some positive real number. Define

$$A = \begin{bmatrix} 0 & K(x,y) + \alpha & -1 \\ -K^T(y,x) - \alpha & 0 & +1 \\ +1 & -1 & 0 \end{bmatrix}. \tag{4.1}$$

Here α also denotes the $n \times m$ matrix, where each component equals α. "$+1$" (or "-1") denote columns or rows of $+1$'s (or -1's) of the appropriate dimensions. "0" denotes the matrix of 0's with the appropriate dimensions. Let γ be the set of all R^{n+m+1} policy functions H. For any H in γ, denote the first n components by F, the next m components by G, and the last component by Λ. The symmetric continuous game $\{\gamma, \gamma, A\}$ is the *symmetrization* of the original game $\{\Delta, \Gamma, K\}$.

LEMMA 4.1. *Suppose $H = (F, G, \Lambda)$ is an optimal strategy for $\{\gamma, \gamma, A\}$. If $\sum_{i=1}^m G_i(1) > 0$, then $\Lambda(1) > 0$.*

Proof. Suppose H is an optimal strategy for player 2. Since the game $\{\gamma, \gamma, A\}$ is symmetric, the value is zero, and thus from (4.1), we observe that for all x in I,

$$\int_0^1 (K(x,y) + \alpha)_i\,dG(y) - \int_0^1 d\Lambda(y) \leqq 0, \qquad i = 1, \cdots, n, \tag{4.2}$$

where the subscript "i" denotes the ith row.

Assume $\sum_{i=1}^{m} G_i(1) = \mu > 0$. If $\Lambda(1) = 0$, then from (4.2) we have

$$\int_0^1 K_i(x, y)\, dG(y) \leqq -\alpha\mu, \qquad i = 1, \cdots, n, \tag{4.3}$$

where K_i is the ith row of K. Denote $(1/\mu)G$ by G^*. Since $\mu > 0$, $G^* \in \Gamma$ (i.e., is an R^m policy function). From (4.3) and the definition of α,

$$\int_0^1 K_i(x, y)\, dG^*(y) \leqq -\alpha = -k - \delta, \qquad i = 1, \cdots, n. \tag{4.4}$$

But (4.4) is impossible since $-k \leqq \int_0^1 K_i(x, y)\, dG^*(y)$. We conclude that $\Lambda(1) > 0$.

THEOREM 4.2. *Suppose $H = (F, G, \Lambda)$ is an optimal strategy for $\{\gamma, \gamma, A\}$. Then*

$$\sum_{i=1}^{n} F_i(1) = \sum_{j=1}^{m} G_j(1) = \mu > 0. \tag{4.5}$$

Furthermore, $F^* = (1/\mu)F$ and $G^* = (1/\mu)G$ are optimal strategies for the game $\{\Delta, \Gamma, K\}$, and the value of the game is

$$v = (1/\mu)\Lambda(1) - \alpha. \tag{4.6}$$

Proof. Denote $K^\alpha = K + \alpha$. Again since H is an optimal solution for the symmetric game $\{\gamma, \gamma, A\}$ and since the game has value zero, it follows that

$$\sum_{j=1}^{m} \int_0^1 K_{ij}^\alpha(x, y)\, dG_j(y) - \int_0^1 d\Lambda(y) \leqq 0, \qquad i = 1, \cdots, n, \tag{4.7}$$

$$\sum_{i=1}^{n} \int_0^1 -K_{ij}^\alpha(y, s)\, dF_i(y) + \int_0^1 d\Lambda(y) \leqq 0, \qquad j = 1, \cdots, m, \tag{4.8}$$

and

$$\sum_{i=1}^{n} \int_0^1 dF_i(y) - \sum_{j=1}^{m} \int_0^1 dG_j(y) \leqq 0. \tag{4.9}$$

Suppose $\sum_{i=1}^{n} \int_0^1 dG_j(y) = 0$. Then by (4.9), $\sum_{i=1}^{n} \int_0^1 dF_i(y) = 0$. But since the $F_i(\cdot)$ are monotone, nondecreasing, and zero at $y = 0$, each $F_i(\cdot)$ is identically zero. But then from (4.8), $\int_0^1 d\Lambda(y) = 0$, and thus,

$$\sum_{i=1}^{n} \int_0^1 dF_i(y) + \sum_{j=1}^{m} \int_0^1 dG_i(y) + \int_0^1 d\Lambda(y) = 0, \tag{4.10}$$

which contradicts the assumption that H is an R^{n+m+1} policy function. Therefore, $\sum_{i=1}^{m} \int_0^1 dG_j(y) = \mu > 0$ and by our preceding lemma, $\int_0^1 d\Lambda(y) = \lambda > 0$. Denote $(1/\mu)G$ by $G^* \in \Gamma$. Integrating (4.7) with respect to $F_i(x)$ and summing, we obtain

$$\int_0^1 dF(x) \cdot \int_0^1 K^\alpha(x, y)\, dG(y) - \lambda \sum_{i=1}^{n} \int_0^1 dF_i(y) \leqq 0. \tag{4.11}$$

Similarly integrating (4.8) with respect to $G_j(x)$ and summing yields

$$-\int_0^1 dG(x) \cdot \int_0^1 (K^\alpha)^T(y, x)\, dF(y) + \lambda \sum_{j=1}^{m} \int_0^1 dG_j(y) \leqq 0, \tag{4.12}$$

or changing the order of integration of (4.12), we have

$$-\int_0^1 dF(x)\cdot\int_0^1 K^{\alpha}(x,y)\,dG(y)+\lambda\sum_{j=1}^{m}\int_0^1 dG_j(y)\leqq 0. \tag{4.13}$$

Since $\lambda>0$, (4.11) and (4.13) imply

$$\begin{aligned}\sum_{j=1}^{m}\int_0^1 dG_j(y) &\leqq \frac{1}{\lambda}\int_0^1 dF(x)\cdot K^{\alpha}(x,y)\,dG(y)\\ &\leqq \sum_{j=1}^{n}\int_0^1 dF_i(y).\end{aligned} \tag{4.14}$$

Combining (4.14) and (4.9), we have

$$\sum_{i=1}^{n}\int_0^1 dF_i(y)=\sum_{j=1}^{m}\int_0^1 dG_j(y)=\mu>0. \tag{4.15}$$

Denote $(1/\mu)F=F^*$. Dividing both sides of (4.7) and (4.8) by μ and recalling the definitions of K^{α}, G^*, and F^*, we see that for all x in I,

$$\sum_{j=1}^{m}\int_0^1 K_{ij}(x,y)\,dG_j^*(y)\leqq\frac{\lambda}{\mu}-\alpha,\qquad i=1,\cdots,n, \tag{4.16}$$

and for all y in I,

$$\sum_{i=1}^{n}\int_0^1 K_{ij}(x,y)\,dF_i^*(x)\geqq\frac{\lambda}{\mu}-\alpha,\qquad j=1,\cdots,m. \tag{4.17}$$

(4.16) and (4.17) show that F^* and G^* are optimal strategies for players 1 and 2, respectively, and that the value of the game is $v=(\lambda/\mu)-\alpha$.

5. The prospects for applications of the theory to the numerical solution of games. In this paper we have shown how a solution to a symmetric L^{∞} or continuous game can be obtained as a limit from the dynamic model (1.3–2). To obtain a solution to an arbitrary L^{∞} or continuous game, we apply our symmetrization described in the previous section. An iterative procedure for numerically solving the dynamical equation could be based upon the contractive map used in § 2.3 to establish the existence of a unique solution. Alternatively, by viewing (1.3–2) as an ordinary differential in a Banach space, the equation could be solved using Runge–Kutta methods. The numerical analysis of such equations is a current area of active research.

We shall now demonstrate an advantage of maintaining the original function space setting. Consider the continuous game $\{\Gamma,\Gamma,K\}$, where K maps $[0,1]\times[0,1]$ into R. Suppose we approximate $\{\Gamma,\Gamma,K\}$ by the matrix game $\{S^n,S^n,A\}$, where A is given by $a_{ij}=K(i/n,j/n)$ $i,j=1,\cdots,n$, and S^n is the n-dimensional simplex of mixed strategies. It can be easily shown that the dynamical model (1.3–2) for $\{S^n,S^n,A\}$ becomes

$$\frac{du_i}{ds}(s)=\phi[Au(s)]-\sum_{i=1}^{n}\phi_i[(Au)(s)]u(s), \tag{5.1}$$

where $i = 1, \cdots, n$ and $u(s) \in S^n$ for all $s \geqq 0$. Equation (5.1) is precisely the system of differential equation used in [1] to study matrix games. Now suppose we approximate the integrals of the continuous kernel K in (1.3–2) by a first order integration method. If the mesh size for the integration is $1/n$, then defining $u_i(s) = ng(i/n, s)$, it can be shown that (1.3–2) also reduces to (5.1). Thus in the sense described above, a matrix game approximation of $\{\Gamma, \Gamma, K\}$ is equivalent to using first order integration methods in (1.3–2). By using higher order integration methods, we should be able to improve on the matrix approximation. It should be remarked that, in general, since (1.3–2) is nonlinear, explicit closed form solutions cannot be found. Even for simple kernels such as $K(x, y) = x - y$, to obtain an explicit solution requires lengthy analysis.

Also of interest in the numerical problem is the selection of the initial function for the dynamic model. Recall the only restriction was that it belongs to the strategy space. In the case of continuous games, an exact solution G_∞ to the symmetric game was obtained as a limit of sequence $\{G(\cdot, s_m)\}$ (see Theorem 1.3–4). Whether finding the appropriate sequence poses any difficulties when computing G_∞ is a practical problem which needs to be studied.

REFERENCES

[1] G. W. Brown and J. von Neumann, *Solutions of games by differential equations*. Contributions to the Theory of Games, H. W. Kuhn and A. W. Tucker, Princeton University Press, Princeton, N.J., pp. 73–79.

[2] Avner Friedmann, *Foundations of Modern Analysis*, Holt, Rinehart, and Winston, New York, 1970.

[3] D. Gale, H. W. Kuhn and A. W. Tucker, *On symmetric games*, Contributions to the Theory of Games, H. W. Kuhn and A. W. Tucker, Princeton University Press, Princeton, N.J., pp. 81–88.

[4] L. M. Graves, *The Theory of Functions of Real Variables*, McGraw-Hill, New York, 1956.

[5] H. W. Kuhn and A. W. Tucker, *Contributions to the Theory of Games*, Annals of Math. Studies, 24, Princeton University Press, Princeton, N.J., 1950.

[6] G. W. Owen, *Game Theory*, W. B. Saunders, Philadelphia, 1968.

[7] R. G. Underwood, *A functional differential equation approach to solving infinite games*, University of South Carolina Mathematics Tech. Rep. 90D05-1, Columbia, 1975.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

COMBINED PRIMAL–DUAL AND PENALTY METHODS FOR CONVEX PROGRAMMING*

BARRY W. KORT† AND DIMITRI P. BERTSEKAS‡

Abstract. In this paper we propose and analyze a class of combined primal–dual and penalty methods for constrained minimization which generalizes the method of multipliers. We provide a convergence and rate of convergence analysis for these methods for the case of a convex programming problem. We prove global convergence in the presence of both exact and inexact unconstrained minimization, and we show that the rate of convergence may be linear or superlinear with arbitrary Q-order of convergence depending on the problem at hand and the form of the penalty function employed.

1. Introduction. In 1968, Powell [19] and Hestenes [10] independently introduced a new algorithm for minimizing a nonlinear function subject to nonlinear equality constraints. Shortly thereafter, Haarhoff and Buys [9] produced the third independent proposal of the same idea. The name "method of multipliers" is due to Hestenes. The original papers offered limited interpretation or analysis. Since 1968, however, a great deal of research has been conducted on the subject (for a fairly complete account see [1], [12] and the survey papers [24], [25]), and by now the behavior and properties of the method are fairly well understood. In particular it has been shown analytically that the method is superior in several ways to ordinary penalty methods [1]–[5].

In this paper we propose a class of methods which generalizes the method of multipliers and contains as a special case the original method. These methods correspond to different members of a class of penalty functions that we introduce. The original method corresponds to a quadratic penalty function. There are several reasons why such more general methods merit consideration. One reason is that when applied to nonlinear programming problems for which Rockafellar's "quadratic growth condition" [23] is not satisfied, the original method of multipliers may fail to solve the problem. The reason for this is that the associated unconstrained minimization problems may fail to have a solution. This difficulty can be avoided by using a suitable penalty function from the class that we introduce. A second reason is that by choosing an appropriate penalty function within our class it is possible to obtain twice differentiable augmented Lagrangians. This feature, which may be helpful in applying superlinearly convergent methods for unconstrained minimization, is not present in the original method. A third reason is that different penalty functions among our class exhibit drastically different behavior with respect to their convergence rate. Their speed of convergence may be much faster or much slower than that of the original method depending on the penalty function employed. This perhaps surprising feature, which is not

* Received by the editors September 14, 1973, and in final revised form March 28, 1975. This work was conducted at Engineering–Economic Systems Dept., Stanford University, Bell Telephone Laboratories, Holmdel, New Jersey and Coordinated Science Laboratory, University of Illinois, Urbana, Illinois, and supported by Bell Telephone Laboratories, Holmdel, New Jersey, the National Science Foundation under Grant GK 32870, and Joint Services Electronics Program, Contract DAAB-07-C-0259.

† Bell Telephone Laboratories, Holmdel, New Jersey 07733.

‡ Department of Electrical Engineering and Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801.

present in ordinary penalty methods, raises the interesting question as to whether it is more efficient for certain types of problems to use methods from our class rather than the original method.

The present paper is an outgrowth of research which started in early 1972. Since then our results have been reported in several papers and reports [11]–[15]. With the exception of an unpublished report [14] and the thesis of the first author [12], which contains some additional results, this is the first time that detailed analysis and proofs are being provided. The main results reported here are the global convergence results of § 3 and the rate of convergence results of § 4. Global convergence of the (quadratic) method of multipliers for the case of a convex programming problem was also proved independently by Rockafellar [22]. An important difference between our convergence result for inexact minimization and the one of Rockafellar is that our stopping rule is computationally implementable. The convergence rate results of § 4 are new and have no counterpart in the existing literature with the exception of a convergence rate analysis of the quadratic method of multipliers [1], [2], which utilizes much stronger assumptions than those of this paper.

Since we are concerned mainly with the convex programming case, we have made extensive use of the theory of convex functions. The excellent book by Rockafellar [20] has provided the foundation for much of our analysis, and the reader will encounter frequent references to notions and theorems from [20]. It is thus inevitable that some familiarity with [20] is required on the part of the reader.

2. A generalized class of multiplier methods. We consider the following convex programming problem:

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \leqq 0, \qquad i = 1, \cdots, m, \end{aligned} \tag{1}$$

where we make the following assumptions which will be in effect throughout this paper:

A1. The function $f_0 : R^n \to (-\infty, +\infty]$ is an extended real-valued closed proper convex function on R^n [20], and the functions $f_i : R^n \to (-\infty, +\infty)$, $i = 1, \cdots, m$, are real-valued convex functions on R^n (R^n is n-dimensional Euclidean space).

A2. Problem (1) has a nonempty and compact solution set denoted by X^*, and a nonempty and compact set of Lagrange multiplier vectors (or Kuhn–Tucker vectors as defined in [20]) denoted by Y^*.

Notice that no differentiability assumptions are imposed on the functions f_i. Furthermore, set constraints of the form $x \in X \subset R^n$ may be incorporated into the objective function f_0 by defining $f_0(x) = +\infty$ for $x \notin X$. In order to simplify the exposition and avoid overburdening the notation we do not consider equality constraints. The presence of affine equality constraints does not alter the basic nature of our results. For the corresponding treatment together with the associated algorithms we refer to [11]–[15].

We now introduce the class P of two-parameter penalty functions $p : R^2 \to R$ satisfying the following properties:

(a) p is continuous on $R \times [0, +\infty)$, continuously differentiable on $R \times (0, +\infty)$ and possesses for all $t \in R$ the right partial derivative

$$\lim_{y \to 0^+} \frac{p(t; y) - p(t; 0)}{y}.$$

(The partial derivative with respect to the first argument is denoted by $\nabla_1 p(\cdot\,;\cdot)$ and the one with respect to the second argument by $\nabla_2 p(\cdot\,;\cdot)$.)

(b) $p(t;\cdot)$ is concave on R for each fixed $t \in R$.

(c) For each fixed $y \in R$, $p(\cdot\,; y)$ is convex on R and satisfies the following strict convexity condition:

$$\text{if (i) } t_0 > 0 \text{ and } y \geqq 0 \text{ or (ii) } \nabla_1 p(t_0; y) > 0, \text{ then}$$
$$p(t; y) - p(t_0; y) > (t - t_0)\nabla_1 p(t_0; y) \quad \forall t \neq t_0.$$

For all $y \in [0, +\infty)$,

(d) $p(0; y) = 0$,

(e) $\nabla_1 p(0; y) = y$,

(f) $\lim_{t \to -\infty} \nabla_1 p(t; y) = 0$,

(g) $\lim_{t \to +\infty} \nabla_1 p(t; y) = +\infty$,

(h) $\inf_{t \in R} p(t; y) > -\infty$.

Several properties of the functions in P which will be used later are given in Proposition A.1 of the Appendix. In Fig. 1 we show the shape of a typical function in P. Note that $p(\cdot\,; 0)$ is the type of function used in many exterior penalty methods. The predominant effect of the parameter y is to alter the slope as $p(\cdot\,; y)$ passes through the origin (properties d and e). For t near zero, $p(t; y) \approx yt$, but elsewhere the penalty effect dominates. The main consideration is that $p(\cdot\,; y)$ passes through the origin with slope y. As $t \to \infty$, $p(t; y)$ grows to infinity with unbounded slope. As $t \to -\infty$, $p(t; y)$ approaches or reaches a finite infimum which is less than or equal to zero.

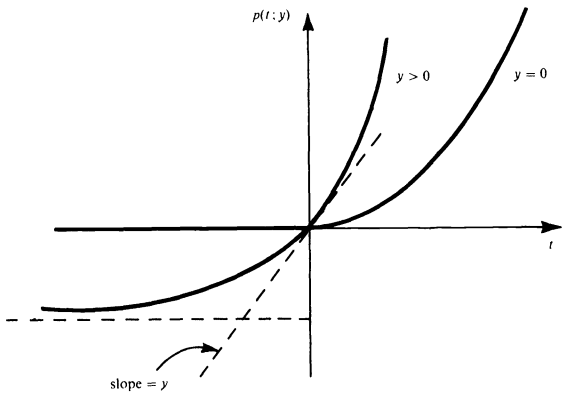


FIG. 1

It is useful to be able to explicitly control the severity of the penalty effect. (In pure penalty methods, that control is the essence.) For each $p \in P$ we can actually

generate a parametric family with a continuum on the severity of the penalty behavior. For any scalar $r > 0$ we define

$$p_r(t\,;y) \triangleq rp(t/r\,;y). \tag{2}$$

It is easy to verify that p_r also satisfies properties (a)–(h). Figure 2 shows the effect of the penalty parameter r. When r is large the penalty effect is small; as r approaches

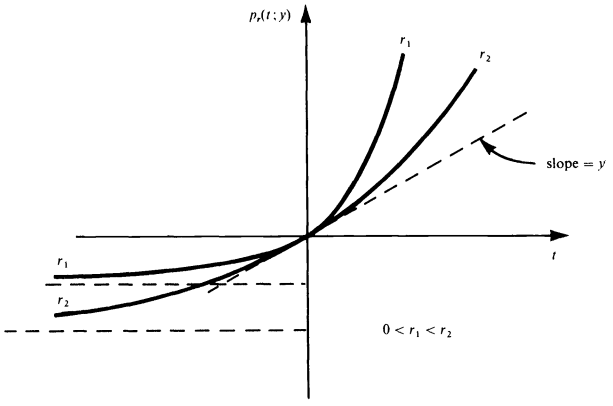


FIG. 2

zero, the penalty behavior becomes increasingly severe. Observe that for $y \in [0, +\infty)$

$$\lim_{r\to\infty} p_r(t\,;y) = yt$$

and

$$\lim_{r\downarrow 0} p_r(t\,;y) = \begin{cases} 0 & \text{if } t \leqq 0, \\ \infty & \text{if } t > 0. \end{cases}$$

Note also, as Fig. 2 illustrates, that

$$0 < r_1 < r_2 \Rightarrow p_{r_1}(t\,;y) \geqq p_{r_2}(t\,;y).$$

Example 1 (quadratic).

$$p(t\,;y) = \begin{cases} yt + \tfrac{1}{2}t^2, & t \geqq -y, \\ -\tfrac{1}{2}y^2, & t \leqq -y, \end{cases} \tag{3}$$

$$\nabla_1 p(t\,;y) = \max[0, y + t].$$

This function has been proposed and analyzed by Rockafellar [21]–[22], and also by Buys [8]. It undoubtedly is one of the best functions in the class P both in terms of its convergence and rate of convergence properties, as will be seen later. It grows quadratically as $t \to \infty$, and this fact makes it unsuitable for certain nonconvex problems in which Rockafellar's quadratic growth condition [23] is not satisfied. We attempted the solution of one such problem (the post office

parcel problem [7], [8]) using this function, and the algorithm failed. Another disadvantage is that the function (3) is not twice differentiable at $t = -y$, a fact which may adversely affect the performance of certain unconstrained minimization methods, particularly for small values of the penalty parameter r, and also for small values of Lagrange multipliers corresponding to active constraints.

Example 2 (ρ-order of growth, $\rho > 2$).

$$p(t;y) = \begin{cases} yt + \frac{1}{2}t^2 + t^\rho, & 0 \leqq t, \\ yt + \frac{1}{2}t^2, & -y \leqq t \leqq 0, \\ -\frac{1}{2}y^2, & t \leqq -y, \end{cases}$$

This function has properties similar to (3) but grows at a larger rate as $t \to \infty$, thus bypassing the shortcoming mentioned in connection with (3).

Example 3 (twice differentiable).

$$p(t;y) = \begin{cases} yt + yt^2 + \frac{1}{6}t^3, & t \geqq 0, \\ yt/(1-t), & t \leqq 0, \end{cases}$$

$$\nabla_1 p(t;y) = \begin{cases} y + 2yt + \frac{1}{2}t^2, & t \geqq 0, \\ y/(1-t)^2, & t \leqq 0, \end{cases}$$

$$\frac{\partial^2 p}{\partial t^2}(t;y) = \begin{cases} 2y + t, & t \geqq 0, \\ 2y/(1-t)^3, & t \leqq 0. \end{cases}$$

Example 4 (class P_E). This subclass of P is defined as the class of functions $p: R^2 \to R$ of the form

$$p(t;y) = \begin{cases} yt + \phi(t) & \text{if } y + \nabla\phi(t) \geqq 0, \\ \min\limits_{\tau \in R} \{y\tau + \phi(\tau)\} & \text{otherwise}, \end{cases} \tag{4}$$

where $\phi: R \to R$ is a function satisfying:

(a) ϕ is continuously differentiable and strictly convex on R,

(b) $\phi(0) = 0$, $\nabla\phi(0) = 0$,

(c) $\lim_{t \to -\infty} \nabla\phi(t) = -\infty$, $\lim_{t \to +\infty} \nabla\phi(t) = +\infty$.

The class of functions ϕ above is basic in the extension of our algorithms to equality constraints as discussed in [11]–[15]. Notice that the corresponding class P_E contains the quadratic function of Example 1 ($\phi(t) = \frac{1}{2}t^2$).

We mention that the class P may be further enlarged by the introduction of "barriers" (i.e., vertical asymptotes) in $p(\cdot\,; y)$. Then p would take on the value $+\infty$ outside the barrier and in condition g the limit would be taken as $t \to a$ where $a \in (0, +\infty]$. We refer the reader to [12] for a discussion of this modification.

The algorithms that we propose are based on exact or approximate unconstrained minimization of the *augmented Lagrangian*

$$\begin{aligned} L_r(x;y) &= f_0(x) + r \sum_{i=1}^{m} p[f_i(x)/r; y_i] \\ &= f_0(x) + \sum_{i=1}^{m} p_r[f_i(x); y_i] \end{aligned} \tag{5}$$

defined for each penalty parameter $r > 0$ and penalty function p from the class P.

ALGORITHM A (exact minimization). Select a penalty function p from the class P, a scalar $r^0 > 0$, and an initial estimate of the Lagrange multiplier vector $y^0 = (y_1^0, \cdots, y_m^0)$ with $y_i^0 \geqq 0$, $i = 1, \cdots, m$.

Step 1. Given y^k, r^k find an x^k solving the problem

$$\min_{x \in R^n} L_{r^k}(x; y^k).$$

Step 2. Set

$$y_i^{k+1} = \nabla_1 p[f_i(x^k)/r^k; y_i^k] \quad \text{for } i = 1, \cdots, m. \tag{6}$$

Stop if $y^{k+1} = y^k$. Otherwise select $r^{k+1} > 0$ and return to Step 1.

In practice, Step 1 in the above algorithm should be carried out only approximately. Not only is this necessary in order for the algorithm to be implementable, but in addition it usually results in substantial computational savings. We provide below an implementable version of the algorithm which employs inexact minimization. Let us denote by $s^k: R^n \to R^m$, $k = 0, 1, \cdots$, the functions given by

$$s_i^k(x) = \nabla_1 p[f_i(x)/r^k; y_i^k], \qquad i = 1, \cdots, m, \tag{7}$$

where $s^k = (s_1^k, \cdots, s_m^k)$.

Denote also by $\Delta_x L_r(x; y)$ the element of minimum Euclidean norm of the subdifferential (with respect to x) $\partial_x L_r(x; y)$ ([20, § 23]) of $L_r(x; y)$ for every x, y for which $\partial_x L_r(x; y)$ is nonempty. We have

$$\|\Delta_x L_r(x; y)\| = \min_{z \in \partial_x L_r(x; y)} \|z\|, \tag{8}$$

where $\|\cdot\|$ denotes the standard Euclidean norm. Note that $\Delta_x L_r(x; y)$ is just the ordinary gradient if $L_r(\cdot\,; y)$ is differentiable at x.

ALGORITHM B (inexact minimization). Select a penalty function p from P, scalars $r^0 > 0$, $\eta^0 \geqq 0$, and an initial estimate y^0 with $y_i^0 \geqq 0$, $i = 1, \cdots, m$.

Step 1. Given y^k, r^k, η^k, find an x^k satisfying

$$\|\Delta_x L_{r^k}(x^k; y^k)\|^2 \leqq \eta^k \sum_{i=1}^{m} \{s_i^k(x^k) f_i(x^k) - r^k p[f_i(x^k)/r^k; y_i^k]\}, \tag{9}$$

where s^k and $\Delta_x L_{r^k}$ are defined in (7), (8).

Step 2. Set

$$y_i^{k+1} = \nabla_1 p[f_i(x^k)/r^k; y_i^k] \quad \text{for } i = 1, \cdots, m. \tag{10}$$

Stop if $y^{k+1} = y^k$. Otherwise select $r^{k+1} > 0$, $\eta^{k+1} \geqq 0$ and return to Step 1.

It is easy to show (see Proposition A.1—Appendix) that the right-hand side of the stopping criterion (9) is nonnegative. Since ordinarily we take the sequence η^k to be decreasing and convergent to zero, the inexact minimization is asymptotically exact. Notice that Algorithm B above is equivalent to Algorithm A if $\eta^k = 0$ for all k, and that both algorithms generate points with $y_i^k \geqq 0, i = 1, \cdots, m$. When $\eta^k > 0$, and y^k is not a Lagrange multiplier vector we shall show in the next section that the stopping criterion (9) will yield for many cases of interest a vector x^k by means of a finite process. We will also show that Algorithm A is globally

convergent if r^k is bounded above while Algorithm B is globally convergent if, in addition, $\eta^k \to 0$ and a certain uniform convexity assumption is satisfied.

3. Convergence analysis. It is convenient to state and prove our results simultaneously for both Algorithms A and B. In this way we avoid duplication of arguments. We introduce the following uniform convexity assumption which is necessary for the results concerning Algorithm B.

A3. (For Algorithm B.) There exists a positive scalar μ such that for all $x', x \in R^n$, $x^* \in \partial f_0(x)$,

$$f_0(x') \geqq f_0(x) + \langle x^*, x' - x \rangle + \frac{\mu}{2}\|x' - x\|^2. \tag{11}$$

Every result for Algorithm A ($\eta^k = 0$ for all k) assumes A1, A2 while every result referring to Algorithm B ($\eta^k > 0$ for some k) assumes A1, A2 *and* A3.

Let us also introduce the *ordinary Lagrangian*

$$L(x; y) = \begin{cases} f_0(x) + \sum_{i=i}^{m} y_i f_i(x) & \text{if } y_i \geqq 0, \quad i = 1, \cdots, m, \\ -\infty & \text{otherwise}, \end{cases} \tag{12}$$

the *ordinary dual functional*

$$g(y) = \inf_x L(x, y), \tag{13}$$

and for $r > 0$, the "*penalized*" *dual functional*

$$g_r(y) = \inf_x L_r(x, y).$$

The function g_r above may be viewed as a dual functional corresponding to appropriate perturbations in the convex programming problem (1) [12]. It is easy to show that for every Lagrange multiplier vector y^* of problem (1) we have

$$g(y^*) = g_r(y^*) = \max_y g(y) = \max_y g_r(y) \tag{14}$$

and furthermore,

$$g(y^*) = g_r(y^*) = \min_x \{f_0(x) | f_i(x) \leqq 0, i = 1, \cdots, m\}.$$

The following proposition shows that Step 1 in both Algorithms A and B can be carried out.

PROPOSITION 1. *For any $r^k > 0$, $y^k \in R^m$, $y_i^k \geqq 0$, $i = 1, \cdots, m$, the set of points minimizing $L_{r^k}(\cdot\,; y^k)$ is nonempty and compact. Under* A3 *this set consists of a single point $\bar{x}^k$ Furthermore, if $\eta^k > 0$, y^k is not a Lagrange multiplier vector of problem* (1), *$\{z^j\}$ is a sequence with $z^j \to \bar{x}^k$ and $\Delta_x L_{r^k}(z^j; y^k) \to 0$, then there exists a vector $x^k \in \{z^1, z^2, \cdots\}$ satisfying the stopping criterion* (9).

Proof. Since for every $r > 0$ and $y \in R^m$, $y_i \geqq 0$, $i = 1, \cdots, m$, $L_r(\cdot\,; y)$ has no directions of recession (see Proposition A.2 of the Appendix) it follows that all the level sets of $L_r(\cdot\,; y)$ are compact and the minimum set is nonempty, which proves the first part. Now under A3 we have for all $x, x' \in R^n$ and $x^* \in \partial_x L_{r^k}(x; y^k)$,

$$L_{r^k}(x'; y^k) \geqq L_{r^k}(x; y^k) + \langle x^*, x' - x \rangle + \frac{\mu}{2}\|x' - x\|^2$$

from which uniqueness of the minimizing point $\bar{x}^k$ follows. If y^k is not a Lagrange multiplier, then it is easy to show using Proposition A.1 of the Appendix that

$$\lim_{j\to\infty} \sum_{i=1}^{m} \{s_i^k(z^j) f_i(z^j) - r^k p[f_i(z^j)/r^k; y_i^k]\}$$

$$= \sum_{i=1}^{m} \{s_i^k(\bar{x}^k) f_i(\bar{x}^k) - r^k p[f_i(\bar{x}^k)/r^k; y_i^k]\} > 0,$$

and the result follows using (9), $\eta^k > 0$ and the fact $\lim_{j\to\infty} \Delta_x L_{r^k}(z^j; y^k) = 0$.

Q.E.D.

We turn now to proving that the points y^k generated by Algorithms A, B (eventually) ascend the ordinary dual functional $g(\cdot)$ of (13). This fact leads to the interpretation of Algorithms A, B as primal-dual methods. Furthermore, this fact is helpful in proving that $\{y^k\}$, $\{x^k\}$ are bounded sequences. To avoid ambiguities we shall adopt the convention that if the algorithm stops at iteration $\bar{k}$, then $y^{k+1} = y^k$ for all $k \geqq \bar{k}$.

PROPOSITION 2. *If* $\{y^k\}$, $\{x^k\}$ *are sequences generated by Algorithm* A *or Algorithm* B, *then*:

(a) *For Algorithm* A *and all* k,

$$g(y^k) \leqq g_{r^k}(y^k) \leqq g(y^{k+1}) \tag{15}$$

with strict inequality if $y^k \neq y^{k+1}$.

(b) *For Algorithm* B *and all* k *such that* $\eta^k < 2\mu$,

$$g(y^k) \leqq L(x^k; y^k) \leqq L_{r^k}(x^k; y^k) \leqq g(y^{k+1}) \tag{16}$$

with strict inequality

$$L(x^k; y^k) < L_{r^k}(x^k; y^k) < g(y^{k+1})$$

if $y^k \neq y^{k+1}$.

Proof. We shall prove part (b). A similar (but simpler) argument proves part (a). It is easy to show that

$$\Delta_x L_{r^k}(x^k; y^k) = \Delta_x L(x^k; y^{k+1}). \tag{17}$$

From A3 we have for all x, $x^* \in \partial_x L(x^k; y^{k+1})$

$$L(x; y^{k+1}) \geqq L(x^k; y^{k+1}) + \langle x^*, x - x^k \rangle + \frac{\mu}{2}\|x - x^k\|^2, \tag{18}$$

from which, taking infima with respect to x, we find that

$$g(y^{k+1}) \geqq L(x^k; y^{k+1}) - \frac{1}{2\mu}\|\Delta_x L(x^k; y^{k+1})\|^2. \tag{19}$$

The stopping rule (9) may also be written as

$$\|\Delta_x L_{r^k}(x^k; y^k)\|^2 \leqq \eta^k \{L(x^k; y^{k+1}) - L_{r^k}(x^k; y^k)\}. \tag{20}$$

Combining (17), (19), (20), we obtain

$$
\begin{aligned}
L(x^k; y^{k+1}) - g(y^{k+1}) &\leqq \frac{1}{2\mu}\|\Delta_x L(x^k; y^{k+1})\|^2 \\
&= \frac{1}{2\mu}\|\Delta_x L_{r^k}(x^k; y^k)\|^2 \\
&\leqq \frac{\eta^k}{2\mu}\{L(x^k; y^{k+1}) - L_{r^k}(x^k; y^k)\} \\
&\leqq L(x^k; y^{k+1}) - L_{r^k}(x^k; y^k),
\end{aligned} \tag{21}
$$

from which

$$L_{r^k}(x^k; y^k) \leqq g(y^{k+1}).$$

The remaining inequalities in (16) follow from the definition (13) and the fact $p(t; y) \geqq yt \ \forall y, t \in R$. The strict inequality part of the proposition follows from the properties of the penalty function p (see Appendix—Proposition A.1). Q.E.D.

COROLLARY 2.1. *A sequence $\{y^k\}$ generated by either Algorithm* A *or Algorithm* B *is bounded, provided (in the case of Algorithm* B*) that there exists a $\bar{k} \geqq 0$ such that $\eta^k < 2\mu$ for all $k \geqq \bar{k}$.*

Proof. In either case, by Proposition 2, y^k belongs to the level set $\{y | g(y) \geqq g(y^{\bar{k}})\}$ for all $k \geqq \bar{k}$. But this set is compact since the set of maximizing points Y^* of g is compact by A2 ([20, Corollary 8.7.1]). Q.E.D.

PROPOSITION 3. *A sequence $\{x^k\}$ generated by either Algorithm* A *or Algorithm* B *is bounded, provided that there exists an $\bar{r} > 0$ such that $0 < r^k \leqq \bar{r}$ for all k, and (in the case of Algorithm* B*) that there exists a $\bar{k} \geqq 0$ such that $\eta^k < 2\mu$ for all $k \geqq \bar{k}$.*

Proof. We show that for $k \geqq \bar{k}$ the vectors x^k belong to a level set of a certain closed, proper, convex function. This convex function has no directions of recession and hence its level sets are compact. Using Corollary 2.1, let M be an upper bound for $\{y_i^k\}$, i.e., $0 \leqq y_i^k \leqq M$, for all i, k. Now we have from properties of p that

$$
\begin{aligned}
p_{r^k}(t; y_i^k) &\geqq p_{r^k}(t; 0) \geqq p_{\bar{r}}(t; 0) \quad \forall t \geqq 0, \\
p_{r^k}(t; y_i^k) &\geqq p_{r^k}(t; M) \geqq p_{\bar{r}}(t; M) \geqq \inf_{\tau} p_{\bar{r}}(\tau; M) \quad \forall t < 0.
\end{aligned}
$$

Using the fact $p_{\bar{r}}(t; 0) = 0, \forall t < 0$ and $\inf_\tau p_{\bar{r}}(\tau, M) \leqq 0$, we have for all i, k,

$$p_{r^k}(t; y_i^k) \geqq p_{\bar{r}}(t; 0) + \inf_{\tau} p_{\bar{r}}(\tau; M) \quad \forall t \in R.$$

Hence

$$
\begin{aligned}
L_{r^k}(x; y^k) &\geqq f_0(x) + \sum_{i=1}^{m} \{p_{\bar{r}}[f_i(x); 0] + \inf_{\tau} p_{\bar{r}}(\tau; M)\} \\
&= L_{\bar{r}}(x; 0) + m \inf_{\tau} p_{\bar{r}}(\tau; M) \quad \forall x \in R^n.
\end{aligned}
$$

Now the function $L_{\bar{r}}(x; 0)$ has no directions of recession (Appendix—Proposition A.2) and hence has bounded level sets. Furthermore, for $k \geqq \bar{k}$ by Proposition 2,

we have that x^k belongs to the level set

$$\{x | L_{\bar{r}}(x; 0) \leqq \max_y g(y) - m \inf_\tau p_{\bar{r}}(\tau; M)\}.$$

Hence $\{x^k\}$ is bounded. Q.E.D.

We can now state our main convergence result.

PROPOSITION 4. *Every limit point of a sequence* $\{(x^k; y^k)\}$ *generated by either Algorithm* A *or Algorithm* B *is an optimal solution–Lagrange multiplier pair for problem* (1) *provided that for some* $\bar{r} > 0$, $0 < r^k \leqq \bar{r}$ *for all* k *and (in the case of Algorithm* B*) there exists a* $\bar{k} \geqq 0$ *such that* $\eta^k < 2\mu$ *for all* $k \geqq \bar{k}$. *Furthermore, at least one such limit point exists.*

Proof. By Corollary 2.1 and Proposition 3 there exists a limit point $(\bar{x}; \bar{y})$. Let $\{(x^k; y^k)\}_{k \in K}$ be a subsequence with $\lim_{k \to \infty} \{(x^k; y^k)\}_{k \in K} = (\bar{x}; \bar{y})$. We first show that the point $\bar{x}$ is feasible. Indeed, if $f_i(\bar{x}) > 0$ for some i, then for some $\delta > 0$, $f_i(x^k) > \delta > 0$ for all $k \in K$, $k \geqq \tilde{k}$, where $\tilde{k}$ is sufficiently large. We then have by Proposition A.1 of the Appendix, for all $k \geqq \tilde{k}$, $k \in K$,

$$\begin{aligned} L_{r^k}(x^k; y^k) - L(x^k; y^k) &\geqq p_{r^k}[f_i(x^k); y_i^k] - y_i^k f_i(x^k) \\ &\geqq p_{r^k}[f_i(x^k); 0] \geqq p_{\bar{r}}(\delta, 0) > 0. \end{aligned}$$

But by the ascent property (16),

$$\lim_{k \to \infty} \{L_{r^k}(x^k; y^k) - L(x^k; y^k)\} = 0,$$

which contradicts the previous inequality. Hence $\bar{x}$ is feasible. Also from the ascent property we have

$$p_{\bar{r}}[f_i(\bar{x}); \bar{y}_i] = \bar{y}_i f_i(\bar{x}), \qquad i = 1, \cdots, m,$$

which by the properties of p (Proposition A.1) implies

$$p_{\bar{r}}[f_i(\bar{x}); \bar{y}_i] = \bar{y}_i f_i(\bar{x}) = 0, \qquad i = 1, \cdots, m. \tag{22}$$

Now by the lower semicontinuity of f_0, the ascent property (16) and (22),

$$\begin{aligned} \max_y g(y) &\geqq \lim_{\substack{k \to \infty \\ k \in K}} L(x^k; y^k) = \lim_{\substack{k \to \infty \\ k \in K}} \left\{ f_0(x^k) + \sum_{i=1}^m y_i^k f_i(x^k) \right\} \\ &\geqq f_0(\bar{x}) + \sum_{i=1}^m \bar{y}_i f_i(\bar{x}) = f_0(\bar{x}). \end{aligned}$$

Since $\bar{x}$ is also feasible, it follows that equality holds throughout in the above inequality. Hence, $\bar{x}$ is optimal, and in view of (22) and the fact $\bar{y}_i \geqq 0, i = 1, \cdots, m$, the vector $\bar{y}$ is a Lagrange multiplier. Q.E.D.

4. Rate of convergence analysis. This section considers the rate at which the sequence $\{y^k\}$ of dual variables converges to the set Y^* of Lagrange multipliers of problem (1). We examine the convergence of y^k to the set Y^* in terms of the distance

$$\|y^k - Y^*\| \triangleq \min_{y^* \in Y^*} \|y^k - y^*\|.$$

Specifically, assuming $\{y^k\}$ does not converge finitely (i.e., $y^k \notin Y^* \forall k$), we wish to estimate

$$\beta = \limsup_{k\to\infty} \frac{\|y^{k+1} - Y^*\|}{\|y^k - Y^*\|}.$$

If $0 < \beta < 1$, we say that the convergence is linear with ratio β. If $1 \leqq \beta < \infty$, the convergence is order 1 but not linear. If $\beta = 0$, the convergence is superlinear; we then consider the set of scalars $\alpha \geqq 1$ such that

$$\limsup_{k\to\infty} \frac{\|y^{k+1} - Y^*\|}{\|y^k - Y^*\|^\alpha} < \infty.$$

The supremum of the set of such α is called the (Q-) order of convergence [17].

We introduce the following new assumptions which are in addition to A1, A2 and A3 (the last of which is again in effect only for the results relating to Algorithm B). We also assume throughout that for some $\bar{r} > 0$, $0 < r^k \leqq \bar{r}$ for all k, and that for some $\bar{k}$ (when considering Algorithm B) $0 \leqq \eta^k < 2\mu$ for all $k \geqq \bar{k}$. In this way, convergence of $\{y^k\}$ to Y^* is guaranteed by Proposition 4.

A4. $p \in P_E$, where the class P_E is defined in Example 4 of § 2.

A5. There exist scalars $M_2 \geqq M_1 > 0$ and $\rho > 1$ such that for some open interval N_0 containing zero

$$M_1|t|^{\rho-1} \leqq \left|\frac{d\phi(t)}{dt}\right| \leqq M_2|t|^{\rho-1}, \qquad \forall\, t \in N_0, \tag{23}$$

where ϕ corresponds to p as in Example 4 of § 2.

A6. There is a neighborhood $B(Y^*; \delta)$ of Y^*, a scalar $\gamma > 0$ and a scalar $q > 1$ such that the concave dual functional g satisfies

$$g(y) - \sup_{y'} g(y') \leqq -\gamma\|y - Y^*\|^q, \qquad \forall\, y \in B(Y^*; \delta). \tag{24}$$

Assumptions A4, A5 *and* A6 *will be assumed to hold throughout this section.* On occasion we will use the following assumption, in which case we will explicitly so state. This assumption is a special case of A5 with $\rho = 2$ and A6 with $q = 2$.

A7. ϕ is twice differentiable on N_0 and $d^2\phi(0)/dt^2 = 1$. (In view of the role of the penalty parameter r, there is no loss of generality in assuming that $d^2\phi(0)/dt^2 = 1$ rather than $d^2\phi(0)/dt^2 > 0$.) Furthermore, $q = 2$ in A6.

Assumption A5 may be explained as a growth assumption on ϕ. Roughly speaking, it states that in a neighborhood of zero, $\phi(t)$ behaves like $|t|^\rho$ for some $\rho > 1$. Similarly, A6 is a growth assumption on the dual g. It says that in a neighborhood of the maximum set Y^*, $g(y)$ grows (downward) at least as fast as $\gamma\|y - Y^*\|^q$. (This assumption is much weaker than typical regularity assumptions which require g to be twice differentiable with negative definite Hessian at a unique maximum. In fact, A6 does not require once differentiability of g or even finiteness of g over the neighborhood $B(Y^*; \delta)$.)

We first introduce some notation and conventions in the following remarks R.1–R.4, and subsequently we prove a few lemmas which set the stage for the proof of the main propositions.

R.1. For each $y \in R^m$, we denote by $\hat{y}$ the unique projection of y on Y^*

$$\|y - \hat{y}\| = \|y - Y^*\| = \min_{y^* \in Y^*} \|y - y^*\|. \tag{25}$$

R.2. When considering results related to Algorithm B we use the notation

$$v^k = \frac{\eta^k}{2\mu}, \qquad k = 0, 1, \cdots. \tag{26}$$

To simplify statements of results, we assume (without essential loss of generality) that for some $\bar{v} > 0$, we have

$$0 \leqq v^k \leqq \bar{v} < 1, \qquad k = 0, 1, \cdots. \tag{27}$$

The results of all lemmas and propositions below where v^k and $\bar{v}$ appear, hold also with $v^k = \bar{v} = 0$ for the case of Algorithm A.

R.3. By A4, p has the form

$$p(t; y) = \begin{cases} yt + \phi(t), & y + \nabla\phi(t) \geqq 0, \\ \min\limits_{\tau} \{y\tau + \phi(\tau)\}, & y + \nabla\phi(t) < 0, \end{cases} \tag{28}$$

where $\phi : R \to R$ is a function satisfying the properties of Example 4 in § 2. Let us consider the conjugate convex function of $p(\cdot\,; y_i)$ for each $y_i \geqq 0$;

$$p^*(s_i; y_i) = \sup_{t_i} \{s_i t_i - p(t_i; y_i)\}, \qquad y_i \geqq 0. \tag{29}$$

It is easy to show that

$$p_r^*(s_i; y_i) = \begin{cases} \phi_r^*(s_i - y_i) = r\phi^*(s_i - y_i), & s_i, y_i \geqq 0, \\ +\infty, & s_i < 0, \end{cases} \tag{30}$$

where $\phi_r(t) \triangleq r\phi(t/r)$.

In (30) ϕ_r^* is the convex conjugate of ϕ_r, and ϕ^* is the convex conjugate of ϕ. Both ϕ^* and ϕ_r^* are real-valued convex functions. We denote for all $t = (t_1, \cdots, t_m)$, $y = (y_1, \cdots, y_m)$ with $y_i \geqq 0$, $i = 1, \cdots, m$, and $r > 0$,

$$h_r[t; y] = \sum_{i=1}^{m} p_r(t_i; y_i). \tag{31}$$

Then the conjugate convex function of h_r with respect to t is given for each y with $y_i \geqq 0$, $i = 1, \cdots, m$ by

$$\begin{aligned} h_r^*[s; y] &= \sum_{i=1}^{m} p_r^*(s_i; y_i) \\ &= \begin{cases} \sum\limits_{i=1}^{m} \phi_r^*(s_i - y_i) = r \sum\limits_{i=1}^{m} \phi^*(s_i - y_i), & s_i \geqq 0, \quad i = 1, \cdots, m, \\ +\infty, & \text{otherwise.} \end{cases} \end{aligned} \tag{32}$$

R.4. In all the results that follow, $\{y^k\}, \{x^k\}$ are assumed to be sequences generated by either Algorithm A or Algorithm B. We denote by u^k the m-vector[1]

$$u^k = \nabla_1 h_{r^k}^*[y^{k+1}; y^k], \tag{33}$$

where $\nabla_1 h_{r^k}^*$ denotes the vector of right partial derivatives of $h_{r^k}^*$ with respect to the first argument. Notice that these derivatives exist by (32) since ϕ^* is everywhere finite and differentiable by virtue of the strict convexity of ϕ and the fact $\lim_{t \to \pm\infty} (d\phi(t)/dt) = \pm\infty$. Notice that by (32), (33) and well-known facts on conjugacy [20], we have

$$y_i^{k+1} - y_i^k = \nabla\phi(u_i^k/r^k), \qquad i = 1, \cdots, m. \tag{34}$$

LEMMA 1. *For all y with $y_i \geqq 0$, $i = 1, \cdots, m$, $r > 0$ and x, we have*

$$L_r(x; y) = \max_s \{L(x; s) - h_r^*[s; y]\}. \tag{35}$$

Furthermore, for all k,

$$L_{r^k}(x^k; y^k) = L(x^k; y^{k+1}) - h_{r^k}^*[y^{k+1}; y^k], \tag{36}$$

$$u^k \in \partial_y L(x^k; y^{k+1}), \tag{37}$$

where u^k, $h_{r^k}^*[y^{k+1}; y^k]$ are given by (32), (33).

Proof. By writing (35) as

$$f_0(x) + \sum_{i=1}^m p_r[f_i(x); y_i] = \max_s \{f_0(x) + \sum_{i=1}^m s_i f_i(x) - h_r^*[s; y]\}, \tag{38}$$

its validity becomes evident via (29), (32). Relations (36), (37) follow from the fact $y_i^{k+1} = \nabla_1 p_{r^k}[f_i(x^k); y_i^k]$, $i = 1, \cdots, m$, (38), (33) and standard conjugacy results. Q.E.D.

LEMMA 2. *For all k sufficiently large, one has*

$$0 \leqq \gamma \|y^{k+1} - Y^*\|^q \leqq h_{r^k}^*[\hat{y}^k; y^k] - (1 - v^k) h_{r^k}^*[y^{k+1}; y^k], \tag{39}$$

where $\hat{y}^k$ is the projection of y^k on Y^* as in (25). *In addition, $\|y^{k+1} - y^k\| \to 0$.*

Proof. Using Lemma 1, we find that

$$\begin{aligned} L_{r^k}(x^k; y^k) &= L(x^k; y^{k+1}) - h_{r^k}^*[y^{k+1}; y^k] \\ &\geqq L(x^k; \hat{y}^k) - h_{r^k}^*[\hat{y}^k; y^k] \geqq \sup_y g(y) - h_{r^k}^*[\hat{y}^k; y^k]. \end{aligned} \tag{40}$$

By (36), (20), the stopping rule (9) may be written as

$$\|\Delta_x L(x^k; y^{k+1})\|^2 \leqq \eta^k h_{r^k}^*[y^{k+1}; y^k], \tag{41}$$

and by using (21) we have

$$\begin{aligned} L(x^k; y^{k+1}) &\leqq g(y^{k+1}) + \frac{1}{2\mu}\|\Delta_x L(x^k; y^{k+1})\|^2 \\ &\leqq g(y^{k+1}) + v^k h_{r^k}^*[y^{k+1}; y^k]. \end{aligned} \tag{42}$$

[1] Equivalently, $u_i^k = \max\{f_i(x^k), \tau_i^k\}$, $i = 1, \cdots, m$, where $\tau_i^k = \arg\min_\tau \{y_i^k \tau + \phi_{r^k}(\tau)\}$. (See [12, § 7]).

Combining (40), (42) and (24) (which holds for sufficiently large k), we obtain

$$\sup_y g(y) - h^*_{r^k}[\hat{y}^k; y^k] + h^*_{r^k}[y^{k+1}; y^k] \leqq L(x^k; y^{k+1})$$

$$\leqq \sup_y g(y) - \gamma \|y^{k+1} - Y^*\|^q + v^k h^*_{r^k}[y^{k+1}; y^k]$$

from which (39) follows. Using (39), (32), it follows that

$$(1 - v^k) \sum_{i=1}^m \phi^*(y_i^{k+1} - y_i^k) \leqq \sum_{i=1}^m \phi^*(\hat{y}_i^k - y_i^k).$$

Now v^k is bounded away from unity by assumption (27) and $\|\hat{y}^k - y^k\| \to 0$, so by the properties of ϕ^*, $y_i^{k+1} - y_i^k \to 0$, $i = 1, \cdots, m$. Q.E.D.

LEMMA 3. *There exists a scalar M_0 such that for all k sufficiently large,*

$$\|y^{k+1} - y^k\| \leqq M_0 \|y^k - Y^*\|.$$

Proof. From A5

$$M_1 |t|^{\rho - 1} \leqq \left| \frac{d\phi(t)}{dt} \right| \leqq M_2 |t|^{\rho - 1}, \qquad t \in N_0$$

and by integration

$$\frac{M_1}{\rho} |t|^\rho \leqq \phi(t) \leqq \frac{M_2}{\rho} |t|^\rho, \qquad t \in N_0.$$

Hence for any scalar s,

$$\sup_{t \in N_0} \left\{ st - \frac{M_1}{\rho} |t|^\rho \right\} \geqq \sup_{t \in N_0} \{st - \phi(t)\} \geqq \sup_{t \in N_0} \left\{ st - \frac{M_2}{\rho} |t|^\rho \right\}.$$

Let $[-a, a] \subset N_0$, $a > 0$. Then if $|s| \leqq M_1 a^{\rho - 1}$, the suprema are attained, and by the definition of the conjugate convex function we obtain

$$\frac{1}{\sigma M_2^{\sigma - 1}} |s|^\sigma \leqq \phi^*(s) \leqq \frac{1}{\sigma M_1^{\sigma - 1}} |s|^\sigma, \qquad |s| \leqq M_1 a^{\rho - 1}, \tag{43}$$

where σ is the conjugate exponent of ρ:

$$\frac{1}{\sigma} + \frac{1}{\rho} = 1, \qquad \sigma = \frac{\rho}{\rho - 1}.$$

Since $\|y^{k+1} - y^k\| \to 0$ (by Lemma 2) and $\|y^k - \hat{y}^k\| \to 0$, for k sufficiently large, $|y_i^{k+1} - y_i^k|$ and $|y_i^k - \hat{y}_i^k|$ are less than $M_1 a^{\rho - 1}$. Now apply Lemma 2, (32), (41). We have:

$$\frac{1}{\sigma} \frac{r^k}{M_2^{\sigma - 1}} \sum_{i=1}^m |y_i^{k+1} - y_i^k|^\sigma \leqq r^k \sum_{i=1}^m \phi^*(y_i^{k+1} - y_i^k)$$

$$= h^*_{r^k}[y^{k+1}; y^k] \leqq \frac{1}{1 - v^k} h^*_{r^k}[\hat{y}^k; y^k]$$

$$\leqq \frac{r^k}{(1 - v^k)\sigma M_1^{\sigma - 1}} \sum_{i=1}^m |\hat{y}_i^k - y_i^k|^\sigma.$$

Hence

$$\|y^{k+1} - y^k\|_\sigma \leqq (M_2/M_1)^{1/\rho}(1 - v^k)^{(1-\rho)/\rho}\|\hat{y}^k - y^k\|_\sigma,$$

where $\|\cdot\|_\sigma$ denotes the l_σ-norm.

Passing to the l_2-norm via the topological equivalence theorem for all norms on R^n, we obtain for k sufficiently large and for some $M_0 > 0$,

$$\|y^{k+1} - y^k\| \leqq M_0\|\hat{y}^k - y^k\|.$$

(In fact the best M_0 that we can obtain by the above analysis can be calculated to be $m^{|1/2-1/\rho|}(M_2/M_1)^{1/\rho}(1 - \bar{v})^{1/\rho-1}$.) Q.E.D.

LEMMA 4. *For all k sufficiently large,*

$$\begin{aligned} M_1\|u^k/r^k\|^{\rho-1} &\leqq \|y^{k+1} - y^k\| \\ &\leqq m^{1-\rho/2}M_2\|u^k/r^k\|^{\rho-1} \quad \text{if } \rho \leqq 2, \end{aligned} \tag{44}$$

and

$$\begin{aligned} m^{1-\rho/2}M_1\|u^k/r^k\|^{\rho-1} &\leqq \|y^{k+1} - y^k\| \\ &\leqq M_2\|u^k/r^k\|^{\rho-1} \quad \text{if } \rho \geqq 2, \end{aligned} \tag{45}$$

where u^k is given by (33).

Proof. By (34) and since $y_i^{k+1} - y_i^k \to 0$, we obtain $\nabla\phi(u_i^k/r^k) \to 0$. It follows by the continuity and strict monotonicity of $\nabla\phi$ that $u_i^k/r^k \to 0$. Hence $u_i^k/r^k \in N_0$ for k sufficiently large. Applying A5, we have for k sufficiently large,

$$\begin{aligned} M_1|u_i^k/r^k|^{\rho-1} \leqq |\nabla\phi(u_i^k/r^k)| = |y_i^{k+1} - y_i^k| \leqq M_2|u_i^k/r^k|^{\rho-1}, \\ M_1^2\sum_{i=1}^m |u_i^k/r^k|^{2(\rho-1)} \leqq \|y^{k+1} - y^k\|^2 \leqq M_2^2\sum_{i=1}^m |u_i^k/r^k|^{2(\rho-1)}. \end{aligned} \tag{46}$$

Now it is easy to prove that if $0 < \rho - 1 \leqq 1$,

$$\left[\sum_{i=1}^m |u_i^k/r^k|^2\right]^{\rho-1} \leqq \sum_{i=1}^m |u_i^k/r^k|^{2(\rho-1)} \leqq m^{2-\rho}\left[\sum_{i=1}^m |u_i^k/r^k|^2\right]^{\rho-1}.$$

Hence (46) yields

$$M_1^2\|u^k/r^k\|^{2(\rho-1)} \leqq \|y^{k+1} - y^k\|^2 \leqq M_2^2 m^{2-\rho}\|u^k/r^k\|^{2(\rho-1)}.$$

Taking square roots, (44) follows. If $1 \leqq \rho - 1$, we have

$$m^{2-\rho}\left[\sum_{i=1}^m |u_i^k/r^k|^2\right]^{\rho-1} \leqq \sum_{i=1}^m |u_i^k/r^k|^{2(\rho-1)} \leqq \left[\sum_{i=1}^m |u_i^k/r^k|^2\right]^{\rho-1}$$

and (46) yields

$$m^{2-\rho}M_1^2\|u^k/r^k\|^{2(\rho-1)} \leqq \|y^{k+1} - y^k\|^2 \leqq M_2^2\|u^k/r^k\|^{2(\rho-1)}.$$

Again taking square roots, we obtain (45). Q.E.D.

LEMMA 5. *For all k sufficiently large and all* $y^* \in Y^*$,

$$\|u^k\| \cdot \|y^{k+1} - Y^*\| \geqq \gamma \|y^{k+1} - Y^*\|^q - \nu^k h^*_{r^k}[y^{k+1}; y^k], \tag{47}$$

$$\begin{aligned} &\|y^{k+1} - y^* - u^k/r^k\| \cdot \|y^{k+1} - Y^*\| \\ &\quad \geqq \|y^{k+1} - Y^*\|^2 + \frac{\gamma}{r^k}\|y^{k+1} - Y^*\|^q - \frac{\nu^k}{r^k} h^*_{r^k}[y^{k+1}; y^k]. \end{aligned} \tag{48}$$

Proof. Take k sufficiently large so that $y^{k+1} \in B(Y^*; \delta)$, where $B(Y^*; \delta)$ is the neighborhood defined in A6. Then by assumption A6, (19), (37), (17), (41) and (26), we have

$$\begin{aligned} \gamma\|y^{k+1} - Y^*\|^q &\leqq \sup g - g(y^{k+1}) \\ &\leqq L(x^k; \hat{y}^{k+1}) - L(x^k; y^{k+1}) + \frac{1}{2\mu}\|\Delta_x L(x^k; y^{k+1})\|^2, \\ &\leqq \langle u^k, \hat{y}^{k+1} - y^{k+1}\rangle + \frac{1}{2\mu}\|\Delta_x L_{r^k}(x^k; y^k)\|^2 \\ &\leqq \|u^k\| \cdot \|Y^* - y^{k+1}\| + \nu^k h^*_{r^k}[y^{k+1}; y^k], \end{aligned}$$

from which (47) follows. The inequality above yields also, for any $y^* \in Y^*$,

$$\begin{aligned} &\langle y^{k+1} - y^*, y^{k+1} - \hat{y}^{k+1}\rangle + \frac{\gamma}{r^k}\|y^{k+1} - Y^*\|^q \\ &\qquad \leqq \langle y^{k+1} - y^* - u^k/r^k, y^{k+1} - \hat{y}^{k+1}\rangle + \frac{\nu^k}{r^k} h^*_{r^k}[y^{k+1}; y^k]. \end{aligned}$$

Using in the above relation the fact

$$\langle y^{k+1} - y^*, y^{k+1} - \hat{y}^{k+1}\rangle \geqq \|y^{k+1} - Y^*\|^2,$$

we find that relation (48) follows. Q.E.D.

Convergence rate of Algorithm A (*exact minimization*).

PROPOSITION 5. *Let* $(\rho - 1)(q - 1) \leqq 1$. *If* $\{y^k\}$ *is generated by Algorithm* A, *the* (*Q*-) *order of convergence of* $\|y^k - Y^*\|$ *is at least* $1/[(\rho - 1)(q - 1)]$.

Proof. Apply Lemmas 3 and 4 together with (47) (with $\nu^k = 0$). For sufficiently large k,

$$\begin{aligned} M_0\|y^k - Y^*\| &\geqq \|y^{k+1} - y^k\| \geqq M_1'\|u^k/r^k\|^{\rho - 1} \\ &\geqq M_1'\left[\frac{\gamma}{r^k}\|y^{k+1} - Y^*\|^{q-1}\right]^{\rho - 1}, \end{aligned}$$

where $M_1' = M_1 \min\{1, m^{1-\rho/2}\}$. Hence

$$\limsup_{k\to\infty} \frac{\|y^{k+1} - Y^*\|}{\|y^k - Y^*\|^\alpha} \leqq (M_0/M_1')^\alpha(\hat{r}/\gamma)^{1/(q-1)} < \infty,$$

where $\alpha = 1/[(\rho - 1)(q - 1)]$ and $\hat{r} = \limsup_{k\to\infty} r^k$. Q.E.D.

Note that $1 < 1/[(\rho - 1)(q - 1)] < \infty$ when $1 < \rho < 1 + 1/(q - 1)$. Hence for a given g, *any* order of convergence is obtainable by appropriate selection of

$p \in P_E$. Notice also that for $1 < q < 2$ the quadratic penalty function ($\rho = 2$) yields superlinear convergence rate and that as $q \to 1$, the order tends to infinity. This is consistent with a result of [6] which shows that for polyhedral convex programs where q can be taken as close to one as desired, the quadratic multiplier method converges in a finite number of iterations (i.e., with order of convergence infinity). The next proposition shows that linear convergence is obtained with the quadratic penalty provided A6 holds with $q = 2$.

PROPOSITION 6. *Let* A7 *hold. If* $\{y^k\}$ *is generated by Algorithm* A, *then* $\{\|y^k - Y^*\|\}$ *converges linearly with convergence ratio*

$$\beta = \limsup_{k \to \infty} \frac{\|y^{k+1} - Y^*\|}{\|y^k - Y^*\|} \leqq \frac{\hat{r}}{\hat{r} + \gamma},$$

where

$$\hat{r} = \limsup_{k \to \infty} r^k \leqq \bar{r}.$$

Proof. By Taylor's theorem,

$$\nabla \phi_{r^k}(t) = \nabla \phi(t/r^k) = t/r^k + o(t/r^k),$$

where $o(\cdot)$ denotes a function with $\lim_{\alpha \to 0} [o(\alpha)/\alpha] = 0$.

By (34)

$$y_i^{k+1} = y_i^k + \nabla \phi(u_i^k/r^k) = y_i^k + u_i^k/r^k + o(u_i^k/r^k).$$

Hence

$$y^{k+1} - \hat{y}^k - u^k/r^k = y^k - \hat{y}^k + o(u^k/r^k).$$

Using (48) (with $v^k = 0$, $q = 2$) and the above equality, we find that

$$\begin{aligned} (1 + \gamma/r^k)\|y^{k+1} - Y^*\| &\leqq \|y^{k+1} - \hat{y}^k - u^k/r^k\| \\ &\leqq \|y^k - Y^*\| + o(u^k/r^k). \end{aligned} \tag{49}$$

But by Lemma 3 and (44),

$$\|u^k/r^k\| \leqq \frac{M_0}{M_1}\|y^k - Y^*\|,$$

so (49) yields

$$\|y^{k+1} - Y^*\| \leqq \frac{r^k}{r^k + \gamma}[\|y^k - Y^*\| + o(\|y^k - Y^*\|)]$$

and

$$\limsup_{k \to \infty} \frac{\|y^{k+1} - Y^*\|}{\|y^k - Y^*\|} \leqq \frac{\hat{r}}{\hat{r} + \gamma}.$$ Q.E.D.

Note that $r^k \to 0$ implies superlinear convergence (i.e., $\beta = 0$) under A7.

Interpretation of results. The last two propositions show that some classes of penalty functions in P_E are more desirable than others. For example, for a fixed q a choice of penalty with $\rho < 1 + 1/(q - 1)$ yields superlinear rate. Furthermore,

when $q > 2$ it can be seen that the quadratic multiplier method may have poor rate of convergence properties relative to a method where $\rho < 1 + 1/(q-1)$. When $q = 2$ it is necessary to use $\rho < 2$ (or $\rho = 2$, $r^k \to 0$) in order to achieve superlinear convergence rate. On the other hand, it should be remembered that the penalized Lagrangian $L_r(\cdot\,; y)$ becomes ill-conditioned if $\nabla\phi$ changes too quickly. This does occur when $\rho < 2$. As a result rapid convergence of $\{y^k\}$ is achieved at the expense of ill-conditioning the unconstrained minimization. However, in situations where one repeatedly solves the same basic problem with minor variations, one may be able to "fine tune" the algorithm by choosing $\{r^k\}, \{\eta^k\}$ in a near optimal fashion. Since good estimates of the solution are already known, the ill-conditioning may not be a problem; then one can exploit the superior convergence rate of the order $\rho < 2$ penalty without incurring undue cost in computing the unconstrained minima. It may be worth bringing to the attention of the reader the fact that our results imply that the order $\rho < 2$ penalties lead to fast convergence only after the method is near convergence. When far from the solution, some geometric arguments indicate [12], [14] that convergence may be slow unless the penalty function ϕ contains, implicitly or explicitly, terms of the form $|t|^{\rho_1}$ where $\rho_1 \geqq 2$. For this reason penalty functions of the form $\phi(t) = |t|^\rho + |t|^{\rho_1}$, $1 < \rho < 2, 2 \leqq \rho_1$, seem to be preferable to functions of the form $\phi(t) = |t|^\rho$, $1 < \rho < 2$.

Comparison to ordinary penalty method. We proceed now to demonstrate that the rate of convergence of multiplier methods is in most cases superior to that of the ordinary exterior penalty method. In the ordinary penalty method, one solves the sequence of unconstrained minimizations

$$\min_{x \in R^n} L_{r^k}(x\,; 0) \quad \text{where } r^k \to 0.$$

The dual update is not used (i.e., $y^k \equiv 0\ \forall k$), but the dual update formula is still relevant since it provides a sequence $\{\tilde{y}^k\}$ of estimates of the Lagrange multiplier. That is,

$$\tilde{y}_i^k = \nabla_1 p_{r^k}[f_i(x^k)\,; 0], \qquad i = 1, \cdots, m.$$

and

$$\|\tilde{y}^k - Y^*\| \to 0.$$

For any $y^* \in Y^*$, using assumption A6 and equation (35), we have for all k such that $\tilde{y}^k \in B(Y^*\,; \delta)$,

$$\begin{aligned} \sup g - h_{r^k}^*[y^*\,; 0] &\leqq g(\tilde{y}^k) - h_{r^k}^*[\tilde{y}^k\,; 0] \\ &\leqq \sup g - \gamma\|\tilde{y}^k - Y^*\|^q - h_{r^k}^*[\tilde{y}^k\,; 0]. \end{aligned}$$

Hence

$$\gamma\|\tilde{y}^k - Y^*\|^q \leqq h_{r^k}^*[y^*\,; 0] - h_{r^k}^*[\tilde{y}^k\,; 0].$$

The sequence $\{\tilde{y}^k\}$ generated by the ordinary penalty method is bounded since g has bounded level sets and $g(\tilde{y}^k) \geqq g_{r^k}(0) \geqq g_{\bar{r}}(0)$, where $r^k \in (0, \bar{r}]$. Furthermore, by (32), $h_r^*[y\,; 0]$ is equal to the real-valued function $r\sum_{i=1}^m \phi^*(y_i)$ on the nonnegative orthant and is hence Lipschitzian on bounded sets contained in this orthant. It

follows from the above inequality that

$$\gamma\|\tilde{y}^k - Y^*\|^q \leqq r^k K\|\tilde{y}^k - y^*\| \quad \forall y^* \in Y^*,$$

where K is the appropriate Lipschitz constant. Hence for all k sufficiently large,

$$\|\tilde{y}^k - Y^*\| \leqq (r^k K/\gamma)^{1/(q-1)}. \tag{50}$$

Convergence bounds similar to (50) with $q = 2$ have been obtained for the quadratic penalty method by Mifflin [16] and Polyak [18] under different assumptions (including differentiability of f_0, f_i). For nonquadratic exterior penalty methods, (50) is the first result of its type. Notice that the scalar ρ does not enter in the estimate (50), and the type of penalty selected does not seem to be material. Recalling Proposition 6, we have that if, for example, $\rho = q = 2$, the rate of convergence of the multiplier method is governed by

$$\|y^{k+1} - Y^*\| \leqq \frac{r^k}{r^k + \gamma}\left[\|y^k - Y^*\| + o(\|y^k - Y^*\|)\right], \tag{51}$$

while Proposition 5 shows that if $(\rho - 1)(q - 1) < 1$, the rate is superlinear. Comparing (51) to (50) with $q = 2$ shows the superiority of the multiplier method. This advantage in speed comes at negligible cost in computational complexity. The dual iteration requires very little additional computer code and an insignificant amount of computer time to execute.

Convergence rate of Algorithm B (*inexact minimization*).

PROPOSITION 7. *Let* $(\rho - 1)(q - 1) \leqq 1$. *If* $\{y^k\}$ *is generated by Algorithm* B, *the* (*Q*-) *order of convergence of* $\{\|y^k - Y^*\|\}$ *is at least* $\rho/[(\rho - 1)q]$.

Proof. By Lemma 4 $\|y^{k+1} - y^k\| \geqq M_1'\|u^k/r^k\|^{\rho-1}$ for sufficiently large k with $M_1' = M_1 \min\{1, m^{1-\rho/2}\}$. Applying (47), we can write

$$\left[\frac{1}{M_1'}\|y^{k+1} - y^k\|\right]^{\sigma-1} \geqq \|u^k/r^k\|$$

$$\geqq \frac{\gamma}{r^k}\|y^{k+1} - Y^*\|^{q-1} - \frac{v^k h_{r^k}^*[y^{k+1}; y^k]}{r^k\|y^{k+1} - Y^*\|},$$

where $\sigma = \rho/(\rho - 1)$. But from (32), (43),

$$\frac{1}{r^k}h_{r^k}^*[y^{k+1}; y^k] = \sum_{i=1}^m \phi^*(y_i^{k+1} - y_i^k)$$

$$\leqq \frac{1}{\sigma M_1^{\sigma-1}}\sum_{i=1}^m |y_i^{k+1} - y_i^k|^\sigma \leqq \frac{d}{\sigma M_1^{\sigma-1}}\|y^{k+1} - y^k\|^\sigma$$

for large k, where $d = \max\{1, m^{1-\sigma/2}\}$. Hence

$$\left[\frac{1}{M_1'}\|y^{k+1} - y^k\|\right]^{\sigma-1} \geqq \frac{\gamma}{r^k}\|y^{k+1} - Y^*\|^{q-1} - \frac{dv^k}{\sigma M_1^{\sigma-1}}\frac{\|y^{k+1} - y^k\|^\sigma}{\|y^{k+1} - Y^*\|}.$$

Equivalently,

$$\gamma M_1^{\sigma-1}\|y^{k+1}-Y^*\|^q - r^k\|y^{k+1}-Y^*\|\left[\frac{M_1}{M_1'}\|y^{k+1}-y^k\|\right]^{\sigma-1} - \frac{dv^k r^k}{\sigma}\|y^{k+1}-y^k\|^\sigma \leqq 0. \tag{52}$$

By Lemma 3 and the triangle inequality, we have

$$\|y^{k+1}-y^k\| \leqq M_0\|y^k-Y^*\|,$$
$$\|y^{k+1}-Y^*\| \leqq \|y^{k+1}-y^k\| + \|y^k-Y^*\| \leqq (1+M_0)\|y^k-Y^*\|.$$

Combining the above three inequalities, we obtain

$$\gamma M_1^{\sigma-1}\|y^{k+1}-Y^*\|^q \leqq \left[r^k(1+M_0)(M_0M_1/M_1')^{\sigma-1} + \frac{dv^kr^k}{\sigma}M_0^\sigma\right]\|y^k-Y^*\|^\sigma.$$

Given that $\sigma = \rho/(\rho-1)$, we have for k sufficiently large,

$$\frac{\|y^{k+1}-Y^*\|}{\|y^k-Y^*\|^{\rho/[(\rho-1)q]}} \leqq M,$$

where M is some scalar, and the proposition is proved. Q.E.D.

Note from (52) that if $\eta^k \to 0$, the order of convergence may increase up to $1/[(\rho-1)(q-1)] = (\sigma-1)/(q-1)$. In particular, if $\{\eta^k/\|y^{k+1}-y^k\|^{(\sigma-q)/(q-1)}\}$ is bounded, one may show that this increased order of convergence is achieved. To see this fact, assume that for all k,

$$\eta^k \leqq c\|y^{k+1}-y^k\|^a, \tag{53}$$

where $c > 0$ and

$$a \geqq \frac{\sigma-q}{q-1} = \frac{1-(\rho-1)(q-1)}{(\rho-1)(q-1)}. \tag{54}$$

Then (52) together with Lemma 3 yields

$$\|y^{k+1}-Y^*\|^q \leqq K_1\|y^{k+1}-Y^*\|\cdot\|y^k-Y^*\|^{\sigma-1} + K_2\|y^k-Y^*\|^{\sigma+a}, \tag{55}$$

where $K_1, K_2 > 0$ are some scalars. Assume that the order of convergence is lower than $(\sigma-1)/(q-1)$ and as a result,

$$\limsup_{k\to\infty}\frac{\|y^{k+1}-Y^*\|}{\|y^k-Y^*\|^{(\sigma-1)/(q-1)}} = \infty.$$

Then for every $M > 0$, there exists a k such that

$$\|y^{k+1}-Y^*\|^{q-1} \geqq M\|y^k-Y^*\|^{\sigma-1}.$$

Using (54) and the above inequality in (55), we obtain

$$\|y^{k+1}-Y^*\|^q \leqq \frac{K_1}{M}\|y^{k+1}-Y^*\|^q + K_2\left(\frac{1}{M}\right)^{(\sigma+a)/(\sigma-1)}\|y^{k+1}-Y^*\|^q,$$

and since M is arbitrarily large we obtain a contradiction. Now it is easy to implement the stopping rule (9) so as to guarantee that (53), (54) hold. Instead of selecting $\{\eta^k\}$ a priori, simply generate it according to the formula

$$\eta^k = \min\{\hat{\eta}^k, c\|s^k(x^k) - y^k\|^a\}, \tag{56}$$

where

$$a \geqq \frac{1 - (\rho - 1)(q - 1)}{(\rho - 1)(q - 1)}, \tag{57}$$

$\{\hat{\eta}^k\}$ denotes a preselected sequence with $\hat{\eta}^k \to 0$, $s^k(\cdot)$ is as defined in (7) and c is an arbitrary positive scalar. The employment of this method for generating $\{\eta^k\}$ restores the order of convergence to $1/[(\rho - 1)(q - 1)]$.

COROLLARY 7.1. *Let* $(\rho - 1)(q - 1) \leqq 1$. *If* $\{y^k\}$ *is generated by Algorithm* B *and the stopping rule* (9) *is operated with* η^k *chosen according to* (56), *then the order of convergence of* $\{\|y^k - Y^*\|\}$ *is at least* $1/[(\rho - 1)(q - 1)]$.

PROPOSITION 8. *Let* A7 *hold. If* $\hat{r}\hat{\eta} < 4\gamma\mu$, *then* $\{\|y^k - Y^*\|\}$ *converges linearly with convergence ratio* β *satisfying*

$$\beta \leqq \frac{1 + \sqrt{1 + \dfrac{2\hat{v}}{1 - \hat{v}}\left(1 + \dfrac{\gamma}{\hat{r}}\dfrac{1}{1 - \hat{v}}\right)}}{2\left(1 + \dfrac{\gamma}{\hat{r}}\dfrac{1}{1 - \hat{v}}\right)}, \tag{58}$$

where $\hat{r} = \limsup_{k\to\infty} r^k$, $\hat{\eta} = \limsup_{k\to\infty} \eta^k$ *and* $\hat{v} = \hat{\eta}/2\mu < 1$. *If* $\hat{r} = 0$, *one has* $\beta = 0$.

Proof. As in Proposition 6,

$$y^{k+1} - \hat{y}^k - u^k/r^k = y^k - \hat{y}^k + o(u^k/r^k).$$

Using (48), we see that

$$\begin{aligned}\left(1 + \frac{\gamma}{r^k}\right)\|y^{k+1} - Y^*\|^2 - \frac{v^k}{r^k}h^*_{r^k}[y^{k+1}; y^k] \\ \leqq \|y^{k+1} - \hat{y}^k - u^k/r^k\| \cdot \|y^{k+1} - Y^*\| \\ \leqq \|y^k - Y^*\| \cdot \|y^{k+1} - Y^*\| + o(\|u^k/r^k\|)\|y^{k+1} - Y^*\|.\end{aligned}$$

By Lemma 3 and (44),

$$\|u^k/r^k\| \leqq \frac{M_0}{M_1}\|y^k - Y^*\|.$$

Also we may bound $h^*_{r^k}[y^{k+1}; y^k]$ using (39) to get

$$\begin{aligned}\left(1 + \frac{\gamma}{r^k}\frac{1}{1 - v^k}\right)\|y^{k+1} - Y^*\|^2 - \frac{v^k}{1 - v^k}\frac{1}{r^k}h^*_{r^k}[\hat{y}^k; y^k] \\ \leqq \|y^k - Y^*\| \cdot \|y^{k+1} - Y^*\| + \|y^{k+1} - Y^*\|o(\|y^k - Y^*\|).\end{aligned}$$

But from (43) and using the fact that $\sigma = \rho = 2$ under A7, we obtain

$$\frac{1}{r^k} h^*_{r^k}[\hat{y}^k; y^k] \leqq \frac{1}{2M_1} \|y^k - Y^*\|^2.$$

Thus

$$\left(1 + \frac{\gamma}{r^k} \frac{1}{1 - v^k}\right) \|y^{k+1} - Y^*\|^2 - \frac{1}{2M_1} \frac{v^k}{1 - v^k} \|y^k - Y^*\|^2$$
$$\leqq \|y^k - Y^*\| \cdot \|y^{k+1} - Y^*\| + \|y^{k+1} - Y^*\| o(\|y^k - Y^*\|).$$

Dividing through by $\|y^k - Y^*\| \cdot \|y^{k+1} - Y^*\|$, we can write

$$\left(1 + \frac{\gamma}{r^k} \frac{1}{1 - v^k}\right) \frac{\|y^{k+1} - Y^*\|}{\|y^k - Y^*\|} - \frac{1}{2M_1} \frac{v^k}{1 - v^k} \frac{\|y^k - Y^*\|}{\|y^{k+1} - Y^*\|} \leqq 1 + \frac{o(\|y^k - Y^*\|)}{\|y^k - Y^*\|}.$$

The above expression is quadratic in the ratio $\|y^{k+1} - Y^*\|/\|y^k - Y^*\|$, and M_1 can be chosen arbitrarily close to unity. Solving the quadratic yields (58) provided $\hat{r} \neq 0$. If $\hat{r} = 0$, one has $\beta = 0$. Q.E.D.

The linear convergence bound (58) is less than one provided $\hat{v} < 2\gamma/\hat{r}$ or equivalently $\hat{r}\hat{\eta} < 4\gamma\mu$. Note that if $\eta^k \to 0$, one has $\hat{v} = 0$ and the bound (58) reduces to $\beta \leqq \hat{r}/(\hat{r} + \gamma)$ which is the same convergence ratio obtained under exact minimization (Proposition 6).

The foregoing analysis shows that if $\eta^k \to 0$, then linear or superlinear convergence rate holds for Algorithm B (inexact minimization) in every case where it has been shown to hold for Algorithm A (exact minimization). However, the bound obtained on the order of convergence (in the case of a superlinear convergence rate) may be worse for Algorithm B than for Algorithm A. Nonetheless, when the sequence $\{\eta^k\}$ is chosen according to (56), then the two bounds on the order of convergence are the same $(1/[(\rho - 1)(q - 1)]$ for a ρ-order penalty with $(\rho - 1) \cdot (q - 1) \leqq 1)$.

5. Computational experience. A number of computer experiments were carried out to test the algorithm under a variety of conditions. The best results were obtained when Algorithm B was used. As expected, it was generally found to be advantageous to take a decreasing sequence of the penalty parameter r^k rather than to keep it fixed. We also compared our algorithm with the ordinary penalty method in which $y^k \equiv 0\ \forall k$.

The computer experiments were performed on the well-known test problem of Rosen and Suzuki:

$$\min f_0(x) = x_1^2 + x_2^2 + 2x_3^2 + x_4^2 - 5x_1 - 5x_2 - 21x_3 + 7x_4$$

subject to:

$$f_1(x) = 2x_1^2 + x_2^2 + x_3^2 + 2x_1 - x_2 - x_4 - 5 \leqq 0,$$
$$f_2(x) = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_1 - x_2 + x_3 - x_4 - 8 \leqq 0,$$
$$f_3(x) = x_1^2 + 2x_2^2 + x_3^2 + 2x_4^2 - x_1 - x_4 - 10 \leqq 0.$$

The optimal solution is $x^* = (0, 1, 2, -1)$ with $f_0(x^*) = -44$. The Lagrange multiplier is $y^* = (2, 1, 0)$, and the first two constraints are active.

Unconstrained minimization was carried out via the Davidon–Fletcher–Powell method, available on the IBM 360 as Scientific Subroutine "FMFP". Table 1 shows the results of a variety of experiments using the quadratic penalty function (3). The initial conditions were $x^0 = (0, 0, 0, 0)$, $y^0 = (0, 0, 0)$. The table shows for each run the number of cycles of the algorithm (i.e., the number of unconstrained minimizations carried out), the total number of line searches required by the unconstrained minimizer and the total number of evaluations of the problem functions. (One evaluation consists of computing f_0, f_i, ∇f_0, ∇f_i, $i = 1, 2, 3$.) Accuracy of x and y refer to the number of decimal places of accuracy to which x^* and y^* were computed. Note that $\eta^k = 0$ corresponds to exact minimization (which in our case corresponds to setting the termination parameters of the FMFP to 10^{-6}).

TABLE 1
Quadratic penalty

Run #	Penalty Constant r^k	Stop Rule Parameter η^k	Multiplier Method # Min Cycle	Multiplier Method # Line Search	Multiplier Method # Fcn Eval.	Multiplier Method x Acc.	Multiplier Method y Acc.	Penalty Method # Min Cycle	Penalty Method # Line Search	Penalty Method # Fcn Eval.	Penalty Method x Acc.	Penalty Method y Acc.
1	$(.1)^k$	$.1 \times (.1)^k$	4	37	128	5	5	7	66	186	6	5
2	$(.1)^k$	$.1 \times (.5)^k$	4	36	130	5	5	8	83	209	6	4
3	$(.1)^k$	$\{(.4)^k\ (.8)^k\}$	4	35	135	5	5	7	68	166	6	5
4	$(.1)^k$	0	4	49	156	5	5	7	92	223	6	5
5	$(.2)^k$	$.1 \times (.2)^k$	5	43	117	6	5	9	82	207	6	5
6	$(.2)^k$	$.1 \times (.5)^k$	5	42	115	6	5	11	93	232	6	5
7	$(.2)^k$	$(.4)^k$	5	41	113	6	5					
8	$(.2)^k$	$(.8)^k$	5	39	114	6	5	11	87	218	6	4
9	$(.2)^k$	1.0	5	40	105	6	5	11	90	225	6	5
10	$(.2)^k$	0	5	60	148	6	5	9	102	257	6	5
11	$(.4)^k$	$(.4)^k$	6	47	113	5	4					
12	$(.4)^k$	$(.8)^k$	6	39	104	5	4	12	79	210	5	4
13	$(.5)^k$	$.1 \times (.5)^k$	7	47	121	5	4	12	89	282	4	3
14	$(.5)^k$	0	7	68	174	5	4	16	156	438	5	4

Table 2 shows a number of computer runs using the order $\rho = \frac{3}{2}$ penalty

$$p(t; y) = \begin{cases} yt + \phi(t) & y + \nabla\phi(t) \geqq 0, \\ \min_\tau \{y\tau + \phi(\tau)\} & y + \nabla\phi(t) < 0, \end{cases}$$

where $\phi(t) = \frac{1}{2}t^2 + \frac{2}{3}|t|^{3/2}$. This penalty is not twice differentiable at zero, so that the penalized Lagrangian is ill-conditioned near the solution. The effects of the ill-conditioning can be seen in Table 2—the amount of computation needed to find the minimum varies considerably from one run to the next. As expected, this penalty generally requires fewer minimization cycles ($q = 2$ for this problem and the order of convergence is 2) but the added difficulty of computing those minimizations may offset the advantage. However, if the algorithm is finely "tuned" by

selecting appropriate sequences $\{r^k\}$ and $\{\eta^k\}$, the order $\rho = \frac{3}{2}$ penalty can outperform the quadratic ($\rho = 2$) as demonstrated by runs 3–6 in Table 2.

TABLE 2
Order $\rho = \frac{3}{2}$ penalty

Run #	Penalty Constant r^k	Stop Rule Parameter η^k	Multiplier Method				
			# Min Cycle	# Line Search	# Fcn Eval.	x Acc.	y Acc.
1	$.1 \times (.1)^k$	$.1 \times (.1)^k$	3	53	181	6	3
2	$.1 \times (.2)^k$	$.1 \times (.2)^k$	3	56	187	6	4
3	$(.1)^k$	$\left\{\begin{matrix}(.4)^k \\ 2 \times (.5)^k\end{matrix}\right\}$	3	33	77	5	5
4	$(.1)^k$	$(.8)^k$	3	32	75	5	4
5	$(.1)^k$	$5 \times (.5)^k$	3	28	81	5	4
6	$(.1)^k$	0	3	41	97	5	4
7	$(.2)^k$	$(.4)^k$	4	50	149	6	4
8	$(.2)^k$	$(.8)^k$	4	49	146	6	4

The computer experiments confirm the superiority of the multiplier method over the ordinary exterior penalty method. Table 1 also suggests that the multiplier method is less sensitive to parameter selection than is the penalty method. Finally, inexact minimization offers a significant reduction in computation over the algorithm with exact minimization.

Appendix. This appendix contains proofs of some results which are important for the proof of our main propositions. The first proposition lists some properties of the functions p in P. Its proof is left to the reader (see also [12], [14]). The other proposition relates to properties of the Lagrangian functions L and L_r.

PROPOSITION A.1. *Let $p \in P$. For all $t \in R$ and $y \geqq 0$, there holds*

1. $\nabla_2 p(t; y) \geqq t$,
2. $t\nabla_1 p(t; y) \geqq p(t; y) \geqq y\nabla_2 p(t; y) \geqq yt$,
3. $p(t; y) - yt \geqq p(t; 0)$.

Furthermore the following are equivalent:

4a. $t\nabla_1 p(t; y) = p(t; y)$,
4b. $p(t; y) = yt$,
4c. $p(t; y) = 0$,
4d. $\nabla_1 p(t; y) = y$,
4e. $t \leqq 0$ *and* $yt = 0$.

Let $h: R^n \to (-\infty, +\infty]$ be a closed, proper convex function. A vector $z \in R^n$, $z \neq 0$ determines a *direction* in R^n, namely the direction of the ray emanating from the origin and passing through z. Fixing $x \in R^n$, the one-dimensional function $\eta(t) = h(x + tz)$, $t \in R$, is a cross section of h through x in the direction z. The direction z is called a *direction of recession* of h if $\eta(t)$ is nonincreasing over the entire real line. It is known that every level set of h is nonempty and bounded and its minimum set is compact if and only if h has no directions of recession

([20, Cor. 8.7.1, Thm. 27.1(d)]). Consider the *recession function* $h0^+$ of h which may be given as [20, Thm. 8.5, Cor. 8.5.2]

$$h0^+(z) = \lim_{t\to\infty} \frac{h(x + tz) - h(x)}{t} = \lim_{t\downarrow 0} th(z/t). \tag{A.1}$$

The direction z is a direction of recession of h if and only if $h0^+(z) \leqq 0$. (In fact, this last statement is usually taken as the definition [20, § 8].)

PROPOSITION A.2. *Under* A1, A2 (of § 2) *for any* $r > 0$, $y \in R^m$ *with* $y_i \geqq 0$, $i = 1, \cdots, m$ *the augmented Lagrangian* $L_r(\cdot\,; y)$ *has no directions of recession.*

Proof. We need to compute the recession function $L_r0^+(\cdot\,; y)$. By Theorem 9.3 of [20], we have

$$L_r0^+(z; y) = f_0 0^+(z) + \sum_{i=1}^{m} h_i 0^+(z),$$

where h_i is given by

$$h_i(x) = p_r[f_i(x); y_i].$$

Using (A.1) we have

$$h_i0^+(z) = \lim_{t\to\infty} \frac{p_r[f_i(x + tz); y_i] - p_r[f_i(x); y_i]}{t}. \tag{A.2}$$

Suppose z is a direction of recession of f_i. Then

$$f_i(x + tz) \leqq f_i(x) \qquad \forall t \geqq 0.$$

We have for all $t \geqq 0$,

$$-\infty < \inf_u p_r(u; y_i) \leqq p_r[f_i(x + tz); y_i] \leqq p_r[f_i(x); y_i].$$

It follows then that the limit in (A.2) is zero. Now suppose z is not a direction of recession of f_i. By (A.1)

$$h_i0^+(z) = \lim_{t\downarrow 0} tp_r[f_i(z/t); y_i] = \lim_{t\downarrow 0} tp_r[tf_i(z/t)/t; y_i].$$

Since $f_i0^+(z) = \lim_{t\downarrow 0} tf_i(z/t) > 0$, we have $tf_i(z/t) \geqq \alpha > 0$, $0 < t \leqq t_0$. Then

$$h_i0^+(z) \geqq \lim_{t\downarrow 0} tp_r(\alpha/t; y_i) = p_r0^+(\alpha; y_i) = +\infty.$$

Hence

$$h_i0^+(z) = \begin{cases} 0 & \text{if } z \text{ is a direction of recession of } f_i, \\ \infty & \text{if } z \text{ is not a direction of recession of } f_i. \end{cases}$$

Consequently

$$L_r0^+(z; y) = \begin{cases} f_0 0^+(z) & \text{if } z \text{ is a direction of recession of each } f_i,\ i = 1, 2, \cdots, m, \\ +\infty & \text{if } z \text{ is not a direction of recession of some } f_i, \\ & i = 1, 2, \cdots, m. \end{cases}$$

But f_0, f_i, $i = 1, 2, \cdots, m$, have no common directions of recession by A2 (of § 2). That is, $f_0 0^+(z) > 0$ if z is a direction of recession of each f_i, $i = 1, \cdots, m$. Hence, $L_r 0^+(z; y) > 0$ $\forall z \neq 0$, and L_r has no direction of recession. Q.E.D.

Notes added in proof.

Note 1. The results contained in this paper have been presented at the IEEE Conference on Decision and Control, San Diego, California, December, 1973; at the SIGMAP–UW Nonlinear Programming Symposium, Madison, Wisconsin, April, 1974; and at the IFAC Sixth Triennial World Congress, Boston, Massachusetts, August, 1975. They have also appeared (with some variations and without proofs) in references [13] and [15] and in:

D. P. BERTSEKAS, *Multiplier Methods: A Survey*, Preprints of IFAC Sixth Triennial World Congress, Part IB, Boston, Mass., Aug. 1975.

Note 2. The class of penalty functions and the multiplier updating formulas introduced in this paper may also be utilized in conjunction with several algorithms of the multiplier type other than those proposed here. The class of penalty functons introduced here finds additional application in algorithms such as those presented in:

D. P. BERTSEKAS, *Nondifferentiable Optimization via Approximation*, Proc. 12th Ann. Allerton Conf. on Circuit and System Theory, Allerton Park, Ill., Oct. 1974, pp. 41–52; also in *Mathematical Programming Study* 3, M. Balinski and P. Wolfe, eds., North-Holland, Amsterdam, to appear.

———, *A General Method for Approximation Based on the Method of Multipliers*, Proc. of 13th Ann. Allerton Conf. on Circuit and System Theory, Allerton Park, Ill., Oct. 1975.

———, *Multiplier Methods for Two-sided Inequality Constraints and Related Algorithms*, Coordinated Science Laboratory Working Paper, Univ. of Illinois, Urbana, Ill., Aug. 1975.

REFERENCES

[1] D. P. BERTSEKAS, *Combined primal–dual and penalty methods for constrained minimization*, EES Dept. Working Paper, Stanford Univ., Stanford, Calif., 1973; this Journal, 13 (1975), pp. 521–544.

[2] ———, *On penalty and multiplier methods for constrained minimization*, EES Dept. Working Paper, Stanford Univ., Stanford, Calif., 1973; this Journal, 14 (1976), pp. 216–235.

[3] ———, *On the method of multipliers for convex programming*, EES Dept. Working Paper, Stanford Univ., Stanford, Calif., 1973; IEEE Trans. Automatic Control, AC-16 (1975), pp. 385–388.

[4] ———, *On penalty and multiplier methods for constrained minimization*, Nonlinear Programming 2 (Proc. SIGMAP-UW Nonlinear Programming Symposium, Madison, Wisc., 1974), O. Mangasarian, S. Robinson and R. Meyer, eds., Academic Press, New York, 1975, pp. 165–191.

[5] ———, *Convergence rate of penalty and multiplier methods*, Proc. of 1973 IEEE Conf. on Decision and Control, San Diego, Calif., IEEE publication no. 73 CHO 806-O SMC, IEEE, New York, 1973, pp. 260–264.

[6] ———, *Necessary and sufficient conditions for a penalty method to be exact*, EES Dept. Working Paper, Stanford Univ., Stanford, Calif., 1973; Math. Programming, to appear.

[7] M. J. BOX, *A comparison of several current optimization methods and the use of transformations in constrained problems*, Comput. J., 9 (1966), pp. 67–77.

[8] J. D. BUYS, *Dual algorithms for constrained optimization*, Ph.D. thesis, Univ. of Leiden, the Netherlands, 1972.

[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. of 1972 IEEE Conf. on Decision and Control, New Orleans, La., IEEE publication no. 72 CHO 705-4 SCS, IEEE, New York, 1972, pp. 162–166.

[12] B. W. KORT, *Combined primal-dual and penalty function algorithms for nonlinear programming*, Ph.D. thesis, Stanford Univ., to appear.

[13] B. W. KORT AND D. P. BERTSEKAS, *Multiplier methods for convex programming*, Proc. 1973 IEEE Conf. on Decision and Control, San Diego, Calif., IEEE publication no. 73 CHO 806-O SMC, IEEE, New York, 1973, pp. 428–432.

[14] ———, *Combined primal-dual and penalty methods for convex programming*, EES Dept. Working Paper, Stanford Univ., Stanford, Calif., 1973.

[15] B. W. KORT, *Rate of convergence of the method of multipliers with inexact minimization*, Nonlinear Programming 2 (SIGMAP-UW Nonlinear Programming Symposium, Madison, Wisc., 1974). O. Mangasarian, S. Robinson and R. Meyer, eds., Academic Press, New York, 1975, pp. 193–214.

[16] R. MIFFLIN, *Convergence bounds for nonlinear programming algorithms*, Tech. Rep. 57, Dept. of Administrate Sciences, Yale Univ., New Haven, Conn., 1972; Math. Programming, 8 (1975), pp. 251–271.

[17] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[18] B. T. POLYAK, *The convergence rate of the penalty function method*, Ž. Vyčisl. Mat. i Mat. Fiz., 11 (1971), pp. 3–11.

[19] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.

[20] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[21] ———, *A dual approach to solving non-linear programming problems by unconstrained optimization*, Math. Programming, 5 (1973), pp. 354–373.

[22] ———, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optimization Theory Appl., 12 (1973), pp. 555–562.

[23] ———, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.

[24] ———, *Penalty methods and augmented Lagrangians in nonlinear programming*, Proc. of the Fifth IFIP Conference on Optimization Techniques, Rome, Springer-Verlag, Berlin, pp. 418–425.

[25] ———, *Solving a nonlinear programming problem by way of a dual problem*, Symposia Mathematica, Academic Press, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

CONVERGENCE AND STABILITY PROPERTIES OF THE DISCRETE RICCATI OPERATOR EQUATION AND THE ASSOCIATED OPTIMAL CONTROL AND FILTERING PROBLEMS*

WILLIAM W. HAGER† AND LARRY L. HOROWITZ‡

Abstract. The convergence properties for the solution of the discrete time Riccati matrix equation are extended to Riccati operator equations such as arise in a gyroscope noise filtering problem. Stabilizability and detectability are shown to be necessary and sufficient conditions for the existence of a positive semidefinite solution to the algebraic Riccati equation which has the following properties: (i) it is the unique positive semidefinite solution to the algebraic Riccati equation, (ii) it is converged to geometrically in the operator norm by the solution to the discrete Riccati equation from any positive semidefinite initial condition, (iii) the associated closed loop system converges uniformly geometrically to zero and solves the regulator problem, and (iv) the steady state Kalman–Bucy filter associated with the solution to the algebraic Riccati equation is uniformly asymptotically stable in the large. These stability results are then generalized to time-varying problems; also it is shown that even in infinite dimensions, controllability implies stabilizability.

1. Introduction. The purpose of this paper is to prove that the convergence and stability properties associated with the Riccati difference equation in finite dimensions also hold for the Riccati operator equation in infinite dimensions. Many of the finite-dimensional results already in the literature will also be strengthened. The Riccati difference equation has been studied by Caines and Mayne [2], Lee, Chow and Barr [9], and Zabczyk [10].

In finite dimensions, the first paper proved that if a stabilizability and an observability assumption held, then the solution to the Riccati difference equation converged to a positive definite matrix solving the algebraic Riccati equation, and furthermore, the solution to the algrebraic equation was unique in the class of positive semidefinite matrices. Their proof, however, required the Heine–Borel theorem (a closed, bounded set of $n \times n$ matrices forms a compact set) so that the proofs could not be extended to the Riccati operator equation.

The paper by Lee, Chow and Barr then showed that in a Hilbert space environment, the solution to the quadratic cost control problem could be expressed in feedback form in terms of the solution to the Riccati operator equation, and when the system dynamics were stable, then there existed a solution to the algebraic Riccati equation.

Zabczyk weakened this stability condition to stabilizability and then showed that if the cost functional was positive definite in the state variable, then the solution to the algebraic Riccati operator equation was unique in the class of positive semidefinite operators and furthermore was the limit (in the operator norm) of the solution to the Riccati equation from any positive semidefinite initial condition.

This paper contains the results above as special cases. The observability condition of Caines and Mayne and the positive definiteness of the cost functional required by Zabczyk are weakened to detectability. The positive definiteness of

* Received by the editors August 22, 1974, and in revised form March 11, 1975.

† Department of Mathematics, University of South Florida, Tampa, Florida 33620.

‡ Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts, 02173.

the Riccati equation solution proved by Caines and Mayne is also proved in the infinite-dimensional framework. Furthermore, it is shown that the solution to the regulator problem associated with the Riccati equation is uniformly asymptotically stable in the large if a detectability condition is satisfied and there exists a positive semidefinite solution to the algebraic Riccati equation.

The paper reaches its climax in § 5, where it is proved that stabilizability and detectability are necessary and sufficient conditions for the existence of a positive semidefinite solution to the algebraic Riccati equation which has the following properties: (i) it is the unique positive semidefinite solution to the algebraic Riccati equation, (ii) it is converged to geometrically in the operator norm by the solution to the discrete Riccati equation from any positive semidefinite initial condition, and (iii) the associated closed loop system converges uniformly geometrically to zero and solves the regulator problem.

The stability of the Kalman–Bucy filter for time-varying infinite-dimensional systems under stabilizability and detectability is also treated in the Appendix. This weakens the conditions of controllability, observability, and nonsingularity of the transition operator that Deyst and Price [3] required in their proof of the stability of the solution to the time-varying filtering problem in finite dimensions.

The paper concludes with an illustration of the use of the Riccati operator equation in filtering the noise additively corrupting a gyroscope's output signal. In this example, the domain of the Riccati operator is an L^2-space.

2. Problem statement. Let $K(S, T, i)$ denote the solution to the *Riccati operator equation* given by

$$K(i-1) = A^*(i)\{K(i) - K(i)B(i)[R(i) + B^*(i)K(i)B(i)]^{-1}B^*(i)K(i)\}A(i) + Q(i) \tag{1}$$

with boundary condition $K(T) = S$, where i is an integer, $i \leqq T$, and the following operators appearing in (1) are uniformly bounded linear mappings on Hilbert spaces Y and U: $Q(i): Y \to Y$, $S: Y \to Y$, $A(i): Y \to Y$, $B(i): U \to Y$, and $R(i): U \to U$. (Throughout this paper, the term operator will mean a bounded linear operator.) The inner products on both Hilbert spaces will be denoted by $(\cdot, \cdot)$—the inner product being used should be clear from context. The norm of a vector $y \in Y$ is given by $\|y\| = (y, y)^{1/2}$ and the norm of a linear operator $P: Y \to Y$ is given by $\|P\| = \sup\{\|Py\| : \|y\| = 1\}$. The operator P^* denotes the adjoint of an operator P. P is said to be positive if it is positive semidefinite and self-adjoint; i.e., $P^* = P$ and $(y, Py) \geqq 0$ for all $y \in Y$. The operators $Q(i)$, $R(i)$, and S are assumed positive, and furthermore, $R(i)$ is assumed uniformly positive definite, i.e., $(u, R(i)u) \geqq a\|u\|^2$ for some $a > 0$ and for all $u \in U$, where "a" is independent of i. The notation $P_1 \geqq P_2$ and $P_1 > P_2$ means that $P_1 - P_2$ is positive semidefinite and positive definite respectively.

Associated with the Riccati equation is the *control problem*:

$$\underset{\{u(i)\}}{\text{Minimize}} \left[(Sy(T), y(T)) + \sum_{i=i_0}^{T-1} \{(y(i), Q(i)y(i)) + (u(i), R(i)u(i))\} \right], \tag{2}$$

$$\text{Subject to} \quad y(i+1) = A(i)y(i) + B(i)u(i), \tag{3}$$

$$y(i_0) = y_0 \in Y, \qquad u(i) \in U.$$

Let $J(S, T, i_0, y_0)$ denote the optimal value for the control problem above. As shown in [1] for finite-dimensional spaces,

$$J(S, T, i_0, y_0) = (y_0, K(S, T, i_0)y_0), \tag{4}$$

and the optimal control in feedback form is given by

$$u(i) = -[R(i) + B^*(i)K(S, T, i+1)B(i)]^{-1}B^*(i)K(S, T, i+1)A(i)y(i). \tag{5}$$

The extension of these results to Hilbert spaces is trivial as noted in [6], since the dynamic programming argument used in the derivation of (4) and (5) does not require finite-dimensionality and can be performed in a Hilbert space environment.

The cost function (2) is nonnegative, so (4) implies that $K(S, T, i) \geqq 0$ for all $i \leqq T$, and hence the inverse appearing in (1) and (5) exists and is bounded since $R(i) > 0$. Thus $K(S, T, i)$ is a positive operator for $i \leqq T$.

When (1) is time-invariant (i.e., $A(i) = A$, $B(i) = B$, etc.), then also associated with (1) is the *algebraic Riccati equation* (abbreviated ARE):

$$K = A^*[K - KB(R + B^*KB)^{-1}B^*K]A + Q. \tag{6}$$

Similarly associated with the control problem when the system is time invariant is the *regulator problem*

$$\underset{\{u(i)\}}{\text{Minimize}} \left[\sum_{i=0}^{\infty} (y(i), Qy(i)) + (u(i), Ru(i)) \right], \tag{7}$$

$$\begin{aligned} &\text{Subject to} \quad y(i+1) = Ay(i) + Bu(i), \\ &\qquad\qquad\quad y(0) = y_0 \in Y, \quad u(i) \in U. \end{aligned} \tag{8}$$

Let $J(y_0)$ denote the optimal cost for the regulator problem above.

The estimation problem, or dual problem corresponding to the control problem, is given in Appendix C.

For future reference, the following abbreviations are used throughout the paper:

ARE	algebraic Riccati equation
UASL	uniformly asymptotically stable in the large
ST	stabilizability
DT	detectability
CT	controllability
OB	observability

3. The assumptions. The following *stabilizability* and *detectability* assumptions will appear in the development. These conditions are first stated for time-invariant problems:

(ST) There exists an integer $r \geqq 1$, a constant q, and an operator L such that

$$\|(A - BL)^r\| < q < 1. \tag{9}$$

(DT) There exist integers $s, t \geqq 0$ and constants $0 \leqq d < 1$, $0 < b < \infty$, such that

whenever $\|A^i y\| \geqq d\|y\|$, then

$$\left(y, \sum_{i=0}^{s} A^{*i} Q A^i y\right) \geqq b(y, y). \tag{10}$$

When the problem is time varying, we replace L in (ST) by a sequence $\{L(i)\}$ of uniformly bounded linear operators and require

$$\left\| \prod_{i=k}^{k+r-1} (A(i) - B(i)L(i)) \right\| < q < 1 \tag{ST'}$$

for $k = 0, r, 2r, \cdots$.

Similarly in (DT) we replace A^i by $C(i + k, k)$, where $C(i, k) = A(i - 1) \cdot A(i - 2) \cdots A(k)$ and $C(i, i) = I$, the identity operator, and require that for all $k \geqq 0$, whenever $\|C(k + t, k)y\| \geqq d\|y\|$, then

$$\left(y, \sum_{i=0}^{s} C(k + i, k)^* Q(k + i) C(k + i, k) y\right) \geqq b(y, y). \tag{DT'}$$

Special cases of (ST) and (DT) are the *controllability* and *observability* conditions:

(CT) There exists an integer $r \geqq 0$ and a constant $0 < a < \infty$ such that

$$\left(y, \sum_{i=0}^{r} A^i B B^* A^{*i} y\right) \geqq a(y, y) \tag{11}$$

for all $y \in Y$.

(OB) There exists an integer $s \geqq 0$ and a constant $0 < b < \infty$ such that

$$\left(y, \sum_{i=0}^{s} A^{*i} Q A^i y\right) \geqq b(y, y) \tag{12}$$

for all $y \in Y$.

Note that (OB) is trivially a special case of (DT). At the end of § 4, it will also be shown that (CT) implies (ST).

Recall that in finite dimensions, the pair of matrices $[A, B]$ are said to be stabilizable if there exists a matrix L such that the spectral radius $\rho(A - BL)$ is less than 1. (A, B, and L are assumed to be $n \times n$, $n \times m$, and $m \times n$ respectively.) Similarly $[C, A]$ is detectable if $[A^*, C^*]$ is stabilizable. Note that it follows immediately that (ST) is equivalent to the condition $\rho(A - BL) < 1$ for some L since $\rho(P) = \lim_{k \to \infty} \|P^k\|^{1/k}$ (see [4, p. 567]).

In Appendix B, it is proved that in finite dimensions, (DT) is equivalent to the condition that $\rho(A^* - C^*L) < 1$ for some L where $Q = C^*C$.

4. The main results. The first lemma gives a uniform bound for the solution $K(S, T, i)$ of the Riccati equation (1).

LEMMA 1. *If* (ST') *holds, then there exists a constant c independent of i and T such that $K(S, T, i) < cI$ and $J(y) < c\|y\|^2$, where $J(y_0)$ is the optimal cost for the regulator problem* (7).

Proof. By the relation (4), the bound on $K(S, T, i)$ will be proved if the optimal cost in the control problem (2) can be bounded in terms of the initial condition y_0. Since the operators $A(\cdot)$, $B(\cdot)$, and $L(\cdot)$ are all uniformly bounded, there exists a constant c such that

$$\prod_{i=j}^{j+m} \|N(i)\| \leqq c \tag{13}$$

for all m satisfying $0 \leqq m \leqq r$, where $N(i) = A(i) - B(i)L(i)$. (Throughout this paper, c will denote a generic constant whose value does not depend on T or i and whose value in different equations may change.)

Using the control $u(i) = -L(i)y(i)$ in the system dynamics leads to the estimates

$$\|y(k+1)\| = \|N(k)y(k)\| = \left\| \prod_{i=0}^{k} (N(i))y_0 \right\| \leqq cq^{k/r}\|y_0\|, \tag{14}$$

where the last inequality follows by grouping the operators $N(i)$ into groups of r factors and then applying the bound (ST'). Since $u(i) = -L(i)y(i)$, then $u(i)$ obeys a similar estimate. Inserting these bounds on $u(i)$ and $y(i)$ into the cost functional (2) leads to a bound on $J(S, T, i, y_0)$ of the form $c\sum_{k=0}^{\infty} q^{2k/r}\|y_0\|^2$. Since $q < 1$, the geometric series is convergent and $J(S, T, i, y_0) < c\|y_0\|^2$ as desired. Since c is independent of T and i, then the bound on $J(y)$ also follows immediately. □

A sequence of operators P_k is said to *converge strongly* to P if $\lim_{k\to\infty} \|(P - P_k)y\| = 0$ for all $y \in Y$. An elementary property of operators is the following (see [4, p. 925]): Suppose $\{P_k\}$ is a sequence of uniformly bounded self-adjoint operators satisfying $P_k \leqq P_{k+1}$ for $k \geqq 0$; then $\{P_k\}$ converges strongly to a self-adjoint operator P satisfying $P_k \leqq P$ for all $k \geqq 0$. The sequence P_k *converges weakly* to P if $\lim_{k\to\infty} (z, (P_k - P)y) = 0$ for all $y, z \in Y$. It can be shown that this last condition is equivalent to requiring $\lim_{k\to\infty} (y, (P_k - P)y) = 0$ for all $y \in Y$.

For the remainder of this section, we will only be dealing with the time-invariant Riccati equation and control problem. In Appendix A, the question of stability for time varying systems is considered. Let $K(T, i)$ denote the solution to the time-invariant Riccati equation when the terminal condition vanishes ($S = 0$).

THEOREM 1. *If* $J(0, T, 0, y) < c\|y\|^2$ *for some* c *independent of* T, *then* $K(T, i)$ *converges strongly as* $T \to \infty$ *to a positive operator* P *that satisfies the* ARE.

Proof. Since (4) holds, then $K(T, i) < cI$ and hence $\|K(T, i)\|$ is uniformly bounded by c. Also, $(y, K(T_1, i)y) = J(0, T_1, i, y) \geqq J(0, T_2, i, y) = (y, K(T_2, i)y)$ whenever $T_1 \geqq T_2$ since increasing the terminal time cannot decrease the optimal cost. Thus by the remarks preceding the theorem, $K(T, i) \to P$ strongly as $T \to \infty$. If $F(K)$ denotes the right-hand side of (6), then (1) can be written as $K(T+1, i) = K(T, i-1) = F(K(T, i))$, where the first equality follows since the equation is time-invariant. Now $K(T+1, i) \to P$ strongly as $T \to \infty$ and furthermore by [4, p. 922], $F(K(T, i)) \to F(P)$ strongly as $T \to \infty$. Thus $P = F(P)$ and hence P solves the ARE. □

Combining Lemma 1 and Theorem 1 yields the following.

COROLLARY 1. *If* (ST) *holds, then* $K(T, i) \to P$ *strongly as* $T \to \infty$, *where* P *solves the* ARE.

Later it will be shown that when (DT) holds and there exists a positive solution to the (ARE), then (ST) holds.

The stability of the solution to the following system when P is a positive solution to the ARE will now be studied:

$$y(i+1) = Ay(i) + Bu(i), \quad y(0) = y_0, \quad u(i) = Fy(i), \tag{15}$$

$$F = -[R + B^*PB]^{-1}B^*PA. \tag{16}$$

The following system of inequalities and equalities plays an important role in the development:

$$-(u(i), B^*PAy(i)) = (u(i), [R + B^*PB]u(i)) \tag{17}$$

$$(y(i), Py(i)) \geqq (y(i), Py(i)) - (y(j), Py(j)) \tag{18}$$

$$= \sum_{k=i}^{j-1} (y(k), Py(k)) - (y(k+1), Py(k+1)) \tag{19}$$

$$= \sum_{k=i}^{j-1} (y(k), Py(k) - A^*PAy(k)) - (u(k), B^*PAy(k)) - (B^*PAy(k), u(k)) - (u(k), B^*PBu(k)) \tag{20}$$

$$= \sum_{k=i}^{j-1} (y(k), Qy(k)) - (u(k), B^*PBu(k)) - (u(k), B^*PAy(k)) \tag{21}$$

$$= \sum_{k=i}^{j-1} (y(k), Qy(k)) + (u(k), Ru(k)) \geqq 0. \tag{22}$$

Above, $j > i$ and (17) follows by multiplying $u(i) = Fy(i)$ by $[R + B^*PB]$ and (18), (20), (21), and (22) follow by the positivity of P, (15), the ARE that P satisfies, and (17), respectively.

THEOREM 2. *$J(0, T, 0, y) < c\|y\|^2$ for some constant c independent of T if and only if there exists a positive solution to the* ARE.

Proof. The theorem in the forward direction was proved by Theorem 1. Now suppose P is a positive solution to the ARE and let $y_s(i)$ and $u_s(i)$ be the state and control generated by (15). Then by the relation (18),

$$(y_0, Py_0) \geqq \sum_{k=0}^{T-1} (y_s(k), Qy_s(k)) + (u_s(k), Ru_s(k)). \tag{23}$$

Since P is bounded, then $J(0, T, 0, y) \leqq \|P\| \|y\|^2$. □

Recall that the dynamical system $x(k+1) = f(x(k), k)$, $x(i_0) = x_0$ is said to be *uniformly asymptotically stable in the large* (abbreviated UASL) with respect to x^* if the following holds [8]:

(i) Given $\varepsilon > 0$, there exists $\delta > 0$ such that $\|x^* - x_0\| \leqq \delta$ implies that $\|x(k) - x^*\| \leqq \varepsilon$ for any k, i_0 satisfying $k \geqq i_0$.

(ii) Given $\delta > 0$, there exists $\varepsilon > 0$ such that $\|x^* - x_0\| \leqq \delta$ implies $\|x(k) - x^*\| \leqq \varepsilon$ for any k, i_0 satisfying $k \geqq i_0$.

(iii) Given δ, $\varepsilon > 0$, there exists T such that $\|x(k) - x^*\| \leqq \varepsilon$ for all k, i_0, x_0 satisfying $k \geqq T + i_0$ and $\|x_0 - x^*\| \leqq \delta$.

THEOREM 3. *If* $K(T,0) \to P$ *strongly,* P *solves the* ARE, *and the system* (15) *is* UASL *with respect to the origin, then* P *is the unique solution to the* ARE *in the class of positive operators and* $K(S,T,i)$ *converges strongly to* P *as* $T \to \infty$ *for any* $S \geqq 0$. *Also the state and the control generated by* (15) *are the optimal solutions for the regulator problem and* (y_0, Py_0) *is the optimal cost.*

Proof. Let $\{y_s(i)\}$ and $\{u_s(i)\}$ be generated by (15) using P. Then (18) implies

$$(24) \quad (y_0, Py_0) \geqq \sum_{k=0}^{T-1} (y_s(k), Qy_s(k)) + (u_s(k), Ru_s(k)) \geqq (y_0, K(T,0)y_0) \leqq J(y_0).$$

Since $K(T,0) \to P$, then as $T \to \infty$, the $\geqq$'s in (24) become $=$'s, and the last $\leqq$ implies that $\{u_s(i)\}$ must actually achieve the optimal cost in the regulator problem. Since the cost function (7) is a strictly convex function of $\{u(i)\}$, then $\{u_s(i)\}$ must be the unique optimal control sequence and (y_0, Py_0) is the optimal cost.

Now consider the following inequalities:

$$(25) \quad (y_s(T), Sy_s(T)) + (y_0, Py_0) \geqq (y_s(T), Sy_s(T)) + \sum_{k=0}^{T-1} [(y_s(k), Qy_s(k)) + (u_s(k), Ru_s(k))]$$

$$(26) \quad \geqq (y_0, K(S,T,0)y_0) \geqq (y_0, K(0,T,0)y_0).$$

The second inequality above follows since $(y_0, K(S,T,0)y_0)$ is the optimal cost in the control problem (2) and the third inequality follows since the optimal cost when $S = 0$ is bounded by the optimal cost when a nonnegative terminal cost is present. By assumption, the right side of (26) converges to (y_0, Py_0) as $T \to \infty$, and since the system (15) is UASL with respect to the origin, then $y_s(T) \to 0$ as $T \to \infty$. Thus all the inequalities in (25) become equalities as $T \to \infty$ and hence $K(S,T,0) \to P$ weakly. An elementary application of the Schwarz inequality for positive operators shows that weak convergence implies strong convergence (see again [4, p. 925]).

If $\bar{P}$ is any positive solution to the ARE, then it is easy to see that $K(\bar{P},T,0) = \bar{P}$ for all T and since $K(\bar{P},T,0) \to P$, then $\bar{P} = P$. □

Now it is shown that if (DT) holds, then the stability condition of Theorem 3 is satisfied.

THEOREM 4. *Suppose* P *is a positive operator solving the* ARE *and* (DT) *holds; then the solution to the system* (15) *is* UASL *with respect to the origin.*

Proof. It is shown that $\|y(k+i)\| \leqq c2^{-i/N}\|y(k)\|$ for some N, $c > 0$ independent of k and i, so that the theorem follows immediately.

Step 1. *Suppose that for some* i, $\|A^i y(i)\| \geqq d\|y(i)\|$, *where* d *was given in* (DT); *then there exists a constant* $m > 0$ *independent of* i *such that*

$$(27) \quad (y(i), Py(i)) - (y(i+s+1), Py(i+s+1)) \geqq m\|y(i)\|^2.$$

Proof of Step 1. Let Δ^2 denote the left side of (27) and let c again denote a generic constant. By (18),

$$(28) \quad \Delta^2 \geqq \sum_{k=i}^{i+s} (u(k), Ru(k)) \geqq a \sum_{k=i}^{i+s} \|u(k)\|^2,$$

where a satisfies $R > aI > 0$.

Letting $z(\cdot)$ denote the solution to $z(k+1) = Az(k), z(k=i) = y(i)$, then the error $e(k) = y(k) - z(k)$ satisfies, for $i \leqq k \leqq i+s$,

$$\|e(k+1)\| \leqq \|A\|\,\|e(k)\| + \|B\|\,\|u(k)\| \leqq \sum_{j=i}^{k} \|A\|^{k-j}\|B\|\,\|u(j)\| \tag{29}$$

$$\leqq c\sum_{j=i}^{k} \|u(j)\| \leqq c\left[\sum_{j=i}^{k} \|u(j)\|^2\right]^{1/2} \leqq c\Delta, \tag{30}$$

where the last set of inequalities follow by the Schwarz inequality and the bound (28) on the control.

The relation (18) also yields

$$\Delta^2 \geqq \sum_{k=i}^{i+s} (y(k), Qy(k)) = \sum_{k=i}^{i+s} (e(k)+z(k), Q(e(k)+z(k))) \tag{31}$$

$$\geqq \sum_{k=i}^{i+s} (y(i), A^{*k-i}QA^{k-i}y(i)) - 2\|e(k)\|\,\|Q\|\,\|A^{k-i}y(i)\| \tag{32}$$

$$\geqq b\|y(i)\|^2 - c\Delta\|y(i)\|, \tag{33}$$

where b was given in (10); the inequality (32) follows by the Schwarz inequality and (33) follows by the bound on $e(k)$ in (30). Completing the square in (33) leads to $\|y(i)\|^2 \leqq c\Delta^2$, the desired result.

Step 2. Suppose that $\|A^t y(i)\| \leqq d\|y(i)\|$ *for* $i = k, k+t, \cdots, k+nt$. *Then there exists a constant* M *independent of* n *and* k *such that* $\|y(i)\|^2 \leqq M\|y(k)\|^2$ *for* $k \leqq i \leqq k+nt$.

Proof of Step 2. For notational convenience, suppose $k = 0$. First let $j = lt$ where $0 \leqq l \leqq n$. Then

$$\|y(j)\| = \left\| A^t y(j-t) + \sum_{i=0}^{t-1} A^i Bu(j-1-i) \right\| \leqq d\|y(j-t)\| + c\sum_{i=0}^{t-1} \|u(j-1-i)\| \tag{34}$$

$$\leqq d\|y(j-t)\| + c\left(\sum_{i=0}^{t-1} \|u(j-1-i)\|^2\right)^{1/2} \tag{35}$$

$$\leqq d^l\|y(0)\| + c\left(\sum_{i=0}^{j-1} \|u(i)\|^2\right)^{1/2}, \tag{36}$$

where the Schwarz inequality was used to derive (35) and the last inequality follows by writing the solution to the difference inequality (35) as the convolution of the forcing term $c(\sum_{i=0}^{t-1} \|u(j-1-i)\|^2)^{1/2}$ with d^i and then applying the Schwarz inequality to the convolution; since $d < 1$, then the $\sum d^{2i}$ factor in the Schwarz inequality is bounded. Now by (18),

$$a\sum_{i=0}^{j-1} \|u(i)\|^2 \leqq \|P\|\,\|y(0)\|^2, \tag{37}$$

where $R > aI$. Inserting this bound in (36) yields the desired estimate for $j = lt$.

For $lt < j < (l+1)t$, the relation $y(k+1) = Ay(k) + Bu(k)$ combined with the bound (37) on the controls and the bound above on $\|y(lt)\|$ proves the estimate.

Step 3. *Suppose that* $s_{j+1} \geqq s_j$, $s_j \to \infty$ *as* $j \to \infty$ *and* $|s_j - s_{j+1}|$ *is bounded independent of* j. *Then there exists a constant* c *independent of* j *such that* $\|y(i)\| \leqq c\|y(s_j)\|$ *for* $s_j \leqq i \leqq s_{j+1}$ *and for all* j.

Proof of Step 3. $y(i) = A^{i-s_j}y(s_j) + \sum_{k=s_j}^{i-1} A^{i-k-1}Bu(k)$. Since $|i - s_j| \leqq |s_{j+1} - s_j|$ is uniformly bounded, then the bound on $y(i)$ follows immediately from a bound of the form (37) on the controls where $y(0)$ is replaced by $y(s_j)$ and the summation is from $k = s_j$ to $i - 1$.

Let $\sqrt{D}$ be the maximum constant given in Step 3 corresponding to those sequences of integers $\{s_j\}$ satisfying $s_{j+1} = s_j + s + 1$. Now choose N_1, N_2, and N_3 large enough that the following conditions hold:

$$\|P\|/mN_1 < \tfrac{1}{4}, \tag{38}$$

$$d^{N_2}\overline{M}^{1/2} + c(\|P\|/aN_3)^{1/2} < \tfrac{1}{2}, \tag{39}$$

$$\text{where } \overline{M} = \max\{M, MD\|P\|/m\}, \tag{40}$$

where m was given in (27), M appeared in Step 2, c is the same constant appearing on the right side of (36), D appeared above at the end of Step 3, and $d < 1$ is given in (DT). Let $N = N_1N_2M_3 \max(s+1, t)$.

Step 4. *There exists* $i \in [k, k+N]$ *such that* $\|y(i)\| < \frac{1}{2}\|y(k)\|$ *for any* $k \geqq 0$.

Proof. For notational convenience, choose $k = 0$. Construct a sequence $\{t_j\}$ and $\{f_j\}$ as follows: $t_0 = 0$; for $j \geqq 0$,

$$\text{if } \|A^t y(t_j)\| \leqq d\|y(t_j)\|, \quad \text{then } t_{j+1} = t_j + t, \quad f_j = 0,$$

$$\text{if } \|A^t y(t_j)\| > d\|y(t_j)\|, \quad \text{then } t_{j+1} = t_j + s + 1, \quad f_j = 1.$$

By (18), $(y(t_j), Py(t_j)) - (y(t_{j+1}), Py(t_{j+1})) \geqq 0$, so combining this with (27) yields

$$(y(t_j), Py(t_j)) - (y(t_{j+1}), Py(t_{j+1})) \geqq f_j m\|y(t_j)\|^2. \tag{41}$$

Let J be the first index with $t_j \geqq N$. Adding the inequalities (41) for $j = 0, 1, \cdots, J-1$ yields

$$\begin{aligned}\|P\|\,\|y(0)\|^2 \geqq (y(0), Py(0)) &\geqq (y(t_J), Py(t_J)) + \sum_{j=0}^{J-1} f_j m\|y(t_j)\|^2 \\ &\geqq \sum_{j=0}^{J-1} f_j m\|y(t_j)\|^2.\end{aligned} \tag{42}$$

If at least N_1 of the f_j do not vanish, then the sum on the right side of (42) is bounded below by $mN_1 \min \|y(t_j)\|^2$, where the min is over j such that $f_j = 1$. If $j = n$ achieves the minimum, then $\|y(t_n)\|^2 \leqq \|P\|\,\|y(0)\|^2/mN_1$. Hence Step 4 would follow by (38).

Now if less than N_1 of the f_j equal 1, then there is a sequence of N_2N_3 consecutive j's with $f_j = 0$ since $N = N_1N_2N_3 \max(s+1, t)$ and hence $J \geqq N_1N_2N_3$. Let $k_1 = t_j$ be chosen such that $f_{j+i} = 0$ for $0 \leqq i \leqq N_2N_3 - 1$ and either $f_{j-1} = 1$ or $t_j = 0$. Let $k_2 = t_l$ mark the end of this sequence of f_i's that vanish. By Step 2,

$\|y(i)\|^2 \leqq M\|y(k_1)\|^2$ whenever $k_1 \leqq i \leqq k_2$. If $f_{j-1} = 1$, then the inequalities $\|y(k_1)\|^2 \leqq D\|y(t_{j-1})\|^2 \leqq D\|P\|\,\|y(0)\|^2/m$ follow by Step 3, the choice of D above and (42). Combining these last two sets of inequalities yields $\|y(i)\|^2 \leqq MD\|P\|\,\|y(0)\|^2/m$ if $k_1 \neq 0$ and $\|y(i)\|^2 \leqq M\|y(0)\|^2$ if $k_1 = 0$. Thus $\|y(i)\|^2 \leqq \overline{M}\|y(0)\|^2$, where $\overline{M}$ is given in (40).

Divide $[k_1, k_2]$ into subintervals of length $N_2 t$. Since $|k_1 - k_2| \geqq N_2 N_3 t$, then there are $\geqq N_3$ of these subintervals. By (37), one of these subintervals $[r_1, r_2]$ must satisfy

$$\sum_{i=r_1}^{r_2} \|u(i)\|^2 \leqq \frac{\|P\|}{aN_3}\|y(0)\|^2 \tag{43}$$

(i.e., the smallest sum of the form (43) is bounded by the average sum).

For $j = r_1 + N_2 t = r_2$, the inequality (34) implies

$$\|y(r_2)\| \leqq d^{N_2}\|y(r_1)\| + c\left\{\sum_{i=r_1}^{r_2} \|u(i)\|^2\right\}^{1/2}. \tag{44}$$

Inserting the bounds above on $\|y(i)\|^2$ and (43) into (44) yields

$$\|y(r_2)\| \leqq \left[d^{N_2}\overline{M}^{1/2} + c\left(\frac{\|P\|}{aN_3}\right)^{1/2}\right]\|y(0)\| < \tfrac{1}{2}\|y(0)\|, \tag{45}$$

where the last inequality follows by (39). This completes Step 4 and the geometric convergence follows by combining Steps 3 and 4. □

COROLLARY 2. *If P is a positive solution to the* ARE *and* (DT) *holds, then P is the unique solution to the* ARE *in the class of positive operators and $K(S, T, i) \to P$ geometrically in the operator norm as $T \to \infty$ for any $S \geqq 0$. Also the state and the control generated by* (15) *are optimal solutions to the regulator problem and the solution to the system* (15) *converges to zero uniformly and geometrically.*

Proof. By Theorems 2 and 1, there exists a solution $\bar{P}$ to the ARE such that $K(T, i) \to \bar{P}$ strongly as $T \to \infty$. By Theorem 4, since (DT) holds, the system (15) is UASL with respect to the origin and hence by Theorem 3, $\bar{P} = P$ and (y_0, Py_0) is the optimal cost for the regulator problem.

Let $y(T, i)$ denote the optimal solution to the control problem (2) in the time-invariant case when $S = 0$ and $i_0 = 0$. It can be shown that $y(T, i) \to 0$ uniformly and geometrically as $T \to \infty$. This follows since (18) holds with $(y(i), Py(i))$ replaced by $(y(T, i), K(T, i)y(T, i))$, and hence all the steps of Theorem 4 are valid with $y(i)$ replaced by $y(T, i)$ and P replaced by $K(T, i)$. Note that the proof of Theorem 4 required a bound on $\|P\|$ and hence will require a uniform bound on $\|K(T, i)\|$ for the finite terminal-time case; however, since (y_0, Py_0) is the optimal cost for the regulator problem by Theorem 3, then $(y_0, Py_0) \geqq (y_0, K(T, i)y_0)$ and $\|P\| \geqq \|K(T, i)\|$. (Berberian [11] shows that if Z is a positive operator, then $\|Z\| = \sup\{(y, Zy) : \|y\| = 1\}$.)

Now

$$(y_0, Py_0) \leqq (y_0, K(T, 0)y_0) + (y(T, T), Py(T, T)).$$

Combining this with (25) yields

$$\begin{aligned}(y_s(T), Sy_s(T)) + (y_0, Py_0) &\geqq (y_0, K(S, T, 0)y_0)\\ &\geqq (y_0, Py_0) - (y(T, T), Py(T, T)).\end{aligned}$$

Since there exist c, q satisfying $\|y(T, T)\|, \|y_s(T)\| \leqq cq^T\|y_0\|$ and $0 < q < 1$, then $|(y_0, Py_0 - K(S, T, 0)y_0)| \leqq cq^{2T}\|y_0\|^2$ for some $c > 0$. Hence Berberian's theorem can now be used to prove that $\|P - K(S, T, 0)\| \leqq cq^{2T}$.

The remaining results in this corollary follow from Theorems 3 and 4. □

To summarize the previous results we have the following theorem.

THEOREM 5. *If* (ST) *and* (DT) *hold, then* $K(S, T, i)$ *converges geometrically in the operator norm as* $T \to \infty$ *to a positive operator* P *that is the unique positive solution to the* ARE. *Also, the control and state generated by* (15) *is* UASL *with respect to the origin and is the unique solution to the regulator problem.*

When the control problem is observable, then any positive solution to the ARE is actually positive definite.

THEOREM 6. *Suppose* P *is a positive solution to the* ARE *and* (OB) *holds. Then* $P > 0$ *and is the unique solution to the* ARE *in the class of positive operators.*

Proof. By Step 1 of Theorem 4, whenever (10) holds, then (27) holds. When the control problem is observable, however, (10) holds all the time so $(y_0, Py_0) \geqq (y(s + 1), Py(s + 1)) + m\|y_0\|^2 \geqq m\|y_0\|^2$ for some $m > 0$. The fact that P is the unique positive solution to the ARE follows by Corollary 2. □

Now cases are presented where the converse of Corollary 1 holds.

THEOREM 7. *If there exists a positive solution* P *of the* ARE *such that the system* (15) *is* UASL *with respect to the origin, then* (ST) *holds.*

Proof. Define $G = A - B[R + B^*PB]^{-1}B^*PA$ and suppose $\|G^k\| \geqq 1$ for all $k \geqq 0$. Then there exists y_k such that $\|G^k y_k\| > \frac{1}{2}$ and $\|y_k\| = 1$. This contradicts condition (iii) in the definition of UASL and so there exists $r \geqq 0$ with $\|G^r\| < 1$. Now (ST) holds for $L = [R + B^*PB]^{-1}B^*PA$. □

COROLLARY 3. *If there exists a positive solution* P *to the* ARE *and* (DT) *holds, then* (ST) *holds.*

Proof. This follows immediately by Theorems 4 and 7.

THEOREM 8. *If* (CT) *holds, then* (ST) *holds.*

Proof. The solution to the system equation (3) is

$$y(r + 1) = A^{r+1}y_0 + \sum_{i=0}^{r} A^iBu(r - i) = A^{r+1}y_0 + M[u(0), \cdots, u(r)], \tag{46}$$

where M is the linear operator on the controls appearing in the middle of (46). Note that the range space of M contains the range space of MM^* and furthermore the operator MM^* is precisely the operator appearing in (11). Thus MM^* is positive definite and hence there exists a solution $\bar{y}$ to the equation $-A^{r+1}y_0 = MM^*y$. Hence the control sequence $M^*\bar{y}$ inserted in (46) yields $y(r + 1) = 0$. From the equation that $\bar{y}$ satisfies and the positive definiteness of MM^*, $a\|\bar{y}\|^2 \leqq (\bar{y}, MM^*\bar{y}) = -(\bar{y}, A^{r+1}y_0) \leqq \|\bar{y}\|\,\|A\|^{r+1}\|y_0\|$ or $\|\bar{y}\| \leqq c\|y_0\|$, where "$a$" is given in (11).

Now choose Q, R to be any positive operators satisfying $R, Q > 0$. Using the control sequence $M^*\bar{y}$ for the controls $\{u(0), \cdots, u(r)\}$ and $u(j) = 0$ for $j > r$ results in $y(j) = 0$ for $j > r$ and the cost function (2) is bounded by $c\|y_0\|^2$ since $\|\bar{y}\| \leqq c\|y_0\|$ and the first $r + 1$ controls are given by $M^*\bar{y}$. By Theorem 1, there exists a positive solution of the ARE and since $Q > 0$, then (DT) holds. Corollary 3 completes the proof. □

Remark. It also follows that the steady state Kalman–Bucy filter for the dual estimation problem corresponding to the control problem (2) in the time-invariant case is uniformly asymptotically stable in the large with respect to the origin when (DT) holds and P solves the ARE. The homogeneous part of the Kalman–Bucy filter (presented in Appendix C) is given by

$$\begin{aligned} x(n+1|n+1) &= (A^* - PB[R + B^*PB]^{-1}B^*A^*)x(n|n) \\ &= (A^* - PB[R + B^*PB]^{-1}B^*A^*)^{n+1}x(0|0) \\ &= \{A[A - B(R + B^*PB)^{-1}B^*PA]^n[I - B(R + B^*PB)^{-1}B^*P]\}^*x(0|0), \end{aligned}$$

where the last equation follows by taking the adjoint of the prior equation twice and then regrouping terms. Theorems 4 and 7 imply that

$$\|[A - B(R + B^*PB)^{-1}B^*PA]^k\| < 1$$

for k large enough. Thus it is easy to see that the homogeneous part of the Kalman–Bucy filter is UASL.

5. Necessary and sufficient conditions. The results of the previous section are now tied together in the following theorem.

THEOREM 9. *The following conditions are all equivalent*:

(a) (ST) *and* (DT) *hold.*

(b) *There exists a unique positive solution* P *to the* ARE. *For any* $S \geqq 0$, $K(S, T, i) \to P$ *geometrically in the operator norm as* $T \to \infty$, *and the solution to* (15) *both solves the regulator problem and is* UASL *with respect to the origin.*

(c) *There exists a positive solution to the* ARE *and* (DT) *holds.*

(d) (DT) *holds and* $J(0, T, 0, y) \leqq c\|y\|^2$ *for some* c *independent of* T.

Proof. By Theorem 2, (c) and (d) are equivalent. By Theorem 5, (a) implies (c) and by Corollary (2), (c) implies (b). The proof will be complete when it is shown that (b) implies (a).

If (b) holds, then by Theorem 7, (ST) holds. Now suppose (DT) is violated and let P be as given in (b). Then given any ε, T, t, there exists $y(\varepsilon, T, t)$ such that $\|y(\varepsilon, T, t)\| = 1$, $\|A^t y(\varepsilon, T, t)\| > \frac{1}{2}$, and $(y(\varepsilon, T, t), M(T)y(\varepsilon, T, t)) \leqq \varepsilon$, where $M(T) = \sum_{i=0}^{T-1} A^{*i}QA^i$.

Now fix t and define $F(P) = A - B[R + B^*PB]^{-1}B^*PA$. It is easy to see that there exist constants $c, \delta > 0$ depending on P such that $\|F(P) - F(P')\| \leqq c\|P - P'\|$ whenever $\|P - P'\| \leqq \delta$. Let $y(\varepsilon, T, t, i)$ and $y_s(\varepsilon, T, t, i)$ denote the solutions to $y(i+1) = F(K(T, i+1))y(i)$, $y(0) = y(\varepsilon, T, t)$ and $z(i+1) = F(P)z(i)$, $z(0) = y(\varepsilon, T, t)$ respectively.

The error $e(\varepsilon, T, t, i) = y_s(\varepsilon, T, t, i) - y(\varepsilon, T, t, i)$ is the solution $e(i)$ to the equation

$$\begin{aligned} e(i+1) &= F(K(T, i+1))e(i) + [F(P) - F(K(T, i+1))]y_s(\varepsilon, T, t, i) \\ &= \sum_{j=0}^{i} \left(\prod_{k=j+2}^{i+1} F(K(T, k)) \right) \delta F(T, j) y_s(\varepsilon, T, t, j), \end{aligned}$$

where $\delta F(T, i) = F(P) - F(K(T, i+1))$ and $e(0) = 0$.

Since the system (15) is UASL with respect to the origin, and $\|y_s(\varepsilon, T, t, 0)\| = \|y(\varepsilon, T, t)\| = 1$, then $\|y_s(\varepsilon, T, t, i)\|$ is bounded uniformly in ε, T, t, and i. By Theorem 3, (y_0, Py_0) is the optimal cost in the regulator problem and hence $P \geqq K(T, i) \geqq 0$ for $i \leqq T$ and $\|K(T, i)\|$ is bounded uniformly in T and i. Also, note that if $a > 0$ satisfies $R > aI$, then $\|[R + B^*ZB]^{-1}\| \leqq 1/a$ for any positive operator Z and hence $\|F(K(T, i))\|$ is uniformly bounded. Combining these uniform bounds with the fact that t is fixed and $\|\delta F(T, i-1)\| = \|F(K(T, i)) - F(P)\| \leqq c\|K(T, i) - P\| \to 0$ as $T \to \infty$, implies that for T sufficiently large, $\|e(\varepsilon, T, t, t)\| \leqq \frac{1}{8}$ (independent of ε).

Now hold T fixed and consider the following lemma.

LEMMA 2. *Suppose $\mathscr{A}(\cdot, \cdot)$ is a continuous bilinear form on $U \times U$ satisfying $\mathscr{A}(u, u) \geqq a\|u\|^2$ for all $u \in U$ and some $a > 0$ independent of u, and $\mathscr{B}(\cdot)$ is a bounded linear operator. Then the problem to minimize $J(u) = \mathscr{A}(u, u) + \mathscr{B}(u)$ over $u \in U$ has a unique solution $u^* \in U$ and $\|u - u^*\|^2 \leqq (J(u) - J(u^*))/a$.*

Proof of lemma. The existence and uniqueness of u^* and the necessary condition $2\mathscr{A}(u - u^*, u^*) + \mathscr{B}(u - u^*) = 0$ for all $u \in U$ follows by [12]. If $J(u)$ is expanded about u^* and the necessary condition is applied, then the following holds: $J(u) - J(u^*) = \mathscr{A}(u - u^*, u - u^*) \geqq a\|u - u^*\|^2$. □

Note that the control problem (2) satisfies the conditions for the lemma since $(u(k), Ru(k)) \geqq a\|u(k)\|^2$ (where $R > aI > 0$), $(y(k), Qy(k)) \geqq 0$, and the cost functional is a quadratic in $\{u(i)\}$ when $\{y(i)\}$ is expressed in terms of $\{u(i)\}$. Thus if $J(\varepsilon, T, t)$ denotes the optimal cost in (2) when the initial condition is $y(0) = y(\varepsilon, T, t)$ and $S = 0$, and $J_0(\varepsilon, T, t)$ is the cost generated by the control sequence $u(k) = 0$ for $k \geqq 0$ starting from the same initial condition, then the relation $\varepsilon \geqq (y(\varepsilon, T, t), M(T)y(\varepsilon, T, t)) = J_0(\varepsilon, T, t) \geqq J(\varepsilon, T, t) \geqq 0$ implies that $\varepsilon/a \geqq (J_0(\varepsilon, T, t) - J(\varepsilon, T, t))/a \geqq \sum_{i=0}^{T-1} \|u(\varepsilon, T, t, i)\|^2$, where $u(\varepsilon, T, t, i)$ is the optimal control sequence for the control problem (2) corresponding to the initial condition $y(0) = y(\varepsilon, T, t)$. (Recall that the solution to $y(i+1) = F(K(T, i+1))y(i)$, $y(0) = y(\varepsilon, T, t)$, which was labeled $y(\varepsilon, T, t, i)$ above, is also the solution to the control problem (2) and so the notation above for the optimal control is compatible with the notation for the optimal state.)

Let $y_0(\varepsilon, T, t, i)$ denote the solution to $y(i+1) = Ay(i)$, $y(0) = y(\varepsilon, T, t)$. Then the error $e_0(\varepsilon, T, t, i) = y(\varepsilon, T, t, i) - y_0(\varepsilon, T, t, i)$ satisfies the equation $e(i+1) = Ae(i) + Bu(\varepsilon, T, t, i)$, $e(0) = 0$. Using the above bound on the controls implies that for ε sufficiently small, $\|e_0(\varepsilon, T, t, t)\| < \frac{1}{8}$.

To summarize,

$$\begin{aligned}\|y_s(\varepsilon, T, t, t) - y_0(\varepsilon, T, t, t)\| &\leqq \|y_s(\varepsilon, T, t, t) - y(\varepsilon, T, t, t)\| \\ &\quad + \|y(\varepsilon, T, t, t) - y_0(\varepsilon, T, t, t)\| \\ &\leqq \tfrac{1}{8} + \tfrac{1}{8} = \tfrac{1}{4}.\end{aligned}$$

By assumption, $\|y_0(\varepsilon, T, t, t)\| = \|A^t y(\varepsilon, T, t)\| > \frac{1}{2}$. Thus for all t, it is possible to choose T large enough and ε small enough so that $\|y_s(\varepsilon, T, t, t)\| > \frac{1}{4}$. However, this violates the assumption that the system (15) is UASL with respect to the origin (see condition (iii) in the definition of UASL). Hence (DT) must hold. □

6. A gyroscope noise filtering problem. A practical problem motivating the study of operator Riccati equations is the gyroscope noise filtering problem which is described briefly below. The additive noise corrupting gyroscopic output readings are observed experimentally to often possess a $1/f$ behavior in power spectral density over a wide band of frequency. To model this noise as the output of a linear system, a continuum of first order linear systems are used with time constants, r, of the linear systems described by a probability density function $p(r)$. The filtering problem is equivalent via duality to solving the following operator control problem.

The state $y(k) \in Y$ is given by a pair $[b(\cdot, k), a(k)]$, where $a(k)$ is an $m \times 1$ vector and $b(\cdot, k) \in L^2([r_1, r_2])$; i.e.,

$$\int_{r_1}^{r_2} b(r, k)^2 \, dr < \infty.$$

The limits r_1 and r_2 satisfy $0 < r_1 < r_2 < \infty$. The inner product on Y is given by

$$(y_1, y_2) = \int_{r_1}^{r_2} b_1(t) b_2(t) \, dt + a_1^* a_2,$$

where $y_1 = [b_1(\cdot), a_1]$ and $y_2 = [b_2(\cdot), a_2]$. The controls $u(k) \in U$ are scalars and the inner product on U is simply multiplication. The operators A and B in the system dynamics (3) are given by

$$A[b(\cdot), a] = [e^{-z/(\cdot)} b(\cdot), \bar{A}a],$$

$$B[u] = [p(\cdot)u, hu],$$

where $p(\cdot)$ is bounded and measurable, $\bar{A}$ is an $m \times m$ matrix, h is an $m \times 1$ vector, and $z > 0$.

The cost functional is

$$\left[(Sy(T), y(T)) + \sum_{k=0}^{T-1} \int_{r_1}^{r_2} Q(r) b(r, k)^2 \, dr + a(k)^* Q a(k) + u(k)^2 d \right],$$

where $Q \geqq 0$ is an $m \times m$ matrix, $Q(r) \geqq c > 0$ is a bounded measurable function, $d > 0$ is a scalar, and $S \geqq 0$ is a positive semidefinite operator.

Note that this problem is not controllable and, in fact, inserting the operators A and B into the controllability condition (11) results in

$$\sum_{i=0}^{r} \left\{ \int_{r_1}^{r_2} p(r) \, e^{-zi/r} b(r) \, dr \right\}^2 \geqq a \int_{r_1}^{r_2} b(r)^2 \, dr$$

for some $a > 0$ and for all $b \in L^2([r_1, r_2])$. This is clearly impossible (for example, consider a sequence of functions $\{b_j(\cdot)\}$ converging to a delta function). The L^2 part of the system dynamics, however, trivially satisfies the stabilizability condition with $L = 0$ since $e^{-z/r} \leqq e^{-z/r_2} < 1$ for $r_1 \leqq r \leqq r_2 < \infty$. The L^2 part of the system dynamics is also observable for $s = 0$ since $Q(r) \geqq c > 0$. Thus if the linear system $a(k+1) = \bar{A}a(k) + hu(k)$ is stabilizable and the matrix $[\sum_{k=0}^{s} \bar{A}^{*k} Q \bar{A}^k] > 0$ for some s, then all the theorems in § 4 apply.

More details on the gyroscope problem are given in [6].

Appendix A. Stability of the time varying Kalman–Bucy filter and control problem solution. The stability result in Theorem 4 can be generalized to the time-varying case, where

$$y(i+1) = A(i)y(i) + B(i)u(i),$$

$$y(0) = y_0 \in Y,$$

$$u(i) = F(S, T, i)y(i),$$

$$F(S, T, i) = -[R(i) + B^*(i)K(S, T, i+1)B(i)]^{-1}B^*(i)K(S, T, i+1)A(i).$$

Note that since $K(S, T, \cdot)$ is only defined on $[-\infty, T]$, then $y(\cdot)$ is only defined on $[0, T]$, and hence it no longer makes sense to ask whether $y(\cdot)$ is stable. However, (ST′) and (DT′) are sufficient to prove the following properties for $y(y_0, T, \cdot)$, the solution to the system above:

(i) Given $\varepsilon > 0$, there exists $\delta > 0$ such that $\|y_0\| \leqq \delta$ implies that $\|y(y_0, T, i)\| \leqq \varepsilon$ whenever $T \geqq i \geqq 0$ (δ independent of T).

(ii) Given $\delta > 0$, there exists $\varepsilon > 0$ such that $\|y(y_0, T, i)\| \leqq \varepsilon$ whenever $\|y_0\| \leqq \delta$ and $T \geqq i \geqq 0$ (ε independent of T).

(iii) Given ε, $\delta > 0$, there exists T' such that $\|y(y_0, T, i)\| \leqq \varepsilon$ whenever $\|y(y_0, T, j)\| \leqq \delta$ and $T \geqq i \geqq j + T'$ (T' independent of T).

This is essentially the same as the definition of UASL except that the index for $y(\cdot)$ must be confined to the range $0 \leqq i \leqq T$. The proof of these results is identical to the proof of Theorem 4. Note that the condition (18) holds with $(y(k), Py(k))$ replaced by $(y(k), K(S, T, k)y(k))$. All the steps of Theorem 4 are valid in the time-varying case with $K(S, T, j)$ replacing P. Since a bound was required on $\|P\|$ in various places in the proof, we must now require that $K(S, T, j)$ be bounded uniformly in T and j. Lemma 1, however, shows that when (ST′) holds, $\|K(S, T, j)\|$ is bounded uniformly.

As in the remark at the end of § 4, it follows that the Kalman–Bucy filter for the dual estimation problem corresponding to the control problem is uniformly asymptotically stable in the large with respect to the origin in the time-varying case when (ST′) and (DT′) hold.

Appendix B. (DT) and detectability. We now show that in finite dimensions, (DT) is equivalent to the condition $\rho(A^* - C^*L) < 1$ for some L where $Q = C^*C$. Hautas proves [5] that this last condition is equivalent to requiring that every unstable eigenvector of A is observable, i.e., when e is an eigenvector of A corresponding to the eigenvalue λ and $|\lambda| \geqq 1$, then $Ce \neq 0$.

PROPOSITION B.1. *Every unstable eigenvector of A is observable if and only if* (DT) *holds for $Q = C^*C$.*

Proof. If (DT) holds, $Ae = \lambda e$, $\|e\| = 1$, and $|\lambda| \geqq 1$, then $\|A^t e\| \geqq \|e\|$ and hence by (10),

$$\left(e, \sum_{i=0}^{s} A^{*i}C^*CA^i e\right) = \sum_{i=0}^{s} |\lambda|^{2i}\|Ce\|^2 \geqq b > 0.$$

Thus $Ce \neq 0$.

Conversely, suppose every unstable eigenvector of A is observable. It is easy to see that the nullspace of $M(k) = \sum_{i=0}^{k} A^{*i}C^*CA^i$ is contained in the nullspace of $M(k-1)$. Thus the nullspace is a decreasing function of k and there exists some integer s such that the nullspace is unchanged for $k \geqq s$.

We only treat the case where there is a complete set of normalized eigenvectors $\{e_k\}$ corresponding to the eigenvalues $\{\lambda_j\}$. The changes necessary for defective eigenvalues are summarized at the end of the proof.

If d is any constant satisfying $0 < d < 1$, then the following result is now proved:

(*) There exists an integer $t \geqq 0$ such that whenever $\|A^t y\| \geqq d\|y\|$, then the expansion $y = \sum a_k e_k$ has $a_k \neq 0$ for some unstable eigenvector e_k.

Form the matrix $N = [e_1, e_2, \cdots, e_n]$. Since the $\{e_k\}$ are independent, then N^{-1} exists and hence if $\|y\| = 1$ and $x = (a_1, a_2, \cdots, a_n)^*$ is defined by $x = N^{-1}y$, then $\|x\|^2 = \sum |a_k|^2 \leqq \|N^{-1}\|^2$. Define $a = \|N^{-1}\|$ and choose t large enough so that if λ_k is a stable eigenvalue, then $|\lambda_k|^t < d/(na)$. If $\|y\| = 1$ and $\|A^t y\| \geqq d$, then expanding y in terms of $\{e_k\}$ leads to $\|A^t y\| = \|A^t \sum a_k e_k\| = \|\sum a_k \lambda_k^t e_k\| \geqq d$. Suppose that a_k vanishes for all the unstable eigenvectors. Then the bounds $|a_k| \leqq a$ and $\|e_k\| = 1$ imply that $\|\sum a_k \lambda_k^t e_k\| \leqq \sum |\lambda_k|^t a < d$, where the last inequality follows since the previous sum is only over stable eigenvalues. This is a contradiction, and hence a_k cannot vanish for all the unstable eigenvectors.

Let f be any vector that minimizes $(y, M(s)y)$ over all real vectors satisfying $\|y\| = 1$ and $\|A^t y\| \geqq d$, and suppose that the optimal value of this minimization problem is zero. If it is not zero, then (DT) is immediately satisfied. Recall that a positive semidefinite matrix can be expressed as $D^T D$ so that $(f, M(s)f) = 0$ if and only if $M(s)f = 0$. Thus $M(k)f = 0$ for $k \geqq s$ since the nullspace of $M(k)$ is invariant for $k \geqq s$. Since $\|A^t f\| \geqq d$, then $a_j \neq 0$ for some unstable component in the expansion $f = \sum a_k e_k$. Let λ_j be the eigenvalue of the biggest modulus such that $a_j \neq 0$, and first let us assume that λ_j is real. Then $\lim_{k\to\infty} \lambda_j^{-k} A^k f = e$, where e is an unstable eigenvector (note that any nonzero linear combination of eigenvectors corresponding to a given eigenvalue is also an eigenvector corresponding to the same eigenvalue). Thus $\lim_{k\to\infty} |\lambda_j|^{-2k}(f, A^{*k}C^*CA^k f) = \|Ce\|^2$. Since $M(k)f = 0$ for $k \geqq s$, then $CA^k f = 0$ for $k \geqq s$ and hence $Ce = 0$. This violates the assumption that none of the unstable eigenvectors of A lies in the nullspace of C.

If λ_j occurs in a complex conjugate pair, then $|\lambda_j|^{-k} CA^k f \to C(e^{ik\theta}e + e^{-ik\theta}\bar{e})$ where $\bar{e}$ is the complex conjugate of e. Since $\theta \neq 0, \pi$, then as $k \to \infty$,we conclude that two linearly independent combinations of $\bar{e}$ and e lie in the nullspace of C (i.e., there exists a subsequence k_j of the k's that converges to a vector in the nullspace of C. Then consider $k_j' = k_j + 1$ and extract another convergent subsequence). Hence $Ce = C\bar{e} = 0$ which is again impossible.

We now summarize the changes for the case of defective eigenvalues. Write A in Jordan canonical form as $A = NDN^{-1}$, where D has eigenvalues on the diagonal and either 1's or 0's on the upper subdiagonal. Let $\{e_k\}$ denote the columns of N. The proof above is almost unaltered until the point where it was shown that $Ce = 0$ which violated the condition that the unstable eigenvectors cannot lie in the nullspace of A. Note now that e may no longer be an eigenvector; however, if e_j is not an eigenvector, then one property of the Jordan decomposition

above is that $Ae_j = \lambda_j e_j + e_{j-1}$. Thus $A^k e_j$ for k large enough will have a component which is an unstable eigenvector. The remainder of the proof (which is still very complicated) involves looking at convergent subsequences as above in the case of a complex conjugate pair of eigenvalues. □

Appendix C. The estimation problem. The estimation problem corresponding to the control problem (2) is now presented. Consider the following linear system and observation sequence $\{z(i)\}$:

$$x(i+1) = A^*(i)x(i) + w(i),$$

$$z(i) = B^*(i)x(i) + v(i),$$

where (i) $w(i)$, $x(i) \in Y$, $z(i)$, $v(i) \in U$, (ii) $x(0)$ is a random variable with mean x_0 and covariance Σ_0 satisfying $E[(Fx(0))^2] = F\Sigma_0 F^*$ for any $F: Y \to R$ = the real numbers, (iii) $\{w(i)\}$ and $\{v(i)\}$ are zero mean white noise with covariances $\{Q(i)\}$ and $\{R(i)\}$ satisfying $E[(Fw(i))^2] = FQ(i)F^*$ and $E[(Gv(i))^2] = GR(i)G^*$ for any $F: Y \to R$ and $G: U \to R$ respectively. Also, $x(0)$, $\{w(i)\}$, and $\{v(i)\}$ are assumed uncorrelated.

The estimation problem is to find a sequence of vectors $\{\hat{x}(i|i)\}$ that minimizes $E[(\hat{x}(i|i) - x(i), \hat{x}(i|i) - x(i))]$, where the estimate $\hat{x}(i|i)$ is based on the observations to time i. The Kalman–Bucy filter corresponding to the estimation problem is given by

$$\hat{x}(n+1|n+1) = A^*(n)\hat{x}(n|n) + \Sigma(n+1|n)B(n+1)[R(n+1) + B^*(n+1)$$
$$\cdot \Sigma(n+1|n)B(n+1)]^{-1}[z(n+1) - B^*(n+1)A^*(n)\hat{x}(n|n)],$$

$$\hat{x}(0|0) = x_0,$$

where $\Sigma(n+1|n)$ is generated by

$$\Sigma(n+1|n) = A^*(n)[\Sigma(n|n-1) - \Sigma(n|n-1)B(n)[R(n) + B^*(n)\Sigma(n|n-1)B(n)]^{-1}$$
$$\cdot B^*(n)\Sigma(n|n-1)]A(n) + Q(n),$$

$$\Sigma(0, -1) = \Sigma_0.$$

REFERENCES

[1] R. Boudarel, J. Delmas and P. Guichet, *Dynamic Programing and its Application to Optimal Control*, Academic Press, New York, 1971, pp. 46–48.

[2] P. E. Caines and D. Q. Mayne, *On the discrete time matrix Riccati equation of optimal control*, Internat. J. Control, 12 (1970), pp. 785–794.

[3] J. J. Deyst, Jr. and C. F. Price, *Conditions for asymptotic stability of the discrete minimum-variance linear estimator*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 702–705.

[4] N. Dunford and T. Schwartz, *Linear Operators*, Interscience, New York, 1963.

[5] M. L. J. Hautas, *Stabilization, controllability, and observability of linear autonomous systems*, Indag. Math., 32 (1970), pp. 448–455.

[6] L. L. Horowitz, *Optimal filtering of gyroscopic noise*, Ph.D. thesis, Mass. Inst. of Tech., Cambridge, Mass., 1974.

[7] R. E. Kalman and J. E. Bertram, *Control system analysis and design via the second method of Lyapunov: I. Continuous time systems*, Trans. ASME, J. Basic Engrg., 82 (1960), pp. 371–393.

[8] ———, *Control system analysis and design via the second method of Lyapunov: II. Discrete time systems*, Ibid., 82 (1960), pp. 394–400.

[9] K. Y. Lee, S. N. Chow and R. O. Barr, *On the control of discrete-time distributed parameter systems*, this Journal, 10 (1972), pp. 361–376.
[10] J. Zabczyk, *Remarks on the control of discrete-time distributed parameter systems*, this Journal, 12 (1974), pp. 721–733.
[11] S. K. Berberian, *Introduction to Hilbert Space*, Oxford University Press, New York, 1961.
[12] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

EXTENSIONS OF RANK CONDITIONS FOR CONTROLLABILITY AND OBSERVABILITY TO BANACH SPACES AND UNBOUNDED OPERATORS*

ROBERTO TRIGGIANI†

Abstract. Generalizations of the familiar rank conditions for controllability and observability of linear autonomous finite-dimensional systems to the general case when both the state space and the control space are infinite-dimensional Banach spaces and the operator A acting on the state is only assumed to generate a strongly continuous semigroup (group) are sought. It is shown that a suitable version of the rank condition, although generally only sufficient for approximate controllability (observability), is however "essentially" necessary and sufficient in two important cases: (i) when A generates an analytic semigroup, (ii) when A generates a group. Such generalization of the rank condition is then used to derive, in turn, easy-to-check tests for approximate controllability (observability) for the important class of normal operators with compact resolvent. In the case of finite number of scalar controls (observations), the tests are expressed by a sequence of rank conditions, using the complete set of eigenvectors of A; moreover, they imply that the minimal number of scalar controls (scalar observations) be not less than the highest multiplicity of the eigenvalues of A. Applications to heat equation as well as wave equation types of systems in finite spatial domains are included. The case when A fails to have a compact resolvent is also analyzed in two examples describing the heat equation in infinite spatial domains, by employing a general procedure.

1. Introduction.

Consider the abstract control system

$$\mathscr{L}: \dot{x} = Ax + Bu \qquad \left(\mathscr{L}_m: \dot{x} = Ax + \sum_{i=1}^{m} b_i u_i, b_i \in X, u_i = \text{scalar}\right),$$

where both X and U are complex[1], separable Banach spaces and $B \in \mathscr{B}(U, X)$, the Banach space of all bounded linear operators from U into X. Unless otherwise stated, X is always intended infinite-dimensional. $\mathscr{L}_m$ refers to the case when $\dim U = m$ or $\dim BU = m$, BU = range of B. When $m = 1$, we shall write b instead of b_1. The operator A is assumed throughout to satisfy the following assumption H1.

H1. A is (closed, linear, with domain $D(A)$ dense in X and range $\mathscr{R}(A)$ in X and) the infinitesimal generator of a strongly continuous semigroup or group (of class C_0) of bounded operators $S(t)$ [3], [9], [10], [5].

If the control function $u = u(t), t \geqq 0$, is (strongly) continuously differentiable[2], $u \in C^1[[0, T], U]$, then the Cauchy problem associated with $\mathscr{L}$ has a unique solution given by

$$x(t, x_0, u) = S(t)x_0 + \int_0^t S(t - \tau)Bu(\tau)\, d\tau \tag{1.1}$$

[1] For application of the results to the case when X and U are *real*, see remark in [8, p. 398] or in [7, p. 694].

[2] For other sufficient conditions on $Bu(t)$ to ensure that (1.1) is the unique solution, perhaps a.e., of the Cauchy problem, see [1], [2, p. 103].

* Received by the editors February 15, 1974, and in revised form November 25, 1974.

† Department of Mathematics, State University of New York at Albany, Albany, New York. Now at Department of Mathematics, Iowa State University, Ames, Iowa 50010. The first draft of this paper was written while the author was with the Center for Control Sciences, University of Minnesota, Minneapolis. Research was then continued at the Control Theory Centre, University of Warwick, England and at S.U.N.Y. at Albany. This research was supported in part by the U.S. Air Force under Grant AF-AFOSR 72-2243.

with the initial point $x_0 \in D(A)$ [10, p. 486]. Notice that the solution $x(T, x_0, u)$ at the finite time T always lies in $D(A)$ and hence (with A unbounded) cannot exhaust all of X when u runs over all $C^1[[0, T], U]$-controls. In other words, the above system $\mathscr{L}$ cannot be exactly controllable in finite time (with the state space X infinite-dimensional and A unbounded). This fact also holds for the "mild solutions" of $\mathscr{L}$, if B is compact [17, § 3], [18]. Let K_t be the set of attainability from the origin of $\mathscr{L}$, i.e., the totality of solution points $x(t, 0, u)$ as u runs over $C^1[[0, t], U]$. We then say that $\mathscr{L}$ is approximately controllable in $[0, T]$, $T < \infty$ (resp. in finite time) in case $\overline{K}_T = X$ (resp. $\bigcup_{0 \leq t} K_t = X$). Within the context of the present paper, the problem of approximate controllability was studied in the fundamental work of Fattorini [8], [7] where the terminology "complete controllability" was used instead. Necessary and sufficient conditions for approximate controllability in finite time of $\mathscr{L}_m$ (more specific than the general characterization (1.3′) below) were given in [8] under the additional assumption that (X be a Hilbert space and) A be a self-adjoint operator (or a normal operator satisfying some further properties). Fattorini's analysis uses the technical apparatus of the so-called ordered representation theory of a Hilbert space for self-adjoint (normal) operators [5, Chap. 12] and, moreover, his criterion requires the knowledge of some ordered representation. The present paper originates from a somewhat different viewpoint and employs, in particular, a less technical apparatus. It intends to explore possible generalizations (or lack thereof) to infinite-dimensional spaces of the familiar rank condition: rank $[B, AB, \cdots, A^{n-1}B] = n$, for controllability when $X = R^n$, $U = R^m$ (here approximate and exact controllability are the same concept).

If the operator A is bounded on X, we have already shown that: (i) the generalization of the rank condition is given by

$$\overline{\text{sp}}\,\{A^n BU, n = 0, 1, \cdots\} = X \quad \text{for } \mathscr{L}, BU = \text{range of } B, \tag{1.2}$$

$$(\text{resp. } \overline{\text{sp}}\,\{A^n b_i, i = 1, \cdots, m; n = 0, 1, \cdots\} = X \text{ for } \mathscr{L}_m) \tag{1.2′}$$

and characterizes approximate controllability (the length of the time interval is immaterial) [17, Thm. 3.1.1]; (ii) moreover, with B compact (and X infinite-dimensional) exact controllability in finite time for $\mathscr{L}$, even within the class of locally L_1-controls, can never arise; this holds, in particular, for $\mathscr{L}_m$ [17, § 3.3] (see also [18]).

The case when the operator A acting on the state is unbounded is known to be theoretically reduced to the bounded operator case, via the introduction of an *associated* system where the operator acting on the state is the (bounded!) resolvent $R(\cdot, A)$ [7, Prop. 2.3] (see Appendix A). (What is believed to be the first example where approximate controllability of a system with unbounded operator is characterized by studying the associated system via (1.2) is given in [17, Example 3.2.7].) In spite of this, we shall explore in § 2 the possibility of extending (an appropriate version of) (1.2) to the *original* system, in the general case when A simply generates a strongly continuous semigroup, or group, $S(t)$.

For simplicity of exposition, we summarize here only the results for $\mathscr{L}_m$, leaving those for the system $\mathscr{L}$—as well as more precise statements—to the sections to follow. We shall basically see that (1.2′) is still a sufficient condition for

approximate controllability of $\mathscr{L}_m$, but it is generally no more necessary. However, in the two important cases when either $S(t)$ is an analytic semigroup for $t > 0$, or else is a group, (1.2′) is necessary and sufficient at least for a class of vectors b_i dense in X. In § 3.2, we shall then show, as an illustration, that for the important class of normal operators with compact resolvent, such a characterization leads, in turn, to easy-to-check tests, requiring knowledge of the complete sets of eigenvectors of A. Also, such tests reveal that the minimum number of scalar controls required for approximate controllability is the highest multiplicity of the eigenvalues for A. In particular, the results for A self-adjoint with compact resolvent are in agreement with Fattorini's results [8]. These tests are checked and/or refined by also using the characterization (1.3) below involving the semigroup or the characterization in Appendix A involving $R(\cdot, A)$. Examples covering heat-equation in finite and infinite spatial domains, as well as wave-equation types of systems, are presented in § 3.1 and §4. Section 5 introduces an observation equation and treats the problem of observability, yielding results that closely parallel those on approximate controllability. In particular, the tests for observability when A is normal with compact resolvent generalize those of Sakawa [15], who treated just a special case of a particular self-adjoint operator. The following consequence of the Hahn–Banach theorem will be used throughout the paper.

PROPOSITION 1.1 [9, p. 31]. *Let X be a normal linear space and E an arbitrary set in X. Then $\overline{\text{sp}}\{E\} = X$ if and only if the zero functional is the only bounded linear functional that vanishes on E.*

It follows easily, via (1.1) and the above proposition, that $\mathscr{L}$ is approximately controllable in $[0, T]$ (in finite time) if and only if

$$x^*(S(t)BU) \equiv 0, \qquad 0 \leqq t \leqq T, \quad (0 \leqq t), \tag{1.3}$$

for all $x^* \in X^*$ (dual of X), implies $x^* = 0$; see also [8] for an equivalent formulation. In particular for $\mathscr{L}_m$, the above condition becomes

$$x^*(S(t)b_i) \equiv 0, \qquad 0 \leqq t \leqq T, \qquad (0 \leqq t), \quad i = 1, \cdots, m, \tag{1.3′}$$

for all $x^* \in X^*$ implies $x^* = 0$.

2. A general result. Before stating the general result extending (1.2) to the case when A is unbounded, we define $D_\infty(A) = \bigcap_{n=1}^{\infty} D(A^n)$ and recall that, for A satisfying H1, $D_\infty(A)$ is a subspace and is still dense in X [3, p. 12]. Also, we call[3] a vector $y \in X$ *analytic* for the semigroup (resp. group) $S(t)$ in case the map: $t \to S(t)y$ is analytic in t for $t > 0$ (resp. for $-\infty < t < \infty$). Let $\eta_a(A)$ denote the totality of analytic vectors for the semigroup (group) generated by A. In the applications that we have in mind, $S(t)$ is either an analytic semigroup for $t > 0$ or is a group. In the first case $\eta_a(A) = X$. In the second case, $\eta_a(A)$ is dense in X [9, p. 310] [23, p. 592], and moreover, the above map is infinitely many times differentiable if and only if $y \in D_\infty(A)$ (since $S(t)$ carries $D_\infty(A)$ into itself); hence $\eta_a(A) \subset D_\infty(A)$ in the group case[4]. For an arbitrary semigroup $\eta_a(A)$ need not be dense in X [23, p. 600].

Define $U_\infty = \{u \in U : Bu \in D_\infty(A)\}$, $U_a = \{u \in U : Bu \in \eta_a(A)\}$ and $U_{a\infty} = U_a$

[3] The author is indebted to Prof. S. K. Mitter for bringing to his attention reference [23].

[4] b is analytic for the *group* $S(t)$, if and only if (i) $b \in D_\infty(A)$ and (ii) $\sum_{n=0}^{\infty} A^n b t^n/n! = S(t)b$.

$\cap\, U_\infty$, i.e., $U_\infty, U_a, U_{a\infty}$ are the largest subspaces of U such that $BU_\infty \subset D_\infty(A)$, $BU_a \subset \eta_a(A), BU_{a\infty} \subset D_\infty(A) \cap \eta_a(A)$, respectively. We have: $BU_{a\infty} = BU_\infty$, if $S(t)$ is an analytic semigroup and $BU_{a\infty} = BU_a$ if $S(t)$ is a group.

We also list here, for convenience, another hypothesis for A, to which we shall refer in the sequel.

H2. The range of $S(t)$ in X belongs to $D(A)$ for each $t > 0$ (differentiable semigroup [28, p. 50]).

It is then known [3, p. 15], that, under assumptions H1 and H2, $S(t)$ is n times continuously differentiable in the uniform operator topology for $0 < t < \infty$ and all integers $n > 0$, and $d^nS(t)/dt^n = A^nS(t) \in \mathscr{B}(X)$. Assumption H2 is weaker than analyticity of $S(t)$ [3, p. 16].

The sought for generalization is provided by the following theorem, where, unless otherwise stated, n runs as follows, $n = 0, 1, 2, \cdots$.

THEOREM 2.1. *Let A satisfy* H1. *A sufficient condition for $\mathscr{L}$ to be approximately controllable on $[0, T]$ is given by*

$$\overline{\text{sp}}\,\{A^nBU_\infty\} = X \tag{2.1}$$

or, more generally, by

$$\overline{\text{sp}}\,\{A^nS(\bar{t})BU_\infty\} = X, \quad \bar{t} \text{ arbitrary in } [0, T]. \tag{2.1'}$$

When A also satisfies H2, *then* (2.1') *can be relaxed as to replace BU_∞ by BU, with $\bar{t}$ arbitrary in $(0, T]$.*

Conversely, assume that $BU_{a\infty}$ is dense in BU. Then a necessary condition for $\mathscr{L}$ to be approximately controllable on $[0, T]$ is given by

$$\overline{\text{sp}}\,\{A^nS(\bar{t})BU_{a\infty}\} = X, \quad \bar{t} \text{ arbitrary} > 0. \tag{2.2}$$

If $S(t)$ is an analytic semigroup for $t > 0$, then (2.2) *can be relaxed as to replace $BU_{a\infty}$ $(= BU_\infty)$ by BU. Also, if $S(t)$ is a group, $\bar{t}$ in* (2.2) *can be any real number, in particular $\bar{t} = 0$, and* (2.2) *simplifies, in this case, $(BU_{a\infty} = BU_a)$, to*

$$\overline{\text{sp}}\,\{A^nBU_a\} = X. \tag{2.2'}$$

Remark 2.1. As stated below, $A^nS(t) = S(t)A^n$ on $D_\infty(A)$; moreover, if assumption H2 also holds, then $\overline{A^nS(t)}$ is a bounded operator on X for $t > 0$ and so coincides with the closure $\overline{S(t)A^n}$.

The important case with a finite number of scalar controls is singled out in the following corollary, where, unless otherwise stated, n and i run as follows: $n = 0, 1, 2 \cdots ; i = 1, \cdots, m$.

COROLLARY 2.2. *Let A satisfy* H1. *A sufficient condition for $\mathscr{L}_m$ to be approximately controllable on $[0, T]$ is given by*

$$\overline{\text{sp}}\,\{A^nb_i\} = X, \qquad b_i \in D_\infty(A), \tag{2.3}$$

or, more generally, by

$$\overline{\text{sp}}\,\{A^nS(\bar{t})b_i\} = X, \qquad b_i \in D_\infty(A), \quad \bar{t} \text{ arbitrary in } [0, T]. \tag{2.3'}$$

When A also satisfies H2, *then the b_i's in* (2.3') *can be relaxed to be any vectors in X with $\bar{t}$ in $(0, T]$.*

Conversely, a necessary condition for $\mathscr{L}_m$ to be approximately controllable on $[0, T]$ is given by

(2.4) $\overline{\text{sp}}\{A^nS(\bar{t})b_i\} = X, \quad \bar{t}$ *arbitrary* $> 0, \qquad$ when $\quad b_i \in \eta_a(A) \cap D_\infty(A)$.

If $S(t)$ *is an analytic semigroup for* $t > 0$, *then the* b_i*'s in* (2.4) *can be relaxed to be any vectors in* X. *Also, if* $S(t)$ *is a group,* $\bar{t}$ *in* (2.4) *can be any real number, in particular* $\bar{t} = 0$, *and* (2.4) *simplifies, in this case, to*

(2.4′) $$\overline{\text{sp}}\{A^nb_i\} = X \quad \text{when} \quad b_i \in \eta_a(A).$$

Proof of Theorem 2.1. First of all, we need the following facts. Let $y \in D_\infty(A)$. Then $S(t)y$ can be differentiated infinitely many times in t and from [3, (1.1.9), Prop. 1.1.6, p. 11] it follows that

(2.5) $$\frac{d^nx^*(S(t)y)}{dt^n} = x^*(S(t)A^ny) = x^*(A^nS(t)y), \qquad t \geqq 0, \quad n = 0, 1, 2, \cdots,$$

for $x^* \in X^*$. However, under assumption H2, we have for all $t > 0$,

(2.5′) $$\frac{d^nS(t)}{dt^n} = A^nS(t) \in \mathscr{B}(X), \qquad n = 0, 1, \cdots,$$

[3, pp. 15–16], and so $A^nS(t)$ can be applied on all of X.

In view of the characterization (1.3), we need only show that (2.1′) implies (1.3); conversely, (1.3) implies (2.2). (2.1′) $\Rightarrow$ (1.3). Suppose by contradiction that $\bar{x}^*(S(t)BU) \equiv 0$ and hence $\bar{x}^*(S(t)BU_\infty) \equiv 0, 0 \leqq t \leqq T$, for some nonzero $\bar{x}^* \in X^*$. Differentiate successively this last identity to show, by induction, using (2.5), that $\bar{x}^*(A^nS(t)BU_\infty) \equiv \bar{x}^*(S(t)A^nBU_\infty) \equiv 0$, $0 \leqq t \leqq T$, $n = 0, 1, \cdots$. Set $t = \bar{t}$ to get, via Proposition 1.1 and the fact that $\bar{x}^*$ is nonzero, a contradiction with (2.1′). (2.1) is obtained for $\bar{t} = 0$. The statement for A satisfying H2 is obtained similarly using (2.5′) instead of (2.5). (1.3) $\Rightarrow$ (2.2): Suppose that (2.2) is false and so, by Proposition 1.1, there is a nonzero $\bar{x}^* \in X^*$ such that $\bar{x}^*(A^nS(\bar{t})BU_{a\infty}) = 0$, $n = 0, 1, \cdots$. By analyticity in t of $S(t)BU_{a\infty}$, it follows, via (2.5), that $\bar{x}^*(S(t)BU_{a\infty}) \equiv 0$ in a neighborhood of $\bar{t}$, and hence for all $t \geqq 0$. Since $BU_{a\infty}$ is dense in BU, it follows by continuity that $\bar{x}^*(S(t)BU) \equiv 0$, $t \geqq 0$, which contradicts (1.3). Q.E.D.

Remark 2.2. When A is bounded and so generates a uniformly continuous analytic *group* $S(t) = \exp(At)$, $-\infty < t < \infty$, $\bar{t}$ in the above theorem and corollary can be taken to be zero. Hence $S(\bar{t}) = S(0) = I$ (identity) and the conditions (2.1) and (2.3) are necessary and sufficient, with $D_\infty(A) = \eta_a(A) = X, U_\infty = U_a = U$ and any $b_i \in X$. Theorem 2.1 reduces therefore to the condition (1.2).

The point is, however, that when $S(t)$ is an analytic semigroup, $t > 0$, in the above *necessary* conditions $\bar{t}$ can be taken to be zero, only when the generator A is bounded, since analytic groups generated by unbounded operators do not exist. The only place in the literature where we were able to find such an assertion is [9, p. 278, also p. 477], as part of a sophisticated treatment on holomorphic semigroups. We therefore sketch here a quick proof of this fact, even under weaker assumptions. Let $S(t)$ be (i) a strongly continuous group of bounded operators on X, (ii) differentiable for $t > \tau \geqq 0$ ($\Leftrightarrow S(t)X \subset S(A)$, for $t > \tau$). Then its infinitesimal generator A is a bounded operator on X. In fact, the closed operator $AS(t)$ is well-defined on X for $t > \tau$ and, by the closed graph theorem, is bounded on X. But then the operator $[AS(t)]S(-t) = A$ is also bounded on X.

Remark 2.3. For simplicity of notation, we limit our comments to $\mathscr{L}_m$, the extension to $\mathscr{L}$ being immediate.

(i) Let $S(t)$ be a group: a characterization for approximate controllability on $[0, T]$ of $\mathscr{L}_m: \langle A, (b_1, \cdots, b_m)\rangle$ is then given by $\overline{\text{sp}}\{A^n b_i\} = X$, at least for vectors b_i in the set $\eta_a(A)$ dense in X.

(ii) Let $S(t)$ be an analytic semigroup, $t > 0$: the same characterization is then given by $\overline{\text{sp}}\{A^n S(\bar{t}) b_i\} = X$, any $b_i \in X$, $\bar{t}$ arbitrary *positive* time. Also let $X_0 = \cup_{0<t} S(t)X$ ($\subset D_\infty(A)$ [3, p. 11]), i.e., $b_i \in X_0$ in case $b_i = S(\tilde{t}_i)\tilde{b}_i$, for some $\tilde{b}_i \in X$ and $\tilde{t}_i > 0$. X_0 is dense in X [9, p. 208]. Now let $b_i \in X_0$: then due to the arbitrariness of $\bar{t}$ in (ii) above (see also Remark 2.5)

$$\overline{\text{sp}}\{A^n b_i\} = \overline{\text{sp}}\{A^n S(\tilde{t})\tilde{b}_i\} = X \Leftrightarrow \overline{\text{sp}}\{A^n S(t) b_i\} = X, \quad t \text{ and } \tilde{t} \text{ arbitrary} > 0,$$

$$\Updownarrow \qquad\qquad \Updownarrow$$

$\langle A, (\tilde{b}_1, \cdots, \tilde{b}_m)\rangle$ approximately controllable on $[0, T]$ — $\langle A, (b_1, \cdots, b_m)\rangle$ approximately controllable on $[0, T]$.

Hence, for a semigroup analytic for $t > 0$, the condition $\overline{\text{sp}}\{A^n b_i\} = X$ is necessary and sufficient for approximate controllability on $[0, T]$ of $\langle A, (b_1, \cdots, b_m)\rangle$ at least for vectors b_i in the set X_0 dense in X.

Generally speaking, however, even with $S(t)$ analytic semigroup, ($\eta_a(A) = X$), (2.3) and (2.4) are not the same (see examples in Remark 2.4). Notice also that $S(t)$ is never a homeomorphism on X (one-to-one, onto) in the present case (which would then imply the equivalence of (2.3) and (2.4)): this follows from [9, Thm. 16.7.5, p. 470] which is not applicable to an analytic semigroup generated by an unbounded operator (see Remark 2.2) in which case we have $0 \notin \rho(S(t))$, $t > 0$.

Remark 2.4. The following example (heat equation in an infinite rod) shows that, when A is an unbounded generator of a semigroup which is not a group, condition (2.3) need not be necessary for approximate controllability, even in the analytic case and with A self-adjoint if $b_i \notin X_0$: hence (2.3) and (2.4) are in general not equivalent for $S(t)$ analytic, $t > 0$.

Let $X = L_2[-\infty, \infty]$, $Af = d^2f/d\xi^2$ (in the sense of distributions) with $D(A) = \{f : f'' \in L_2[-\infty, \infty]\}$ as in Example 1 in [8]. (Notice that in such a case A is self-adjoint but does not have a compact resolvent and in fact $\sigma(A) = (-\infty, 0]$, [8], [5, p. 639]: compare with § 3.2.) Using the ordered representation theory, Fattorini showed that the minimum number of controls to make A approximately controllable is two and this happens, for instance, when $b_1(\xi)$ is different from zero and has compact support and $b_2(\xi) = b_1(\xi - h)$, $h \neq 0$. We shall rederive such a result (actually its generalization) in § 3.1, using our Corollary 2.2. Choose in particular $b_1(\xi)$ to vanish identically outside a finite interval, and to be arbitrarily smooth in the interval as to define a C^∞-function on $[-\infty, \infty]$. Say: $b_1(\xi) = \exp[(\xi^2 - 1)]^{-1}$, $-1 < \xi < 1$, and 0 for $|\xi| \geqq 1$. Then $b_1(\cdot)$ and $b_2(\cdot)$ belong to $D_\infty(A)$ and, moreover, they vanish identically together with all their derivatives outside $[-1, 1 + h]$ for $h > 0$ ($[-1 + h, 1]$ for $h < 0$). Hence, in this case with $m = 2$, the left-hand side of (2.3) is a proper subspace of $L_2[-\infty, \infty]$ and yet the system defined by A, b_1, b_2 is approximately controllable on $[0, T]$. Notice that

$b_i \notin X_0$ (see Remark 2.3(ii)). One can also verify directly that the necessary condition (2.4) is, in this case, indeed satisfied, in agreement with Fattorini's result. See a more complete treatment of this example in § 3.1. A similar counter-example on the necessity of the condition (2.3) can be obtained by using the self-adjoint operator A with no eigenvalues, defined in Example 3 in [8, p. 401, 2nd line from the top].

Remark 2.5. The above proof of Theorem 2.1 also shows that the subspace $\overline{\text{sp}}\{A^n S(t)BU_{a\infty}\}$ is independent on t for $-\infty < t < \infty$ if $S(t)$ is a group, and for $t > 0$ if $S(t)$ is only a semigroup. Similarly, if $S(t)$ is an analytic semigroup, $t > 0$, the subspace $\overline{\text{sp}}\{A^n S(t)BU\}$ is independent on t for $t > 0$.

Remark 2.6. One can obviously write (2.2) as (see Remark 2.1)

$$\text{Cl sp}\{S(\bar{t})A^n BU_{a\infty}\} = \text{Cl } S(\bar{t}) \text{ sp}\{A^n BU_{a\infty}\} = \text{Cl } S(\bar{t})\{\text{Cl sp}\{A^n BU_{a\infty}\}\}$$

since $S(\bar{t})$ is bounded. (Here the closure has been denoted by Cl.) Now let $S(t)$ be analytic for $t > 0$. In this case ($BU_{a\infty} = BU_{\infty}$), if (2.1) and (2.2) both hold, then the range of $S(\bar{t})$ is necessarily dense in X and so $\lambda = 0$ cannot belong to $R\sigma(S(\bar{t}))$, the residual spectrum of $S(\bar{t})$.

Recalling Remark 2.3(ii) above, one concludes: when $S(t)$ is analytic, $t > 0$, and (2.1) is equivalent to (2.2), then $\lambda = 0$ is neither in the resolvent set nor in the residual spectrum of $S(t)$, $t > 0$; conversely, if (2.1) holds and the range of $S(\bar{t})$ is dense in X, then (2.2) follows: this is the case, e.g., with A normal, to which we shall turn in § 3.2.

3. Illustrations. The present section deals with the application of the general result of § 2, the extension of the rank condition, to classes of examples of physical interest. Our illustrations cover, in particular, all the self-adjoint examples treated in [8], for which necessary and sufficient conditions were given, using the ordered representations of Hilbert spaces, plus knowledge of such a representation in each single case. We rederive such conditions using our Corollary 2.2. Moreover, our illustrations include also hyperbolic systems.

3.1. Two self-adjoint examples in infinite spatial domains.

Example 3.1 (heat equation in an infinite slab) [8, example 1, p. 399]. Let $X = L_2[-\infty, \infty]$, $Af = d^2f/d\xi^2$ with $D(A) = \{f \in X, Af \in X\}$, where Af is understood in the sense of distributions. A is self-adjoint, and so the associated semigroup is self-adjoint and analytic for $t > 0$ [9, p. 588]. We shall now use the extension of the rank condition, Corollary 2.2 (see also Remark 2.3(ii)), to derive that

(a) the minimal number of scalar controls which make A approximately controllable is two;

(b) the pair $\langle A, (b_1, b_2)\rangle$, $b_i \in X$ is approximately controllable if and only if

$$\text{(A)} \qquad \hat{b}_1(\omega)\hat{b}_2(-\omega) - \hat{b}_1(-\omega)\hat{b}_2(\omega) \neq 0 \qquad \text{a.e. in } \omega \geqq 0,$$

Here

$$\hat{f}(\omega) = \underset{N\to\infty}{\text{l.i.m.}} (2\pi)^{-1/2} \int_{|\xi| \leqq N} e^{i\omega\xi} f(\xi)\, d\xi$$

is the Fourier–Plancherel transform (isometric isomorphism of $L_2[-\infty, \infty]$ onto itself.) [21, Corollary VI.2, p. 154]. These results are in agreement with [8], where

they were derived using instead the ordered representation theory of X with respect to A. Our proof proceeds through the following steps. First let b_i, $i = 1, \cdots, m$, be arbitrary vectors in X and $\bar{t}$ be an arbitrary positive time.

(i) We have the known fact that

$$\widehat{[S(\bar{t})b_i]}(\omega) = e^{-\omega^2 \bar{t}} \hat{b}_i(\omega),$$

as it follows by solving the homogeneous equation by Fourier–Plancherel transform.

(ii) Also, $f \in D(A)$ if and only if $\omega^2 \hat{f}(\omega) \in L_2[-\infty, \infty]$ and

$$\widehat{Af}(\omega) = -\omega^2 \hat{f}(\omega)$$

[8, p. 399], [26, Chap. I, 1.7].

(iii) But $S(t)X \subset D(A)$, $t > 0$, since $S(t)$ is analytic for $t > 0$ [3, pp. 15–16]. Hence

$$A^n S(t) b_i = S(t/2) A^n S(t/2) b_i \in D(A), \qquad t > 0.$$

(iv) Applying step (ii) repeatedly yields

$$\widehat{[A^n S(\bar{t}) b_i]}(\omega) = (-1)^n \omega^{2n} e^{-\omega^2 \bar{t}} \hat{b}_i(\omega).$$

(v) Since the Fourier–Plancherel transform defines an isometric isomorphism of X onto itself [21, p. 154], we have that

$$(A^n S(\bar{t}) b_i, g) = 0, \qquad n = 0, 1, \cdots, \quad i = 1, \cdots, m, \quad g \in X,$$
$$\Rightarrow g = 0$$

if and only if

$$(\widehat{A^n S(\bar{t}) b_i}, \hat{g}) = 0, \qquad n = 0, 1, \cdots, \quad i = 1, \cdots, m,$$
$$\Rightarrow \hat{g} = 0$$

with $(\cdot, \cdot)$ inner product on X.

(vi) We then have that

$$0 = (\widehat{A^n S(\bar{t}) b_i}, \hat{g}) = (-1)^n \int_{-\infty}^{\infty} \omega^{2n} e^{-\omega^2 \bar{t}} \hat{b}_i(\omega) \bar{\hat{g}}(\omega)\, d\omega$$

$$= (-1)^n \int_0^{\infty} \omega^{2n} e^{-\omega^2 \bar{t}} [\hat{b}_i(-\omega) \bar{\hat{g}}(-\omega) + \hat{b}_i(\omega) \bar{\hat{g}}(\omega)]\, d\omega$$

implies

$$\text{(B)} \qquad \hat{b}_i(-\omega) \bar{\hat{g}}(-\omega) + \hat{b}_i(\omega) \bar{\hat{g}}(\omega) \equiv 0 \qquad \text{a.e. in } \omega \geqq 0$$

[19, p. 62] [27, p. 107] (the value of $\bar{t}$ is immaterial as long as $\bar{t} > 0$).

(vii) It is then easily seen that, for $m = 1$, the identity (B) in (vi) does not imply $\hat{g}(\omega) \equiv 0$ a.e. $-\infty < \omega < \infty$.

Also, for $m = 2$, the identity (B) in (vi) can be written as

$$\begin{vmatrix} \hat{b}_1(\omega) & \hat{b}_1(-\omega) \\ \hat{b}_2(\omega) & \hat{b}_2(-\omega) \end{vmatrix} \begin{vmatrix} \hat{g}(\omega) \\ \bar{\hat{g}}(-\omega) \end{vmatrix} \equiv 0 \quad \text{a.e. in } \omega \geqq 0$$

and implies $\hat{g}(\omega) \equiv 0$ a.e. $-\infty < \omega < \infty$ if and only if (A) holds. Claims (a) and (b) are proved.

As remarked in [8], it is easy to check that (A) holds if $b_1(\xi)$ is different from zero and has compact support and $b_2(\xi) = b_1(\xi - h)$, with $h \neq 0$. This is the special case we have used in Remark 2.4 (which of course can also be proved through a direct use of the extension of the rank condition, i.e., first implication in step (v), without introducing the Fourier–Plancherel transform.)

Notice that the same procedure used above, when applied to the space $L_2(R^r), r \geqq 2$, shows that the heat equation on $R^r, r \geqq 2$, is never approximately controllable, if it is acted upon only by a finite number (no matter how large) of *scalar* controls, in agreement with [8, example 1]. For instance, for R^3, the analogs of steps (i) and (ii) above are now

$$\widehat{[S(\bar{t})b_i]}(\omega_1, \omega_2, \omega_3) = e^{-(\omega_1^2+\omega_2^2+\omega_3^2)\bar{t}}\hat{b}_i(\omega_1, \omega_2, \omega_3)$$

and

$$\widehat{[A^n S(\bar{t})b_i]}(\omega_1, \omega_2, \omega_3) = -(\omega_1^2 + \omega_2^2 + \omega_3^2)^n \widehat{[S(\bar{t})b_i]}(\omega_1, \omega_2, \omega_3),$$

respectively. However, the implication in step (vi) above now fails, since the functions $\{(\omega_1^2 + \omega_2^2 + \omega_3^2)^n e^{-(\omega_1^2+\omega_2^2+\omega_3^2)}\}, n = 0, 1, 2, \cdots$, are not complete in $L_2(R_+^3)$, R_+^3 being the positive 3-space.

Example 3.2 (heat equation in a semi-infinite slab) [8, example 2, p. 400]. Let $X = L_2[0, \infty]$, $Af = d^2f/d\xi^2$ with $D(A) = \{f, Af \in X, f(0) = 0\}$. A is self-adjoint [5, p. 1384], with $\sigma(A) = (-\infty, 0]$ (so that A does not have a compact resolvent) and generates a self-adjoint analytic semigroup for $t > 0$ [9, pp. 588–589]. We shall now use the extension of the rank condition, Corollary 2.2, (see also Remark 2.3 (ii)), to derive that: the pair $\langle A, b\rangle$ is approximately controllable if and only if

$$\tilde{b}(\omega) \neq 0 \quad \text{a.e. in } \omega \geqq 0.$$

Here

$$\tilde{f}(\omega) = \underset{N\to\infty}{\text{l.i.m.}} (2/\pi)^{1/2} \int_0^N (\sin \omega\xi) f(\xi)\, d\xi$$

is the Fourier sine transform (isometric isomorphism of $L_2[0, \infty]$ onto itself [5, p. 1388]). This same result was proved in [8], using instead the ordered representation theory of X with respect to A.

Our proof follows the pattern used in the previous examples 3.1. Let b be an arbitrary vector in X and $\bar{t}$ be an arbitrary positive time. We have the known fact that

$$\widetilde{[S(\bar{t})b]}(\omega) \doteq e^{-\omega^2\bar{t}}\tilde{b}(\omega)$$

($\doteq$ is the proportionality sign) as it follows by solving the homogeneous equation by the Fourier-sine transform. Moreover, from

$$A^n S(t)b = S(t/2)A^n S(t/2)b \in D(A), \qquad t > 0, \quad n = 0, 1, \cdots,$$

(implied by the analyticity of $S(t)$ [3, pp. 15–16]) and from

$$\widetilde{(Af)}(\omega) = -\omega^2 \tilde{f}(\omega)$$

[5, p. 1388], it follows that

$$[\widetilde{A^n S(\bar{t})b}](\omega) \doteq \omega^{2n} e^{-\omega^2 \bar{t}} \tilde{b}(\omega), \qquad n = 0, 1, \cdots.$$

Also, since the Fourier-sine transform is an isometric isomorphism of X onto itself, we have that

$$(A^n S(\bar{t})b, g) = 0, \qquad n = 0, 1, \cdots, \quad g \in X,$$
$$\Rightarrow g = 0$$

if and only if

$$(\widetilde{A^n S(\bar{t})b}, \tilde{g}) = 0, \qquad n = 0, 1, \cdots,$$
$$\Rightarrow \tilde{g} = 0,$$

where $(\cdot, \cdot)$ is the inner product on X. Then

$$0 = (\widetilde{A^n S(\bar{t})b}, \tilde{g}) \doteq \int_0^\infty \omega^{2n} e^{-\omega^2 \bar{t}} \tilde{b}(\omega) \bar{\tilde{g}}(\omega)\, d\omega$$

implies easily $\tilde{b}(\omega)\bar{\tilde{g}}(\omega) \equiv 0$ a.e. in $\omega \geqq 0$ [19, p. 62] [27, p. 107]. (The value of $\bar{t}$ is immaterial as long as $\bar{t}$ is positive). This, in turn, implies $\tilde{g}(\omega) \equiv 0$ a.e. in $\omega \geqq 0$ if and only if $\tilde{b}(\omega) \neq 0$ a.e. in $\omega \geqq 0$. Our claim is proved.

It appears that the procedure followed in the above two examples is quite general. Moreover, it shows that knowledge of $S(\bar{t})$ is not necessarily needed in order to apply the characterization $\overline{\text{sp}}\{A^n S(\bar{t})b_i\} = X, \bar{t} > 0$, for approximate controllability on $[0, T]$ in the analytic case. See also § 3.2.

3.2. The case when A is a normal operator with compact resolvent. Throughout the present section X will be specialized to be a Hilbert space and the operator A is assumed to satisfy, in addition to H1, one or both of the following assumptions.

H3. A is normal and $R(\mu_0, A)$ is compact as an operator on X for some μ_0 (this implies $R(\mu, A)$ compact for all μ in $\rho(A)$ [10, p. 187]);

H4. the eigenvalues of A are contained in some sector $\Sigma = \{\lambda; |\arg(\lambda - a)| \leqq \pi/2 + \beta\}$ a real, $0 < \beta < \pi/2$. (This is automatically true, if A is self-adjoint and satisfies H1.)

The semigroup of A satisfying H1 and H3 is normal (self-adjoint, if A is self-adjoint) and given explicitly below in (3.1); it is analytic if and only if H4 also holds, in which case the domain of analyticity is the sector $|\arg \lambda| < \beta$. See [9, pp. 589–99], [21, pp. 254–9]. So H4 implies H2, but not conversely. Under H2, the spectrum of the generator is contained in a logarithmic sector [28, Thm. 4.9, p. 57]. The class of operators satisfying H1 and H3 usually arises in classical boundary problems [10, p. 187].

In this section, we shall purposely use the general results of § 2 to derive explicitly verifiable tests for approximate controllability on $[0, T]$ for the class of operators satisfying H1, H3, H4, expressed solely in terms of the coordinates of b_i with respect to the natural basis of the eigenvectors of A. Such conditions agree with those derived by Fattorini in [8, example 4] when A is self-adjoint using the ordered representation theory or in [7, Cor. 3.3] using the theory of spectral sets. Extensions and/or refinements of such tests, when H4 is removed as to cover the

import case of unitary groups (wave equation) are also included using either the characterization (1.3) based on $S(t)$ or the characterization in Appendix A, based on $R(\cdot, A)$. In all cases, the obtained tests are a natural generalization to the present situation of finite-dimensional facts, as explained in Remark 3.2.

In view of the assumption H3, the following holds [10, p. 277] (see also [13, p. 487] [16, p. 343] [5, p. 1330]).

(a) There is an infinite sequence $\{\lambda_j\}$, $j = 1, 2, \cdots$, of distinct isolated eigenvalues of A, $|\lambda_j| \to \infty$, $j \to \infty$, each with finite multiplicity r_j equal to the dimensionality of the corresponding eigenmanifold. In view of assumption H1, the eigenvalues λ_j have real parts uniformly bounded above; moreover the spectrum $\sigma(A)$ of A consists only of the λ_j's (point spectrum).

(b) There is a correspondent complete orthonormal set $\{x_{jk}\}$ of eigenvectors of A, $k = 1, \cdots, r_j$.

(c) From the (unique) expansion $x = \sum_{j=1}^{\infty} \sum_{k=1}^{r_j} (x, x_{jk}) x_{jk}$ one gets

$$Ax = \sum_{j=1}^{\infty} \lambda_j \sum_{k=1}^{r_j} (x, x_{jk}) x_{jk} \quad \text{for } x \in D(A) = \{x \in X : \sum_{j=1}^{\infty} |\lambda_j|^2 \sum_{k=1}^{r_j} |(x, x_{jk})|^2 < \infty\}.$$

(d) for λ not in $\sigma(A)$ and each y in X we have[5]

$$R(\lambda, A)y = (\lambda - A)^{-1} y = \sum_{j=1}^{\infty} \frac{1}{\lambda - \lambda_j} \sum_{k=1}^{r_j} (y, x_{jk}) x_{jk}.$$

Such a resolvent is compact.

One then verifies that the semigroup $S(t)$ is given by

$$S(t)x = \sum_{j=1}^{\infty} e^{\lambda_j t} \sum_{k=1}^{r_j} (x, x_{jk}) x_{jk}, \qquad t \geqq 0, \quad x \in X. \tag{3.1}$$

By induction, one finds from (c), that[6] ($n = 0, 1, 2, \cdots$; $i = 1, \cdots, m$)

$$(A^n b_i, x_{pq}) = \lambda_j^n (b_i, x_{pq}), \tag{3.2}$$

$$A^n b_i = \sum_{j=1}^{\infty} \lambda_j^n \sum_{k=1}^{r_j} (b_i, x_{jk}) x_{jk}, \qquad b_i \in D_\infty(A). \tag{3.3}$$

Now let $\bar{t} > 0$. From (3.1) and (3.3) one gets

$$A^n S(\bar{t}) b_i = \sum_{j=1}^{\infty} \lambda_j^n e^{\lambda_j \bar{t}} \sum_{k=1}^{r_j} (b_i, x_{jk}) x_{jk}, \qquad b_i \in D_\infty(A), \tag{3.4}$$

or any $b_i \in X$, if H2 (in particular, if H4) holds.

Remark 3.1. It follows directly from (3.1) that $\lambda = 0 \notin P\sigma(S(t))$, $t > 0$. Hence $S^{-1}(t)$ exists on $\mathscr{R}(S(t))$, $t \geqq 0$, and in fact,

$$S^{-1}(t)y = \sum_{j=1}^{\infty} e^{-\lambda_j t} \sum_{k=1}^{r_j} (y, x_{jk}) x_{jk}, \qquad y \in \mathscr{R}(S(t)), \quad t \geqq 0,$$

[5] Conversely, if $R(\lambda, A)$ satisfies (d) with $\{x_{jk}\}$ an orthonormal set and $1/|\lambda - \lambda_j| \to 0$, then $R(\lambda, A)$ is normal and compact [19, p. 208], in particular, self-adjoint if λ and λ_j are real [16, p. 342].

[6] Here read $\lambda_j^0 = 0$, if $\lambda_j = 0$ for some j.

as one can check directly. From the general theory [16, p. 251] one has $\mathrm{Cl}\,\mathscr{R}(S(\bar{t})) = X$. Notice that for A satisfying also H2, $S^{-1}(t), t > 0$, is unbounded since $-\operatorname{Re}\lambda_j \to \infty$ as $j \to \infty$. Hence $\lambda = 0 \in C\sigma(S(t))$. (See Remark 2.6.)

In the following, we shall need to refer to the following definition: $\sup r_j \triangleq r \leqq \infty$; i.e., r is the highest multiplicity (if $r < \infty$) of the eigenvalues of A. If $r = 1$, we simply write $\{x_j\}$ instead of $\{x_{j1}\}$. Also, for each j and each m-tuple b_i of vectors in X, we define the $r_j \times m$ matrix B_j as follows:

$$B_j = \begin{vmatrix} (b_1, x_{j1}), \cdots, (b_m, x_{j1}) \\ (b_1, x_{j2}), \cdots, (b_m, x_{j2}) \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ (b_1, x_{jr_j}), \cdots, (b_m, x_{jr_j}) \end{vmatrix}$$

and denote by B_j^T its transpose.

Remark 3.2. Before proving the results of this section, we shall illustrate them in the finite-dimensional case, $\dim X = n$, within the framework of the classical theory and linear algebra. Let A be a normal operator defined on R^n. Then a basis of eigenvectors can be chosen, so that the matrix representing A (denoted with the same symbol) is diagonal, with each diagonal element an eigenvalue: $A = \operatorname{diag}[\lambda_1, \cdots, \lambda_n]$. If b is a vector in R^n with coordinates $b^1, \cdots, b^n$, then

$$\det[b, Ab, \cdots, A^{n-1}b] = b^1 \cdot b^2 \cdots b^n \cdot V(\lambda_1, \cdots, \lambda_n),$$

where $V(\lambda_1, \cdots, \lambda_n)$ is the Vandermonde determinant of the numbers $\lambda_1, \cdots, \lambda_n$. Therefore

(i) If $\lambda_1, \cdots, \lambda_n$ are distinct, it follows that the pair $\langle A, b\rangle$ is controllable if and only if: $(*) b^i \neq 0$, $i = 1, \cdots, n$. So, in this case, A can be made controllable with just one scalar control.

(ii) If $\lambda_1, \cdots, \lambda_n$ are not all distinct, one needs more than one scalar control in order to make A controllable. More precisely, let $\lambda_1, \cdots, \lambda_s$ be the distinct eigenvalues of A with multiplicity $r_1, \cdots, r_s$, $\sum r_j = n$ and $r = \max r_j$. Denote by $\{x_{jk}\}$ the basis of correspondent eigenvectors $j = 1, \cdots, s$; $k = 1, \cdots, r_j$. Let $b_1, \cdots, b_m$ be the column vectors of the $n \times m$ matrix B and partition B in blocks B_j of dimensions $r_j \times m$, where B_j is defined as before. Then the matrix B and the controllability matrix $[B, AB, \cdots, A^{n-1}B]$ are given by

$$B = \begin{vmatrix} B_1 \\ B_2 \\ \vdots \\ B_s \end{vmatrix},$$

$$[B, AB, \cdots, A^{n-1}B] = \begin{vmatrix} B_1 & \lambda_1 B_1 \cdots \lambda_1^{n-1} B_1 \\ B_2 & \lambda_2 B_2 \cdots \lambda_2^{n-1} B_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ B_s & \lambda_r B_s \cdots \lambda_r^{n-1} B_s \end{vmatrix},$$

respectively, and the latter has full rank n if and only if: (**) rank $B_j = r_j$, $j = 1, \cdots, s$. (Such results (*) and (**) are believed to be known, although we could explicitly find in the literature only (*) for A symmetric [12, p. 102].) We now proceed to generalize (*) and (**) to infinite-dimensional state spaces.

First, for vectors b_i, $i = 1, \cdots, m$ in X, consider the following condition:

$$\text{rank } B_j = r_j, \qquad j = 1, 2, \cdots, \tag{3.5}$$

which in turn implies $m \geqq r$.

For $m = 1$, (3.5) means that all the coordinates of b are nonzero:

$$(b, x_j) \neq 0, \qquad j = 1, 2, \cdots. \tag{3.5'}$$

We start with a proposition in the case when A is a normal compact operator to get a result which is of interest in itself and which will also be applied in the sequel to the unbounded operator case.

PROPOSITION 3.1. *Let A be a normal compact operator for which zero is not an eigenvalue. Then $\mathscr{L}_m : \langle A, (b_1, \cdots, b_m)\rangle$ is approximately controllable on $[0, T]$ if and only if* (3.5) *holds.*

Proof. Denote by $\{\mu_j\}$ the eigenvalues of A of multiplicity r_j. The associated eigenvectors $\{x_{jk}\}, j = 1, 2, \cdots, k = 1, \cdots r_j$, form a basis for X and we have

$$A^n b_i = \sum_{j=1}^{\infty} \mu_j^n \sum_{k=1}^{r_j} (b_i, x_{jk}) x_{jk}, \qquad n = 0, 1, \cdots, \quad i = 1, \cdots, m, \tag{3.6}$$

in agreement with (d) above [19, pp. 207–208]. By (1.2), we must show that $N = \overline{\text{sp}}\,\{A^n b_i\}$ is the whole space X if and only if (3.5) holds.

Only if: if, by contradiction, rank $B_j^T < r_j$ for some j (B_j^T transpose of B_j), i.e., sp $B_j^T \subsetneqq X_j$ (X_j-eigenspace spanned by $x_{j1}, \cdots, x_{jr_j}$); it then follows from (3.6) that $N \subsetneqq X$.

If: let, instead, $N \subsetneqq X$. N is invariant for A and so $N^\perp$ is invariant for the normal compact operator A^*, hence for $(A^*)^* = A$ [24, Thm. 7]. Hence [24, Thm. 1] $N^\perp$ contains an eigenvector $x_{j\bar{k}}$ of A. Equation (3.6) then yields

$$0 = (A^n b_i, x_{j\bar{k}}) = \mu_j^n (b_i, x_{j\bar{k}})$$

with $\mu_j \neq 0$, $i = 1, \cdots, m$, and so rank $B_j < r_j$: a contradiction. Q.E.D.

COROLLARY 3.2. *Let A be as in Proposition* 3.1. *Then $\mathscr{L}_1 : \langle A, b\rangle$ is approximately controllable on $[0, T]$ if and only if* (3.5') *holds.*

Actually, the above proof of Proposition 3.1 only requires that A be normal and have a set of eigenvectors forming a basis, so that (3.6) holds and zero not be an eigenvalue (e.g., A = normal compact operator and identity).

We next apply the above results to the general unbounded case.

PROPOSITION 3.3. *Let A satisfy* H1, H3 *and* H4. *Then* $\overline{\text{sp}}\,\{A^n S(\bar{t}) b_i\} = X$, $\bar{t} > 0$, $b_i \in X$, *if and only if* (3.5) *holds.*

Proof. Only if. If, instead, rank $B_j^T < r_j$ for some j, then there is a nonzero r_j-dimensional vector $\bar{v}_j$, with coordinates $[\bar{x}^*(x_{j1}), \cdots, \bar{x}^*(x_{jr_j})]$ for some nonzero $\bar{x}^* \in X^*$, $\bar{x}^*(x_{jk}) \equiv 0, j \neq j, k = 1, \cdots, r_j$, such that $B_j^T \bar{v}_j = 0$. Hence, (3.4) implies $\bar{x}^*(A^n S(\bar{t}) b_i) = 0$, $n = 0, 1, \cdots$; $i = 1, \cdots m$; $\bar{t} > 0$, and we have a contradiction, via Proposition 1.1.

If (i) The bounded operator $AS(t)$, $t > 0$, is normal on $\mathscr{D}(A) = \mathscr{D}(A^*)$, hence on X, since $S(t)$ and $S^*(t)$, as well as A and $S^*(t)$, A^* and $S(t)$, and A and A^* commute on $\mathscr{D}(A)$. On $\mathscr{D}(A)$, we then have ([10, p. 168] for the first step)

$$[AS(t)][AS(t)]^* = AS(t)S^*(t)A^* = S^*(t)A^*AS(t) = [AS(t)]^*[AS(t)].$$

(ii) $AS(t)$ is compact on X for $t > 0$: this follows from the expansion (3.4) with $n = 1$ coupled with the fact that $|\lambda_j e^{\lambda_j t}| = |\lambda_j| e^{\mathrm{Re}\,\lambda_j t} \to 0$ as $j \to \infty$ for positive, fixed but otherwise arbitrary t [19, p. 207, also footnote 5].

(iii) Let $\bar{t} > 0$ and set $b_i' = S(\bar{t})b_i$. Then $(b_i', x_{jk}) = e^{\lambda_j \bar{t}}(b_i, x_{jk})$. If B_j' denotes the matrix B_j after Remark 3.1 for the vectors b_i', then rank $B_j' =$ rank B_j. Hence, by Proposition 3.1 applied to $AS(\bar{t})$, $M = \overline{\mathrm{sp}}\,\{[AS(\bar{t})]^n b_i'\}$, $n = 0, 1, \cdots, i = 1, \cdots m$, is all of X if and only if (3.5) holds.

(iv) By Remark 2.5, $N = \overline{\mathrm{sp}}\,\{A^nS(\bar{t})b_i\} = \overline{\mathrm{sp}}\,\{A^nS(t)b_i$, for all $t > 0\}$ and so $M \subset N$. Finally, if $N \subsetneq X$, then $M \subsetneq X$ and by the "if" part of (iii), (3.5) also fails. Q.E.D.

Remark 3.3. For A satisfying H1, H3 and H4 the following results are contained in the proof of Proposition 3.3 (here $n = 0, 1, \cdots$; $j = 1, 2, \cdots$; $i = 1, \cdots, m$):

$$\overline{\mathrm{sp}}\,\{[AS(t)]^n b_i'\} = X \Leftrightarrow \text{rank } B_j' = r_j \Leftrightarrow \text{rank } B_j = r_j \Leftrightarrow \overline{\mathrm{sp}}\,\{A^nS(t)b_i\} = X$$

with $b_i' = S(\bar{t})b_i$ and t and $\bar{t}$ arbitrary positive times.

Proposition 3.3 and Corollary 2.2 imply (see also Remark 2.3 (ii)) the following.

THEOREM 3.4. *Let A satisfy* H1, H3 *and* H4. *Then $\mathscr{L}_m : \langle A, (b_1, \cdots, b_m)\rangle$ is approximately controllable on* $[0, T]$ *if and only if* (3.5) *holds.*

Remark 3.4. It follows from Theorem 3.4 that the operator A *can* be made approximately controllable with a finite number m of scalar controls if and only if $r < \infty$, in which case $m \geqq r$.

COROLLARY 3.5. *Let A satisfy* H1, H3 *and* H4. *Then $\mathscr{L}_1 : \langle A, b\rangle$ is approximately controllable on* $[0, T]$ *if and only if* (3.5′) *holds.*

Theorem 3.4 was arrived at as an application of the general result in Corollary 2.2 to the present class of operators. We next complement Theorem 3.4. We remove the assumption H4 of analyticity for the semigroup and so our conclusion refers to approximate controllability in finite time (as opposed to approximate controllability on $[0, T]$), unless the vectors b_i are analytic vectors: $b_i \in \eta_a(A)$.

THEOREM 3.6. *Let A satisfy* H1 *and* H3. *Then $\mathscr{L}_m : \langle A, (b_1, \cdots, b_m)\rangle$ is approximately controllable in finite time* (on $[0, T]$, *if $b_i \in \eta_a(A)$*) *if and only if* (3.5) *holds.*

COROLLARY 3.7. *Let A satisfy* H1 *and* H3. *Then $\mathscr{L}_1 : \langle A, b\rangle$ is approximately controllable in finite time* (on $[0, T]$ *if $b \in \eta_a(A)$*) *if and only if* (3.5′) *holds.*

Proof of Theorem 3.6 *based on the explicit expression of the semigroup.* We apply the general characterization (1.3′) to the particular form (3.1) of the semigroup generated by A. Hence, we need only show that (3.5) holds if and only if the condition

$$(3.7) \qquad x^*(S(t)b_i) = \sum_{j=1}^{\infty} e^{\lambda_j t} \sum_{k=1}^{r_j} (b_i, x_{jk})x^*(x_{jk}) \equiv 0, \qquad t \geqq 0, \quad i = 1, \cdots, m,$$

for all $x^* \in X^*$, implies $x^* = 0$. *If*: Suppose by contradiction that rank $B_j^T < r_j$ for some j. Then there is a nonzero r_j-dimensional vector $\bar{v}_j$, with coordinates $[\bar{x}^*(x_{j1}) \cdots \bar{x}^*(x_{jr_j})]$ for some nonzero $\bar{x}^* \in X^*$, $\bar{x}^*(x_{jk}) \equiv 0$, $j \neq j$, $k = 1, \cdots r_j$,

such that $B_j^I \bar{v}_j = 0$. Hence, in view of (3.7), it follows that $\bar{x}^*(S(t)b_i) \equiv 0$, $t \geqq 0$, $i = 1, \cdots, m$, and $\mathscr{L}_m$ is not approximately controllable, since $\bar{x}^*$ is nonzero.

Only if: Suppose, by contradiction, that

$$\sum_{j=1}^{\infty} e^{\lambda_j t}\alpha_{ij} \equiv 0, \qquad t \geqq 0, \quad i = 1, \cdots, m, \tag{3.8}$$

for $\alpha_{ij} = \sum_{k=1}^{r_j} (b_i, x_{jk})\bar{x}^*(x_{jk})$ and $0 \neq \bar{x}^* \in X^*$. We shall show below that (3.8) implies $\alpha_{ij} \equiv 0$, $i = 1, \cdots, m$; $j = 1, 2, \cdots$, i.e., in compact form, $B_j^T \bar{v}_j = 0$, $j = 1, 2, \cdots$, with $\bar{v}_j = [\bar{x}^*(x_{j1}), \cdots, \bar{x}^*(x_{jr_j})]^T$. Since $\{x_{jk}\}$ form a complete set in X, and $\bar{x}^*$ is nonzero, then $\bar{v}_j$ is a nonzero vector for some $j = j$, via Proposition 1.1. Hence rank $B_j^T < r_j$ (which is automatically true if $m < r_j$) and this is a contradiction! It remains to show $\alpha_{ij} \equiv 0$.

For λ with $\operatorname{Re}\lambda > \omega_0$ we have from (3.8), [5, Thm. 11, p. 622] and (d) above,

$$0 \equiv \int_0^{\infty} e^{-\lambda t}\bar{x}^*(S(t)b_i)\, dt = \bar{x}^*(R(\lambda, A)b_i) = \sum_{j=1}^{\infty} \frac{\alpha_{ij}}{\lambda - \lambda_j},$$

and by analytic continuation of the right-hand side we have

$$\bar{x}^*(R(\lambda, A)b_i) = \frac{\alpha_{i1}}{\lambda - \lambda_1} + \sum_{j=2}^{\infty} \frac{\alpha_{ij}}{\lambda - \lambda_j} \equiv 0 \quad \text{for all } \lambda \neq \lambda_j$$

(the series is (pointwise) unconditionally convergent, i.e., absolutely convergent at each $\lambda \neq \lambda_j$, since the expression (d) above for $R(\cdot, A)y$ is independent on the order of the index j).

Next, since the λ_j's are isolated, pick a circle Γ_1 centered at λ_1 with sufficiently small radius as to enclose only λ_1 and leave the other λ_j's outside. Multiply the above identity by $(2\pi i)^{-1}$ and integrate over Γ_1. Using standard Cauchy's theorems we get: $\alpha_{i1} + 0 = 0$, since the summation from $j = 2$ is analytic within Γ_1. By induction it follows that $\alpha_{ij} \equiv 0$. Q.E.D.

Remark 3.5. When $\{\operatorname{Re}\lambda_j\}$ are isolated: $\cdots < \operatorname{Re}\lambda_2 < \operatorname{Re}\lambda_1$; $|\operatorname{Re}\lambda_j - \operatorname{Re}\lambda_{j+k}| \geqq \delta > 0$ for each j and k,—in particular when A is self-adjoint—a simpler argument can be given to show $\alpha_{ij} \equiv 0$. (i) Assume first that $\operatorname{Re}\lambda_1 < 0$. Then (3.8) gives $\alpha_{i1} + \sum_{j=2}^{\infty} e^{-(\lambda_1 - \lambda_j)t}\alpha_{ij} \equiv 0$, $t \geqq 0$. From here it follows $\alpha_{i1} = 0$, since the sum in j is bounded above by $e^{-\delta t}\sum_{j=2}^{\infty}|\alpha_{ij}|$, which goes to zero as $t \to \infty$ (the series of the α_{ij}'s is convergent since the expression (3.1) is independent of the order of the index j). (ii) The general case $\operatorname{Re}\lambda_1 < C$ is reduced to the previous one (i), by the translation $\lambda_j' = \lambda_j - C$ and application of (i) to $\{\lambda_j'\}$. Q.E.D.

We give next another proof of Theorem 3.6 which makes use of the general result on approximate controllability in finite time as explained in Appendix A. Such a result reduces the unbounded operator case to the bounded operator case.

Proof of Theorem 3.6 *based on the explicit expression of the resolvent* $R(\lambda, A)$. According to (A.1) in the Appendix we must show that: (3.5) holds if and only if $\overline{\operatorname{sp}}\{R^n(\eta_0, A)b_i\} = X$, where η_0 is a fixed point in the connected component of the resolvent set of A containing $\operatorname{Re}\lambda \geqq \omega_0$ (e.g., $\eta_0 > \operatorname{Re}\lambda_j$) and

$$R^n(\eta_0, A)b_i = \sum_{j=1}^{\infty} \frac{1}{(\eta_0 - \lambda_j)^n} \sum_{k=1}^{r_j} (b_i, x_{jk})x_{jk}, \qquad n = 0, 1, 2, \cdots.$$

Application of Proposition 3.1 to the normal compact operator $R(\eta_0, A)$ yields the conclusion. Q.E.D.

We finally treat the case when the multiplicities r_j of the eigenvalues λ_j of A tend to infinity as $j \to \infty$, (e.g., heat equation on the unit square $0 \leqq \xi_1 \leqq 1$, $0 \leqq \xi_2 \leqq 1$ in R^2 [8, p. 402]: here a system of type $\mathscr{L}_m$ is never approximately controllable, and an appropriate operator B with infinite-dimensional range is needed).

We confine ourselves to giving the statement of the analogue of the most general result for $\mathscr{L}_m$, namely, Theorem 3.6. An analysis of its proof(s) and a geometric interpretation of its statement shows that condition (3.5) for $\mathscr{L}_m$ is replaced, when the system $\mathscr{L}$ is considered, by the more general condition

$$P_j BU = X_j, \qquad j = 1, 2, \cdots, \quad BU = \text{range of } B, \tag{3.9}$$

where X_j is the r_j-dimensional eigenspace associated with the eigenvalue λ_j and P_j is the orthogonal projection $X \to X_j$. We have the following.

THEOREM 3.8. *Let A satisfy* H1 *and* H3. *Then $\mathscr{L}:\langle A, B\rangle$ is approximately controllable in finite time if and only if* (3.9) *holds.*

Notice that (3.9) can always be achieved, by choosing B with $\mathscr{R}(B) = X$, in particular, for $X = Y$, by choosing $B = I$.

Remark 3.6. For simplicity, the following considerations are given for the system $\mathscr{L}_1$, the corresponding extensions to the systems $\mathscr{L}_m$ being obtained by replacing (3.5′) by (3.5). Let $b \in X_0$, i.e., $b = S(\tilde{t})\tilde{b}$ for some $\tilde{b} \in X$ and $\tilde{t} > 0$. $(b, x_j) = e^{\lambda_j \tilde{t}}(\tilde{b}, x_j)$. Under the assumptions H1, H3 and H4, as in Proposition 3.3, one has $X_0 \subset D_\infty(A)$ and

$$\overline{\text{sp}}\,\{A^n b\} = X \Leftrightarrow (\tilde{b}, x_j) \not\equiv 0 \Leftrightarrow (b, x_j) \not\equiv 0 \Leftrightarrow \overline{\text{sp}}\,\{A^n S(\bar{t})b\} = X$$

for $b \in X_0$ and $\bar{t}$ arbitrary > 0. On the other hand, if H4 is omitted and $b \in \eta_a(A)$, it follows from Corollary 2.2 and Corollary 3.7, that $\overline{\text{sp}}\,\{A^n b\} = X \Leftrightarrow (b, x_j) \not\equiv 0$: the implication $\Rightarrow$ is valid in fact for *any* $b \in D_\infty(A)$, as one realizes using the same argument employed in the "only if" part of Proposition 3.1, when applied to the expansion (3.3).

4. Examples for § 3.2. The first two examples refer to a self-adjoint operator A generating, therefore, a (self-adjoint) analytic semigroup for $t > 0$ (heat equation type of systems). The last two examples refer to a normal operator A of the form $A = iF$, with F self-adjoint, generating a unitary group (wave equation), in agreement with Stone's theorem [5, p. 1243]. For the first two examples Theorem 3.4 or Corollary 3.5 are adequate, while for the last two one must resort to Theorem 3.6 or Corollary 3.7. Since the heat equation in a rectangle of R^n was already treated in [8, Example 4], we start with the heat equation on a disk.

Example 4.1 (heat equation on a disk). Let $X = L_2(S)$, where S is the unit disk in R^2, centered at the origin. Let $A = \Delta$, with $D(\Delta)$ consisting of all $f \in X$, with $\Delta f \in X$ and $f = 0$ on the boundary of S[7]. The eigenvalues and the nonnormalized eigenfunctions of Δ are expressed in polar coordinates (ρ, θ) by [4, p. 304]

$$\{\lambda_{n,m} = -k_{n,m}^2\} \quad \text{and} \quad \{J_n(k_{n,m}\rho)\cos n\theta,\ J_n(k_{n,m}\rho)\sin n\theta\},$$
$$n = 0, 1, \cdots, \quad m = 1, 2, \cdots,$$

[7] Δ is then self-adjoint with compact resolvent since S is bounded [5, p. 1330].

where $k_{n,m}$ are the real zeros of the Bessel function $J_n: J_n(k_{n,m}) = 0$. Except for $n = 0$, all the eigenvalues are at least double, since they have associated the linearly independent functions $J_n \cos n\theta$ and $J_n \sin n\theta$. Actually all the nonzero $k_{n,m}$ are distinct [20, pp. 484–5]. The highest multiplicity r in this case is clearly equal to two.

Hence, by Theorem 3.4, at least two functions $b_1(\rho, \theta)$, $b_2(\rho, \theta)$ in $L_2(S)$ are needed to make A approximately controllable, and this happens if and only if

$$\operatorname{rank} \begin{vmatrix} \iint b_1(\rho,\theta) J_n(k_{n,m}\rho) \cos n\theta \, d\rho \, d\theta, & \iint b_2(\rho,\theta) J_n(k_{n,m}\rho) \cos n\theta \, d\rho \, d\theta \\ \iint b_1(\rho,\theta) J_n(k_{n,m}\rho) \sin n\theta \, d\rho \, d\theta, & \iint b_2(\rho,\theta) J_n(k_{n,m}\rho) \sin n\theta \, d\rho \, d\theta \end{vmatrix} = 2,$$

$$n = 1, 2, \cdots, \quad m = 1, 2, \cdots,$$

and either

$$\iint b_1(\rho,\theta) J_0(k_{0,m}\rho) \, d\rho \, d\theta \neq 0,$$

or

$$\iint b_2(\rho,\theta) J_0(k_{0,m}\rho) \, d\rho \, d\theta \neq 0, \qquad m = 1, 2, \cdots.$$

Here $\iint$ indicates the integral extended over S.

Example 4.2 (Sturm–Liouville operator). Let $p(\xi)$, $p'(\xi)$, $q(\xi)$ and $\omega(\xi)$ be continuous real-valued functions on the finite interval $[a, b]$, and assume that $p(\xi) > 0$ and $\omega(\xi) > 0$ on $[a, b]$. Let X be the complex Hilbert space

$$\left\{ f(\cdot) : \int_a^b |f(\xi)|^2 \omega(\xi) \, d\xi < \infty \right\}$$

with inner product

$$(f, g) = \int_a^b f(\xi) \bar{g}(\xi) \omega(\xi) \, d\xi.$$

Consider the Sturm–Liouville operator $Af = [(pf')' - qf]/\omega$ with domain $D(A) = \{f \text{ is } C^2 \text{ in } X \text{ and satisfies (B.C.)}\}$.

(B.C.) $$\beta_1 f(a) + \gamma_1 f'(a) = 0, \qquad \beta_2 f(b) + \gamma_2 f'(b) = 0,$$

β_i and γ_i real constants and $|\beta_1| + |\gamma_1| > 0$, $|\beta_2| + |\gamma_2| > 0$. Then A is symmetric [13, p. 499] and has a self-adjoint extension still denoted by A since its coefficients are real [13, p. 536], [5, p. 1295]. Also, A has compact resolvent [13, p. 501], eigenvalues $\lambda_j, j = 1, 2, \cdots$, bounded above [13, p. 502] and an orthonormal basis of eigenvectors. See also [10, p. 274]. Each eigenvalue has at most multiplicity two [13, p. 502] and hence at most two appropriate vectors b_1 and b_2 in X (as in Theorem 3.4) make A approximately controllable on $[0, T]$. Actually, except for periodic boundary conditions, all the eigenvalues are simple [4, p. 293]: here then only one appropriate vector b in X (as in Corollary 3.5) suffices to make A

approximately controllable on $[0, T]$. For instance if $\omega \equiv p \equiv 1, q \equiv 0$; $\beta_1 = \beta_2 = 0; \gamma_1 = \gamma_2 = 1; a = 0, b = 1$, the eigenvalues and the normalized eigenfunctions are [5, p. 1383] $\{-(j\pi)^2\}$ and $\{\cos j\pi\xi\}, j = 0, 1, \cdots$, respectively. By Corollary 3.5, a vector b in X makes the correspondent operator A approximately controllable if and only if $\int_0^1 b(\xi) \cos j\pi\xi \, d\xi \neq 0, j = 0, 1, \cdots$.

Example 4.3. Let $X = L_2[0, 2\pi]$ and let $Ff = (1/i)f'$, with $D(F)$ given by all f in X that are absolutely continuous on $[0, 2\pi]$ and such that f' is also in X and $f(0) = f(2\pi)$. Finally let $A = iF$; see [5, p. 1381], [16, p. 269]. Then A is closed [16, p. 176], F is self-adjoint with compact resolvent [5, p. 1381], so that A is normal with compact resolvent and the spectrum of A consists entirely of simple eigenvalues $\lambda_j = ij$ on the imaginary axis, with (normalized) eigenvectors $x_j(\xi) = (1/\sqrt{2\pi})\, e^{i(j\xi)}$, $j = 0, \pm 1, \pm 2, \cdots$, forming a complete orthonormal set in X. A generates the unitary group $S(t)y = \Sigma_j e^{ijt}(y, x_j)x_j$. Corollary 3.7 is applicable, but not Corollary 3.5, and the necessary and sufficient condition for $\mathscr{L}_1 : \langle A, b\rangle$ to be approximately controllable in finite time (in $[0, T]$, if, moreover, b is an analytic vector) is that: $(\#) : (b, x_j) = \int_0^{2\pi} b(\xi)\, e^{ij\xi}\, d\xi \not\equiv 0$. Actually, using directly (1.3′), one gets the more refined result that $(\#)$ is in fact necessary and sufficient for $\langle A, b\rangle$, b not an analytic vector, to be approximately controllable on $[0, T]$, $T \geqq 2\pi$. In fact, let, by contradiction, $\Sigma_j e^{ijt}\alpha_j \equiv 0$, $0 \leqq t \leqq 2\pi$, $\alpha_j = (b, x_j)\bar{x}^*(x_j)$, $0 \neq \bar{x}^* \in X^*$; multiply both sides by e^{ijt} and integrate in t over $[0, 2\pi]$. Since the series is absolutely convergent as in Remark 3.6, we may interchange the integration with the infinite sum. We then get $\alpha_{\bar{j}} = 0$ and, by induction $\alpha_j \equiv 0$; hence $(b, x_j) = 0$ for some $\bar{j}$, since $\bar{x}^*$ is nonzero, via Proposition 1.1. This is a contradiction. Q.E.D.

If $X = L_2[0, l]$, the same argument shows that the minimum time for approximate controllability is, in fact, l.

Example 4.4 (wave equation).[8] Let $X = \mathring{H}_1(\Omega) \oplus L_2(\Omega)$, endowed with the energy inner product. The operator A and the energy inner product are

$$A = \begin{vmatrix} 0 & 1 \\ \Delta & 0 \end{vmatrix} \quad \text{and} \quad ([f_1, g_1], [f_2, g_2])_X = \int_\Omega \{\nabla f_1 \cdot \overline{\nabla f_2} + g_1 \cdot \overline{g_2}\}\, d\xi,$$

respectively, where ξ is the spatial coordinate and Ω is some bounded sufficiently smooth spatial domain in R^n. In this case, $D(A) = \{H_2(\Omega) \cap \mathring{H}_1(\Omega)\} \oplus \mathring{H}_1(\Omega)$ and $\dot{x} = Ax$ represents the wave equation with homogeneous Dirichlet boundary conditions. A is normal and, in fact, can be written as $A = iF$, with F self-adjoint:

$$(F[f_1, g_1], [f_2, g_2])_X = \int_\Omega \{\nabla(-ig_1) \cdot \overline{\nabla f_2} - i\Delta f_1 \cdot \overline{g_2}\}\, d\xi$$

$$= \int_\Omega \{(ig_1) \cdot \overline{\Delta f_2} + (i\nabla f_1) \cdot \overline{\nabla g_2}\}\, d\xi = ([f_1, g_1], F[f_2, g_2])_X$$

as it follows by Green's theorem and using the zero boundary conditions. Moreover A has compact resolvent, with eigenvalues $\lambda_j = \sqrt{\lambda_j^\Delta}$ and eigenvectors $x_j = [x_j^\Delta, -\lambda_j x_j^\Delta]$, where λ_j^Δ and x_j^Δ are the eigenvalues and the eigenvectors of Δ

[8] For wave equation types of systems with control acting on the *boundary*—as opposed to the distributed control considered here—Russell has shown, e.g., [14], that approximate controllability can be achieved only after a critical minimal time.

on $L_2(\Omega)$ [25, Remark 2.2]. $\sigma(A)$ consists of isolated eigenvalues on the imaginary axis, and A generates a unitary group as in Example 4.3. For instance, if $n = 1$ and $\Omega = (0, \pi)$, then $\lambda_j^\Delta = -j^2$, $x_j^\Delta(\xi) = \sin j\xi, j = 1, 2, \cdots$. Hence the eigenvalues $\lambda_j = \pm ij$ of A are all simple. According to Corollary 3.7, one (suitable) function $b(\xi) = [b_1(\xi), b_2(\xi)] \in X$ is sufficient to make A approximately controllable in finite time (in $[0, T]$, if furthermore, $b(\xi)$ is as in footnote 4), and this happens if and only if

$$(4.1) \qquad (b, x_j)_X = j \int_0^\pi \{b_1'(\xi) \cos j\xi \mp i b_2(\xi) \sin j\xi\}\, d\xi \not\equiv 0, \quad j = 1, 2, \cdots.$$

Actually, with $b(\cdot)$ arbitrary, the above condition (4.1) is necessary and sufficient for approximate controllability on $[0, T]$, $T \geqq \pi$ (in general, $T \geqq l$, if $\Omega = [0, l]$): see the argument at the end of the previous Example 4.3, which still applies.

The wave equation $\omega_{tt} = \Delta\omega + \gamma(\xi)u(t)$ can be written in the above framework with $b_1(\xi) \equiv 0$ and $b_2(\xi) = \gamma(\xi)$. In this case, the test (4.1) becomes

$$(\gamma, x_j)_{L_2} = \int_0^\pi \gamma(\xi) \sin j\xi\, d\xi \not\equiv 0, \qquad j = 1, 2, \cdots,$$

which is precisely the same as the test for approximate controllability on an arbitrary interval $[0, T]$ of the heat equation $\omega_t = \Delta\omega + \gamma(\xi)u(t)$ on $\Omega = [0, \pi]$. Finally, let us notice that $(\gamma, x_j)_{L_2} \not\equiv 0$ is precisely one of the two conditions in [29]—namely condition (2.18), the other is condition (2.17)—required, according to [29], to steer the initial state of the wave equation (initial position and initial velocity) *exactly* to zero over the time interval $[0, T]$, $T \geqq 2l$.

5. Observability. By observed process O(resp. O_p) we shall mean the complex of the state equation $\mathscr{L}$, or $\mathscr{L}_m$, subject to assumption H1, plus the output equation: $y = Hx$(resp. $y = [h_1, \cdots, h_p]x$, $h_q \in X^*$, $q = 1, \cdots, p$) where H is a bounded linear operator from the Banach space X (state space) into the normed linear space Y (output space). O_p corresponds to the physically significant case of $Y = R^p$, in which p sensors perform measurements on the system and reveal data from the global spatial distribution of the state. For $p = 1$, we write h instead of h_1. We call O, or O_p, observable[9] (resp. observable on $[0, T]$, $0 < T < \infty$) in case the initial state can be recovered from knowledge of the input $u(t)$ and output $y(t)$ for $t \geqq 0$ (resp. $0 \leqq t \leqq T$); i.e., by linearity of $\mathscr{L}$, in case the null output $y(t) = HS(t)x_0 \equiv 0$, $t \geqq 0$(resp. $0 \leqq t \leqq T$) implies $x_0 = 0$.

Remark 5.1. When A is bounded on X, observability of O(resp. O_p) is independent on the time interval length and, in particular, if the state space X is reflexive, is equivalent to

$$(5.1) \qquad \overline{\text{sp}}\, \{(A^*)^n H^* Y^*, n = 0, 1, \cdots\} = X^*,$$

$$(5.1') \qquad (\text{resp. } \overline{\text{sp}}\, \{(A^*)^n h_q, q = 1, \cdots, p; n = 0, 1, \cdots\} = X^*).$$

[9] See [6] for a definition of observability, to be called perhaps continuous observability, where the map from the observed trajectories $y(t), 0 \leqq t \leqq T$, as elements of some prescribed function space into $x(T)$ in X is required not only to be well-defined (as in our definition here) but also continuous. In [6] interesting results for continuous observability for the one-dimensional heat equation, are deduced using a very different approach from the one of the present paper. See also [15] and its references.

See [17]. For finite-dimensional systems, $X = R^n$, $Y = R^p$, (5.1′) reduces to the classical rank condition: rank $[H^*, A^*, \cdots (A^*)^{n-1}H^*] = n$.

It is documented in Appendix B that, like for approximate controllability, observability of O when the operator A acting on the state is unbounded is reduced to, and is equivalent to, the observability of the *associated* system with the (bounded!) resolvent $R(\cdot, A)$ acting on the state. The same holds for observability on $[0, T]$ when, e.g., the semigroup is analytic or A and H commute.

Remark 5.2. The investigation of conditions for observability directly in terms of the "coefficients" A and H of the null observed process will lead us to consider semigroups on the dual space X^*. Therefore, in order to simplify the exposition, we make henceforth in this section the assumption that the Banach state space X be reflexive. In this case: (i) if A is the infinitesimal generator of a strongly continuous semigroup (group) $S(t)$ on X, then the dual operator A^* is also (linear, closed, with domain $D(A^*)$ strongly dense in X^*) and the infinitesimal generator of the strongly continuous semigroup (group) $S^*(t)$ on X^* [3, Cor. 1.4.8, p. 52 with remark after Prop. 1.4.6, p. 50], [9, Thm. 14.4.1, p. 427 with Cor. p. 429 and remark after definition 14.2.1, p. 422], (ii) moreover, $S(t)$ is differentiable on X if and only if $S^*(t)$ is differentiable on X^*. This follows either from the equivalence between the strong limit defining the generator and its weak form [28, Thm. 1.1, p. 41] or from the characterization of a differentiable semigroup in terms of the location of the spectrum of its generator in a logarithmic sector and the growth condition of its resolvent operator [28, Thm. 4.3, p. 57]. These two conditions hold for A if and only if they hold for A^* [10, Thm. 6.22, p. 184]. So A satisfies H2 on X if and only if A^* satisfies H2 on X^*, (iii) finally, if $S(t)$ is analytic with holomorphic extension in the sector $\Delta = \{\lambda : \operatorname{Re}\lambda > 0, -\pi/2 \leqq \alpha < \arg\lambda < \beta \leqq \pi/2\}$, then $S^*(\cdot)$ is also analytic in Δ [3, p. 49], [9, § 3.10]. If the state space X is not reflexive, then $D(A^*)$ need not be strongly dense in X^* [3, Prop. 1.4.2, p. 46], [9, Thm. 2.11.9, p. 43]. However, in this case, a dual semigroup theory with the desired continuity properties can be carried out on the strong closure of $D(A^*)$; see [3, § 1.4] and [9, Chap. 14] for two different approaches. Our method for studying observability still works by replacing X^*, A^*, $T^*(\cdot)$ with, respectively, X_0^*, A_0^*, $T_0^*(\cdot)$ in [3] or $X^\odot$, $A^\odot$, $T^\odot(\cdot)$ in [9]. See also [17, Remark 5.1.1] for the condition of observability with X nonreflexive and A bounded.

As at the beginning of § 2, we define $D_\infty(A^*) = \bigcap_{n=1}^{\infty} D((A^*)^n)$, $Y_\infty^* = \{y^* \in Y^* : H^*y^* \in D_\infty(A^*)\}$, $\eta_a(A^*)$ to be the set of all vectors in X^* analytic for the semigroup (group) $S^*(t)$ generated by A^*, $Y_a^* = \{y^* \in Y^* : H^*y^* \in \eta_a(A^*)\}$, $Y_{a\infty}^* = Y_a^* \cap Y_\infty^*$, and we simply remark that analogous comments apply. The generalization of (5.1) is provided by the following theorem ("dual" of Theorem 2.1 for approximate controllability), where unless otherwise stated n runs as follows: $n = 0, 1, 2, \cdots$.

THEOREM 5.1. *Let X be a reflexive Banach space and let A satisfy* H1. *A sufficient condition for O to be observable on $[0, T]$ is given by*

$$\overline{\operatorname{sp}}\{(A^*)^n H^* Y_\infty^*\} = X^*, \tag{5.2}$$

or more generally, by

$$\overline{\operatorname{sp}}\{(A^*)^n S^*(\bar{t}) H^* Y_\infty^*\} = X^*, \quad \bar{t} \text{ arbitrary in } [0, T]. \tag{5.2′}$$

When A also satisfies H2, *then* (5.2′) *can be relaxed as to replace $H^*Y^*_\infty$ by H^*Y^* with $\bar{t}$ arbitrary in* $(0, T]$.

Conversely, assume $H^*Y^*_{a\infty}$ dense in H^*Y^*. Then a necessary condition for O to be observable on $[0, T]$ *is given by*

$$\overline{\text{sp}}\,\{(A^*)^n S^*(\bar{t})H^*Y^*_{a\infty}\} = X^*, \quad \bar{t} \text{ arbitrary} > 0. \tag{5.3}$$

If $S(t)$ is an analytic semigroup for $t > 0$, then (5.3) *can be relaxed as to replace $H^*Y^*_{a\infty}$ $(= H^*Y^*_\infty)$ by H^*Y^*. Also if $S(t)$ is a group, $\bar{t}$ in* (5.3) *can be any real number, in particular, $\bar{t} = 0$, and* (5.3) *simplifies, in this case, to*

$$\overline{\text{sp}}\,\{(A^*)^n H^*Y^*_a\} = X^*. \tag{5.3′}$$

In the next corollary, referring to O_p, unless otherwise stated, n and q run as follows: $n = 0, 1, 2, \cdots$, $q = 1, \cdots, p$.

COROLLARY 5.2. *Let X be reflexive and let A satisfy* H1. *A sufficient condition for O_p to be observable on* $[0, T]$ *is given by*

$$\overline{\text{sp}}\,\{(A^*)^n h_q\} = X^*, \qquad h_q \in D_\infty(A^*), \tag{5.4}$$

or, more generally, by

$$\overline{\text{sp}}\,\{(A^*)^n S^*(\bar{t})h_q\} = X^*, \qquad h_q \in D_\infty(A^*), \quad \bar{t} \text{ arbitrary in } [0, T]. \tag{5.4′}$$

When A also satisfies H2, *then the h_q's in* (5.4′) *can be relaxed to be any vectors in X^* with $\bar{t}$ in* $(0, T]$.

Conversely, a necessary condition for O_p to be observable on $[0, T]$ *is given by*

$$\overline{\text{sp}}\,\{(A^*)^n S^*(\bar{t})\,h_q\} = X^*, \quad \bar{t} \text{ arbitrary} > 0, \quad \text{when } \bar{t}h_q \in \eta_a(A^*) \cap D_\infty(A^*). \tag{5.5}$$

If $S(t)$ is an analytic semigroup for $t > 0$, then the h_q's in (5.5) *can be relaxed to be any vectors in X^*. Also, if $S(t)$ is a group, $\bar{t}$ in* (5.5) *can be any real number, in particular $\bar{t} = 0$, and* (5.5) *simplifies, in this case, to*

$$\overline{\text{sp}}\,\{(A^*)^n h_q\} = X^* \quad \text{when} \quad h_q \in \eta_a(A^*). \tag{5.5′}$$

Proof of Theorem 5.1. Using the reflexivity of X, one sees that $HS(t)x_0 \equiv 0$, $t \geqq 0$ (resp. $0 \leqq t \leqq T$) implies $x_0 = 0$ if and only if

$$H^{**}S^{**}(t)x^{**} \equiv x^{**}(S^*(t)H^*Y^*) \equiv 0,$$

$t \geqq 0$ (resp. $0 \leqq t \leqq T$) implies $x^{**} = 0$. This last implication is the counterpart ("dual") of (1.3) for approximate controllability. Since $S^*(t)$ is a semigroup (group) on X^*, the proof now proceeds exactly as in the proof of Theorem 2.1 except that it is carried out on X^* rather than X. Use the content of Remark 5.2. Q.E.D.

It is plain at this point, by comparing Theorem 2.1 with Theorem 5.1, that all the remarks in § 2 for approximate controllability can be rephrased in similar remarks for observability. We write explicitly only the counterpart of Remark 2.3, because it illustrates the generalization of the classical rank condition for observability.

Remark 5.3. For simplicity of notation we limit our comments only to O_p, the extension to O being immediate.

(i) Let $S(t)$ be a group: a characterization for observability on $[0, T]$ of O_p: $\langle A, (h_1, \cdots, h_p)\rangle$ is then given by $\overline{\text{sp}}\,\{(A^*)^n h_q\} = X^*$, at least for vectors h_q in the set $\eta_a(A^*)$ dense in X^*.

(ii) Let $S(t)$ be an analytic semigroup, $t > 0$: the same characterization is then given by $\overline{\text{sp}}\,\{(A^*)^n s^*(\bar{t}) h_q\} = X^*$, any $h_q \in X^*$, $\bar{t}$ arbitrary positive time. Also let $X_0^* = \cup_{0<t} S^*(t)X^*$ ($\subset D_\infty(A^*)$). X_0^* is dense in X^* and for $h_q = S^*(\tilde{t}_q)\tilde{h}_q$, $\tilde{h}_q \in X^*$ and $\tilde{t}_q > 0$, we have that $\overline{\text{sp}}\,\{(A^*)^n h_q\} = X^* (\Leftrightarrow \overline{\text{sp}}\,\{A^*)^n S^*(t)\tilde{h}_q\} = X^*, t$ arbitrary > 0) is necessary and sufficient for observability of O_p on $[0, T]$.

Remark 5.4. More specific duality relations can be given. For instance, let $h_q \in \eta_a(A^*)$, $0 = 1, \cdots, p$, if A generates a group or let $h_q \in X_0^*$ if A generates an analytic semigroup for $t > 0$. Then: $\dot{x}^* = A^*x^* + \sum_{q=1}^{p} h_q u$ is approximately controllable on $[0, T]$ if and only if the null system $\dot{x} = Ax$, $y = [h_1, \cdots, h_p]x$ is observable on $[0, T]$.

It is also evident that all the results given in § 3.2 for approximate controllability have a counterpart for observability. In fact, let X be a Hilbert space ($= X^*$): A is normal with compact resolvent if and only if the same holds for A^*; ($R(\lambda, A)$ is compact if and only if $R^*(\lambda, A) = R(\lambda, A^*)$ is compact [10, p. 159; p. 184]). Also, in this case, $D(A) = D(A^*)$ [10, p. 276] and hence $D_\infty(A) = D_\infty(A^*)$; A and A^* have the same eigenvectors $\{x_{jk}\}$ and the eigenvalues of one are the complex conjugates of the eigenvalues of the other with same multiplicity r_j, so A satisfies H4 if and only if A^* does also. We therefore confine ourselves to report explicitly below only the counterpart of the main results of § 3, namely Theorems 3.6 and 3.8 and Corollary 3.7. For the rest of this section, X is a Hilbert space. First, for vectors $h_1, \cdots, h_p$ in $X^* = X$, define the following sequence of $r_j \times p$ matrices:

$$\mathscr{H}_j = \begin{vmatrix} (h_1, x_{j_1}), \cdots, (h_p, x_{j_1}) \\ (h_1, x_{j_2}), \cdots, (h_p, x_{j_2}) \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ (h_1, x_{j_{r_j}}), \cdots, (h_p, x_{j_{r_j}}) \end{vmatrix}, \qquad j = 1, 2, \cdots,$$

and consider the following condition

$$\text{rank } \mathscr{H}_j = r_j, \qquad j = 1, 2, \cdots, \tag{5.6}$$

which implies $p \geqq r$; in particular for $p = 1$, we have

$$(h, x_j) \not\equiv 0, \qquad j = 1, 2, \cdots. \tag{5.6'}$$

In the statement of the next two results it must be noticed that $\eta_a(A^*) = X^*$, if A (hence A^*) satisfies H4. The dual of Theorem 3.6 is the following.

THEOREM 5.3. *Let A (i.e., A^*) satisfy* H1 *and* H3. *Then O_p: $\langle A, (h_1, \cdots, h_p)\rangle$ is observable (observable on $[0, T]$ if $h_q \in \eta_a(A^*)$, $q = 1, \cdots, p$) if and only if* (5.6) *holds.*

Hence the operator A can be made observable with a finite number p of sensors if and only if $r < \infty$, in which case $p \geqq r$. The case of $p = 1$ is dealt with in the following corollary, dual of Corollary 3.7.

COROLLARY 5.4. *Let A satisfy* H1 *and* H3. *Then $O_1:\langle A, h\rangle$ is observable (observable on $[0, T]$ if $h \in \eta_a(A^*)$) if and only if* (5.6′) *holds.*

Remark 5.5. Our Theorem 5.2 is the abstract version containing, in particular, the special case treated by Sakawa [15, Thm. 1] of a *particular* self-adjoint operator A with compact resolvent under even further assumptions; for instance, we impose no requirement on $\{1/(\lambda_i - \lambda_1)\}$, λ_i eigenvalues of A, being in l_2. See assumption 1 in [15]. It is also plain that the examples discussed in § 4 for approximate controllability can be used to generate observable systems, as well as approximately controllable *and* observable systems.

When $r_j \to \infty$ as $j \to \infty$ we get the following result dual of Theorem 3.8. Consider the condition

$$P_j H^* Y^* = X_j, \qquad H^* Y^* = \text{range of } H^* \text{ on } X^* = X, \tag{5.7}$$

where X_j is the r_j-dimensional eigenspace associated with the eigenvalue λ_j of A and P_j is the orthogonal projection $X \to X_j$ Equations (5.7) generalize (5.6). Then we have the following.

THEOREM 5.5. *Let A satisfy* H1 *and* H3. *Then $O:\langle A, H\rangle$ is observable if and only if* (5.7) *holds.*

Notice that (5.7) can always be achieved by choosing $\mathcal{R}(H^*) = X^* = X$ ($\Leftrightarrow H^{-1}$ exists and is continuous [16, p. 233]), in particular, for $X = U$, by choosing $H = I$.

Remark 5.6. It is an obvious matter to translate the results of the present paper for approximate controllability on the state space X to analagous results on the output space Y.

We do not insist. For the bounded operator case, see [17].

Appendix A. Associate with the original system $\mathcal{L}$ the system $\mathcal{L}_{\eta_0}: \dot{x} = R(\eta_0, A)x + Bu$ defined on the same spaces X and U. Here $R(\eta_0, A)$ is the (bounded) resolvent of A, η_0 any point in the connected component $\rho_0(A)$ of the resolvent set $\rho(A)$ of A containing $\operatorname{Re}\lambda \geqq \omega_0$. We have

$$\operatorname*{Cl}_{0 \leqq t < \infty} K_t(\mathcal{L}) = \operatorname*{Cl}_{0 \leqq t < \infty} K_t(\mathcal{L}_{\eta_0}) = \operatorname{Cl} K_T(\mathcal{L}_{\eta_0}) = \overline{\operatorname{sp}}\,\{R^n(\eta_0, A)BU, n = 0, 1, \cdots\},$$

where T is arbitrary but fixed, $0 < T < \infty$. The first equality was shown in [8, Prop. 2.3] and motivated the assumption of bounded operator in [17], where the other equalities are proved. Therefore $\mathcal{L}$(resp. $\mathcal{L}_m$) is approximately controllable in finite time if and only if

$$\overline{\operatorname{sp}}\,\{R^n(\eta_0, A)BU\} = X \quad (\text{resp. } \overline{\operatorname{sp}}\,\{R^n(\eta_0, A)b_i\} = X, i = 1, \cdots, m), \tag{A.1}$$
$$n = 0, 1, \cdots.$$

Appendix B. To the null observed system O of § 5: $\dot{x} = Ax$, $y = Hx$, associate the system $O_{\eta_0}: \dot{x} = R(\eta_0, A)x$, $y = Hx$, with $R(\eta_0, A)$ as in Appendix A. Essentially the argument employed by Fattorini for approximate controllability can be adopted to show that *O is observable if and only if O_{η_0} is observable on* $[0, T]$, thereby reducing the observability problem in the unbounded operator case to the observability in the bounded operator case (see (5.1)). A concise proof of this fact follows. We have to show that $HS(t)x_0 \equiv 0$, $t \geqq 0$, implies $x_0 = 0$ if and only if $He^{R(\eta_0, A)t}x_0 \equiv 0$, $t \geqq 0$, implies $x_0 = 0$.

If. If, by contradiction, $HS(t)\bar{x}_0 \equiv 0$, $t \geqq 0$, for $\bar{x}_0 \neq 0$, then it follows, via

$$R^n(\lambda, A)x = \frac{1}{(n-1)!}\int_0^\infty t^{n-1} e^{-\lambda t} S(t)x\, dt, \tag{B.1}$$

$$n = 1, 2, \cdots, \quad \operatorname{Re}\lambda > w_0 \text{ [5, p. 623]},$$

and $HS(0)\bar{x}_0 = H\bar{x}_0$, that $HR^n(\lambda, A)\bar{x}_0 = 0$, $n = 0, 1, \cdots$, for all λ in $\rho_0(A)$ (by analytic continuation of $R(\lambda, A)$): contradiction.

Only if. If by contradiction, $He^{R(\eta_0, A)t}\bar{x}_0 \equiv 0, t \geqq 0$, and $\bar{x}_0 \neq 0$, then $HR^n(\eta_0, A)\bar{x}_0 = 0, n = 0, 1, \cdots$. Hence using the standard expansion [16, p. 260] for $R(\lambda, A)$ in terms of $R^{n+1}(\mu, A)$ and analytic continuation of $R(\cdot, A)$, it follows that $HR(\lambda, A)\bar{x}_0 \equiv 0$ for all λ in $\rho_0(A)$. Then by (B.1) with $n = 1$ and the uniqueness of the Laplace transform [5, p. 626], via an arbitrary $x^* \in X^*$, one gets $HS(t)\bar{x}_0 \equiv 0, t \geqq 0$: contradiction. Q.E.D.

The above reduction result holds also true for "observability on $[0, T]$" of $\mathscr{L}$ if: (i) the semigroup $S(t)$ is analytic, $t > 0$; or if (ii) H and A commute (in the sense of [10, p. 171]; here we take $Y = X$), so that H and $S(t)$ commute (see below). In both cases, in fact, $HS(t)x_0 \equiv 0$, $0 \leqq t \leqq T$, implies $HS(t)x_0 \equiv 0$, $t \geqq 0$. Proof of (ii): H and $R(\lambda, A)$ commute for every $\lambda \in \rho(A)$ [10, p. 173]; hence, making use of (B.1) for $n = 1$, one gets $\int_0^\infty e^{-\lambda t} x^*(HS(t)x - S(t)Hx)\, dt = 0$, $\operatorname{Re}\lambda > \omega_0$, and hence $HS(t)x = S(t)Hx$, $t \geqq 0$. Q.E.D.

Appendix C. Corollary 3.2 allows one to make the following considerations regarding the stabilizability of controllable systems under perturbation. In the classical finite dimensional theory $X = R^n$, $U = R^m$, it is well known that controllable systems are open and dense in the totality of autonomous linear systems [12, p. 100]. For infinite-dimensional systems, instead, we have already given an example in [17 remark 3.2.1] involving a Volterra (compact) integral operator V on a Hilbert space, where the openness property fails, even if V is left unperturbed and $\|V\| < 1$ (hence $\|b - b'\| < \varepsilon$ implies $\|V^n b - V^n b'\| < \varepsilon, n = 0, 1, 2, \cdots$); yet $\langle V, b\rangle$ is an approximately controllable pair, while $\langle V, b'\rangle$ is not.

Another more illuminating counterexample to the openness property for infinite-dimensional systems stems from Corollary 3.2 and is presented here. Let A be a normal, compact operator on a Hilbert space X, with simple nonzero eigenvalues. If $\{x_j\}$ denote the associated eigenvectors (forming a complete orthonormal basis), let b be a vector in X, $b = \sum_{j=1}^\infty (b, x_j)x_j$, satisfying $(b, x_j) \neq 0$, $j = 0, 1, \cdots$. Hence, the pair $\langle A, b\rangle$ is approximately controllable by Corollary 3.2. Now, given $\varepsilon > 0$, there is an index $J(\varepsilon)$ such that $\sum_{J(\varepsilon)}^\infty |(b, x_j)|^2 < \varepsilon$. Define then a vector b' in X by

$$(b', x_j) = (b, x_j), \qquad j = 1, 2, \cdots, J(\varepsilon) - 1$$

and

$$(b', x_j) = 0 \quad \text{for } j \geqq J(\varepsilon).$$

Then $\|b - b'\| < \varepsilon$, and yet $\langle A, b'\rangle$ is not an approximately controllable pair by Corollary 3.2. Also, one can check that, with A left unperturbed, the denseness property does hold (as for the Volterra operator V in [17, remark 3.2.1]).

Note added in proof. After the present paper was submitted, the following relevant work appeared: Y. Sakawa, Controllability for partial differential equations of parabolic type, this Journal, 12 (1974), pp. 389–400, which is the "dual" of [15] (see Remark 5.5). In Sakawa's paper a *particular* self-adjoint operator with compact resolvent is treated with finitely many controls. His Theorem 3 in § 3 is a special case of the abstract version given by our Theorem 3.4. Sakawa does not approach the problem from the viewpoint of extending the rank conditions, as we do in the present paper.

The present paper was presented in a lecture at the Stefan Banach International Mathematical Centre, Warsaw, Poland, in the Semester on Mathematical Questions of Optimal Control, December, 1973.

Also, we remark that the proof of the general result in § 2 exploits a quasi-analyticity type of property rather than full analyticity.

REFERENCES

[1] A. V. BALAKRISHNAN, *Optimal control problems in Banach spaces*, this Journal, 3 (1965), pp. 152–180.

[2] ———, *Introduction to Optimization Theory in a Hilbert Space*, Lecture Notes, Springer-Verlag, Berlin, 1971.

[3] P. L. BUTZER AND H. BERENS, *Semigroups of Operators and Approximations*, Springer-Verlag, Berlin, 1967.

[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. 1, Interscience, New York, 1963.

[5] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Parts 1 and 2, Interscience, New York, 1959 and 1963.

[6] S. DOLECKI, *Observation for the one-dimensional heat equation*, Studia Math., 48 (1973), pp. 291–305.

[7] H. O. FATTORINI, *Some remarks on complete controllability*, this Journal, 4 (1966), pp. 686–694.

[8] ———, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.

[9] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, American Mathematical Society, Providence, R.I., 1958.

[10] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, Berlin, 1966.

[11] G. E. LADAS AND V. LAKAMIKANTHAM, *Differential Equations in Abstract Spaces*, Academic Press, New York, 1972.

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

[13] T. H. NAYLOR AND G. R. SELL, *Linear Operators in Engineering and Science*, Holt, Rinehart and Winston, New York, 1971.

[14] D. L. RUSSELL, *Boundary value control of the higher-dimensional wave equation*, Parts 1 and 2, this Journal, 9 (1971), pp. 29–42 and 9 (1971), pp. 401–419.

[15] Y. SAKAWA, *Observability and related problems for partial differential equations of parabolic type*, this Journal, 13 (1975), pp. 14–27.

[16] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.

[17] R. TRIGGIANI, *Controllability and observability in Banach space with bounded operators*, this Journal, 13 (1975), pp. 462–491.

[18] ———, *On the lack of exact controllability for mild solutions in Banach space*, J. Math. Anal. Appl., 50 (1975), pp. 438–446.

[19] G. HELMBERG, *Introduction to Spectral Theory in Hilbert Space*, American Elsevier, New York, 1969.

[20] G. N. WATTSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, 1944.

[21] K. YOSIDA, *Functional Analysis*, Springer, Berlin-Göttingen, 1965.

[22] L. A. LIUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Frederick Ungar, New York, 1961.
[23] E. NELSON, *Analytic vectors*, Ann. of Math., 70 (1959), pp. 572–615.
[24] J. WERMER, *On invariant subspaces of normal operators*, Proc. Amer. Math. Soc., 3 (1959), pp. 270–277.
[25] M. SLEMROD, *A note on complete controllability and stabilizability of linear control systems in Hilbert space*, this Journal, 12 (1974), pp. 500–508.
[26] L. HORMANDER, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1964.
[27] G. SZEGÖ, *Orthogonal Polynomials*, Colloquium Publications, American Mathematical Society, Providence, R.I., 1959.
[28] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Dept. of Mathematics Lecture Note 10, University of Maryland, College Park, Maryland, 1974.
[29] D. L. RUSSELL, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–559.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

ON SERIES AND PARALLEL COUPLING OF A CLASS OF DISCRETE TIME INFINITE-DIMENSIONAL SYSTEMS*

PAUL A. FUHRMANN†

Abstract. Necessary and sufficient conditions for controllability and observability of series and parallel coupling of a class of infinite-dimensional realizations are obtained in terms of factorizations of the related transfer functions.

1. Introduction. In this paper we study the coupling in series and parallel of a class of infinite-dimensional realizations. We shall be interested in the controllability and observability of the coupled system and the conditions under which these hold. The analytical tools for this study are mainly the study of ranges of Hankel operators and certain related factorizations of operator-valued functions. This paper extends the results obtained in [12] and generalizes the finite-dimensional results obtained in [2], [11]. As a by-product of the analysis carried out in this paper we get theorems about the similarity of certain operators to restricted shift operators which are of intrinsic interest.

We shall now give a short survey of the main ideas of system theory used in this paper. We shall restrict ourselves to the case of discrete time systems. Let A be an operator-valued analytic function in a neighborhood of the origin. Thus A has a Taylor expansion $A(z) = \sum_{n=0}^{\infty} A_n z^n$. We assume $A(z): U \to Y$ where U and Y are a pair of, usually finite-dimensional, Hilbert spaces. We call U the input space and Y the output space. We consider the function A, a transfer function, as carrying the information about the input/output properties of the system. By this we mean that given a sequence of inputs $\{u_n | n \geqq 0, u_n \in U\}$ we obtain a sequence of outputs $\{y_n | n \geqq 0, y_n \in Y\}$ where

$$y_n = \sum_{i=0}^{n-1} A_{n-i-1} u_i. \tag{1.1}$$

Relation (1.1) gives an external description of the system behavior without giving insight into the internal mechanism that produces the input/output relations.

By an internal description of a system we mean a triple of bounded operators $\{F, G, H\}$ where $F \in B(K, K)$, K being a Hilbert space to which we refer as the state space, $G \in B(U, K)$ and $H \in B(K, Y)$. We do not restrict K to be finite-dimensional but we do make that assumption about the input space U and the output space Y. Thus in the general case we deal with finite input/finite output infinite-dimensional systems. The triple $\{F, G, H\}$ is taken to represent a set of dynamical equations of the form

$$\begin{aligned} x_{n+1} &= Fx_n + Gu_n, \\ y_n &= Hx_n, \end{aligned} \tag{1.2}$$

* Received by the editors July 23, 1974, and in revised form March 25, 1975.

† Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts. Now at Department of Mathematics, Ben Gurion University of the Negev, Beersheva, Israel. This work was supported by the U.S. Office of Naval Research under the Joint Services Electronics Program by Contract N00014-67-A-0298-0006.

with initial condition $x_0 \in K$. Clearly the solution of the system of equations (1.2) is given by

$$x_n = F^n x_0 + \sum_{i=0}^{n-1} F^{n-1-i} G u_i, \tag{1.3}$$

and hence,

$$y_n = HF^n x_0 + \sum_{i=0}^{n-1} HF^{n-1-i} G u_i. \tag{1.4}$$

Now if we assume that the initial condition x_0 satisfies $x_0 = 0$, then the input/output description (1.4) reduces to (1.1) with

$$A_n = HF^n G, \qquad n \geqq 0. \tag{1.5}$$

Thus we say that $\{F, G, H\}$ is a realization of the transfer function A if (1.5) is satisfied. This is equivalent to the fact that for z in some neighborhood of the origin we have

$$A(z) = \sum_{n=0}^{\infty} A_n z^n = \sum_{n=0}^{\infty} HF^n G z^n = \mathrm{H} \sum_{n=0}^{\infty} (zF)^n G$$

or

$$A(z) = H(I - zF)^{-1} G. \tag{1.6}$$

The realization is finite-dimensional if the dimension of K is finite. This clearly implies the rationality of A. The transfer functions dealt with in this paper are nonrational in the generic case.

It is easy to see that the characteristic function introduced and studied in detail by Sz.-Nagy and Foias [20], [22] is nothing but the transfer function of a special kind of system, called standard unitary system by Helton [14]. However, in our formulation the operators G and H are much more loosely related to the operator F than in the case of the characteristic operator-valued function.

Realization theory is concerned with producing internal description of the form (1.2) starting from input/output descriptions of the form (1.1) as well as the study of the relation between two different realizations of the same input/output relation which satisfy some additional minimality conditions. To this end we introduce the notions of controllability and observability. We refer to [15] for an exhaustive treatment of these notions in the finite-dimensional context and which contain some interesting notes on the development of the subject.

We say that a realization $\{F, G, H\}$ is controllable if the set of vectors of the form $\{\sum_{i=0}^{n} F^i G u_i | u_i \in U, n \geqq 0\}$ is dense in the state space. Thus the system $\{F, G, H\}$ is controllable if given any vector x in the state space K we can, starting with the initial condition $x_0 = 0$, find a sequence of controls $\{u_i\}$ such that for some index x_n is arbitrarily close to x. Clearly, using a standard density argument, the controllability of the system is equivalent to the following condition:

$$\bigcap_{n=0}^{\infty} \ker G^* F^{*n} = \{0\}. \tag{1.7}$$

We recall that a vector g is a cyclic vector for an operator F if the set of all linear combinations of the vectors $F^n g$, $n \geqq 0$, is dense in the space. This is equivalent to $\vee_{n=0}^{\infty} F^n M_g = K$ where M_g is the one-dimensional subspace spanned by g. Analogously a subspace M of K will be called a cyclic subspace for F if $\vee_{n=0}^{\infty} F^n M = K$. It follows that the system $\{F, G, H\}$ is controllable if and only if range G is a cyclic subspace for F. Hence controllability is but a simple generalization of the notion of cyclicity.

We define the adjoint system of $\{F, G, H\}$ to be the system $\{F^*, H^*, G^*\}$. We say that $\{F, G, H\}$ is observable if the adjoint system is controllable. Thus the observability of $\{F, G, H\}$ is equivalent to

$$\bigcap_{n=0}^{\infty} \ker HF^n = \{0\}. \tag{1.8}$$

A realization $\{F, G, H\}$ is canonical if it is both controllable and observable. In the sequel a stronger notion of controllability and observability will play an important role. Following Helton [14] we define the controllability operator $\mathscr{C}$ on the dense set of finitely nonzero sequences in $l^2(0, \infty; U)$ by

$$\mathscr{C}(\{u_n\}) = \sum_{n=0}^{\infty} F^n G u_n. \tag{1.9}$$

There is no convergence problem as there are only a finite number of nonzero elements in the sequence. If $\mathscr{C}$ has an extension to a bounded operator from $l^2(0, \infty; U)$ onto K we say that the system $\{F, G, H\}$ is exactly controllable. Essentially, exact controllability means that every vector in the state space can be reached from the origin by a sequence of controls having finite energy where the energy is defined by $\sum_{i=0}^{\infty} \|u_i\|^2 < \infty$.

In an analogous way we define the observability operator and the notion of exact observability. Analytic criterias for controllability and observability and their exact counterparts have been developed by the author for the case of shift systems and we refer to [7], [8] for a full exposition.

Now suppose we are given two realizations $\{F_1, G_1, H_1\}$ and $\{F_2, G_2, H_2\}$ with input spaces U_i, output spaces Y_i and state spaces K_i, respectively. If $U_2 = Y_1$ we can define the series connection of the two systems by letting the state space K be the direct sum $K_1 \oplus K_2$ and the dynamical equation be given by

$$\begin{pmatrix} x_{n+1}^{(1)} \\ x_{n+1}^{(2)} \end{pmatrix} = \begin{pmatrix} F_1 & 0 \\ G_2 H_1 & F_2 \end{pmatrix} \begin{pmatrix} x_n^{(1)} \\ x_n^{(2)} \end{pmatrix} + \begin{pmatrix} G_1 \\ 0 \end{pmatrix} u_n, \tag{1.10}$$

$$y_n = (0 \quad H_2) \begin{pmatrix} x_n^{(1)} \\ x_n^{(2)} \end{pmatrix}.$$

Similarly if $U_1 = U_2$ and $Y_1 = Y_2$, we let $K = K_1 \oplus K_2$ and define the parallel connection of the two realizations to be given by the set of equations

$$\begin{pmatrix} x_{n+1}^{(1)} \\ x_{n+1}^{(2)} \end{pmatrix} = \begin{pmatrix} F_1 & 0 \\ 0 & F_2 \end{pmatrix} \begin{pmatrix} x_n^{(1)} \\ x_n^{(2)} \end{pmatrix} + \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} u_n, \tag{1.11}$$

$$y_n = (H_1 \quad H_2) \begin{pmatrix} x_n^{(1)} \\ x_n^{(2)} \end{pmatrix}.$$

If A_1 and A_2 are the transfer functions of $\{F_1, G_1, H_1\}$ and $\{F_2, G_2, H_2\}$ respectively, then the series connection given by (1.10) has $\chi A_2 A_1$ as transfer function whereas the parallel connection given by (1.11) has $A_1 + A_2$ as transfer function. It is clear that the realizations obtained by coupling canonical systems in series or parallel need not be controllable nor observable. It is of interest to find conditions under which the coupled systems are controllable or observable, and this is the central issue of this paper. The finite-dimensional problem has been solved recently [2], [10] by completely different methods. In the infinite-dimensional case the problem cannot be solved in its full generality and we will have to restrict ourselves to a special class of realizations, the shift realizations [6], [8], [14]. They will be introduced in the next section. For that case we have a natural substitute for the dimension function which will be utilized extensively in the sequel.

2. Hankel operators, left invariant subspaces and the shift realization. Let M, N be separable complex Hilbert spaces. We denote by $L^2(N)$ the space of all (equivalence classes) weakly measurable functions from the unit circle to N having finite norm. The norm in $L^2(N)$ is induced by the inner product

$$(f, g) = \frac{1}{2\pi} \int (f(e^{it}), g(e^{it}))_N \, dt.$$

Functions in $L^2(N)$ have Fourier expansions of the form $f(e^{it}) = \sum_{n=-\infty}^{\infty} f_n e^{int}$ with $\sum_{n=-\infty}^{\infty} \|f_n\|_N^2 < \infty$. The f_n are the vectorial Fourier coefficients of f and are given through $f_n = (1/(2\pi)) \int_0^{2\pi} f(e^{it}) e^{-int} \, dt$. We let $H^2(N)$ denote the subspace of $L^2(N)$ of all functions whose negative indexed Fourier coefficients are zero, i.e., $f_n = 0$ for $n < 0$. We recall [13] that $H^2(N)$ functions have analytic extensions to the open unit disc from which they can be recaptured by radial limits almost everywhere. Thus if $f \in H^2(N)$ and $f(e^{it}) = \sum_{n=0}^{\infty} f_n e^{int}$ the analytic extension is given simply by $f(z) = \sum_{n=0}^{\infty} f_n z^n$. We systematically use the same letter to denote both the function of $H^2(N)$ as defined on the unit circle as well as its analytic extension to the open unit disc. By χ we denote the identity function on the closed unit disc, i.e., $\chi(z) = z$. We define the right shift S in $H^2(N)$ by $Sf = \chi f$. We note that its adjoint S^*, to which we refer as the left shift, is given by $(S^* f)(z) = (f(z) - f(0))/z$. The shift terminology comes from the way the actions of S and S^* are reflected in the sequence of Fourier, or Taylor, coefficients of the function f.

We let $B(M, N)$ denote the space of all bounded linear operators from M to N, and let $L^\infty(B(M, N))$ be the space of weakly measurable, essentially bounded functions from the unit circle to $B(M, N)$ equipped with the sup norm. Elements of $L^\infty(B(M, N))$ have Fourier expansions [13] and we will denote by $H^\infty(B(M, N))$ the functions in $L^\infty(B(M, N))$ whose negative indexed Fourier coefficients are all zero. Functions in $H^\infty(B(N, M))$ have analytic extensions to the open unit disc and can be recaptured as strong radial limits almost everywhere. For A in $L^\infty(B(N, M))$ we let $\tilde{A}(z) = A(\bar{z})^*$.

A subspace of $H^2(N)$ will be called *right invariant* or *left invariant* according to whether it is invariant under the right or left shift respectively. The structure of right and left invariant subspaces of $H^2(N)$ is known. To describe it we need the

notion of a partial isometry. An operator W in a Hilbert space H is a partial isometry if for some subspace M of H we have

$$\|Wx\| = \|x\| \quad \text{if } x \in M$$

and

$$Wx = 0 \quad \text{if } x \in M^{\perp}.$$

M is called the initial space of the partial isometry and WM its final space. An operator W is an isometry if it is a partial isometry with the initial space coinciding with H. W is unitary if it is an isometry and its final space coincides with H.

A subspace K of $H^2(N)$ is right invariant if and only if it has a representation $K = PH^2(N)$ for some $P \in H^{\infty}(B(N, N))$ for which almost everywhere $P(e^{it})$ is a partial isometry with a fixed initial space. If P is almost everywhere unitary then P is called *inner*. The invariant subspaces for which P is inner are called invariant subspaces of full range [13]. For an inner function P in $H^{\infty}(B(N, N))$ we denote by $H(P)$ the left invariant subspace $\{PH^2(N)\}^{\perp}$. Thus we have the following direct sum decomposition:

$$H^2(N) = H(P) \oplus PH^2(N). \tag{2.1}$$

We let $P_{H(P)}$ be the orthogonal projection of $H^2(N)$, and sometimes of $L^2(N)$, onto $H(P)$. We define an operator $S(P)$ in $H(P)$ by

$$S(P)f = P_{H(P)}\chi f \tag{2.2}$$

for all f in $H(P)$. $S(P)$ is called the restricted right shift and we have $S(P)^* = S^*|H(P)$. An inner function P in $H^{\infty}(B(N, M))$ is a left inner factor of a function A in $H^{\infty}(B(N, M))$ if $A = PA'$ for some A' in $H^{\infty}(B(N, M))$. Two functions A in $H^{\infty}(B(M, N))$ and A_1 in $H^{\infty}(B(M_1, N))$ are relatively *left prime* if A and A_1 have no common trivial left inner factor. We shall use $(A, A_1)_L = I_N$ to denote the left primeness of A and A_1. We shall say that A and A_1 are *strongly relatively left prime* if there exists a $\delta > 0$ such that for all $z, |z| < 1$,

$$\inf\{\|A(z)^*\xi\| + \|A_1(z)^*\xi\| \,|\xi \in M, \|\xi\| = 1\} \geqq \delta. \tag{2.3}$$

We shall use $[A, A_1]_L = I_M$ to denote the strong left primeness of A and A_1. Similarly, given A in $H^{\infty}(B(N, M))$ and A_1 in $H^{\infty}(B(N, M_1))$ we define right primeness. We clearly have $(A, A_1)_R = I_N$ if and only if $(\tilde{A}, \tilde{A}_1)_L = I_N$ and $[A, A_1]_R = I_N$ if and only if $[A, A_1]_L = I_N$. Thus $[A, A_1]_R = I_N$ is equivalent to the existence of a $\delta > 0$ such that for all $z, |z| < 1$,

$$\inf\{\|A(z)\xi\| + \|A_1(z)\xi\| \,|\xi \in N, \|\xi\| = 1\} \geqq \delta. \tag{2.4}$$

If $[A, A_1]_L = I_N$, then by a matrix version of the Carleson corona theorem [4] we have the existence of functions $B \in H^{\infty}(B(N, M))$ and $B_1 \in H^{\infty}(B(N, M_1))$ such that

$$A(z)B(z) + A_1(z)B_1(z) = I_N.$$

Thus A and A_1 are left prime for any common left inner divisor of A and A_1 would be a left divisor of I_N and hence necessarily a trivial divisor that is a constant unitary matrix. Thus strong relative primeness implies left primeness and this justifies the terminology.

We define a map J in $L^2(N)$ by letting $(Jf)(e^{it}) = f(e^{-it})$ for all f in $L^2(N)$. For A in $H^\infty(B(N, M))$ we define the Hankel operator induced by it as the operator H_A from $H^2(N)$ into $H^2(M)$ given by $H_A f = P_{H^2(M)} A(Jf)$. Here $P_{H^2(M)}$ is the orthogonal projection of $L^2(M)$ onto $H^2(M)$. The range closure of H_A, $\overline{\text{range } H_A}$, is a left invariant subspace of $H^2(M)$. We shall say that A is *strictly noncyclic* if $\{\text{range } H_A\}^\perp$ is an invariant subspace of full range.

Let Ω be a domain in the complex plane, that is an open connected set. A $B(N, M)$-valued function F is meromorphic of bounded type in Ω if $F = G/g$ where G is a bounded $B(N, M)$-valued analytic function in Ω and g is a bounded scalar-valued analytic function in Ω. A principal tool in all that follows will be the following theorem [9]. For simplicity we will assume from now that M and N are finite-dimensional and hence the notion of determinant of inner functions is well defined.

THEOREM 2.1. (a) *The following statements are equivalent*:

(i) *A function A in $H^\infty(B(N, M))$ is strictly noncyclic.*

(ii) *A is a strong radial limit of a $B(N, M)$-valued meromorphic function of bounded type in $D_e = \{z | 1 < |z| \leqq \infty\}$.*

(iii) *On the unit circle A admits the factorizations*

$$A = \bar{\chi} P C^* = \bar{\chi} C_1^* P_1 \tag{2.5}$$

which satisfy the primeness relations

$$(P, C)_R = I_M \quad \text{and} \quad (P_1, C_1)_L = I_N. \tag{2.6}$$

Here P in $H^\infty(B(M, M))$ and P_1 in $H^\infty(B(N, N))$ are inner functions and C and C_1 are in $H^\infty(B(M, N))$.

(b) *The inner functions P and P_1 in the prime factorizations* (2.5) *are quasi-equivalent* [18], *and in particular*,

$$\det P = \det P_1 \tag{2.7}$$

holds.

In terms of the factorizations (2.5) we have $\overline{\text{range } H_A} = H(P)$ and $\ker H_A = \tilde{P}_1 H^2(N)$. By results of [7] we have range $H_A = H(P)$ if and only if the primeness condition $(P, C)_R = I_M$ is replaced by the strong primeness condition $[P, C]_R = I_M$. The factorizations (2.5) are the generalization of writing a rational function as the quotient of two relatively prime polynomials. They furthermore generalize the polynomial matrix factorizations appearing in Rosenbrock's theory of linear systems [19].

For an inner function P in $H^\infty(B(M, M))$ we define two operators $\Gamma(P)$ and $\gamma(P)$ from M into $H(P)$ by letting

$$\Gamma(P)\xi = P_{H(P)} \bar{\chi} P \xi \tag{2.8}$$

and

$$\gamma(P)\xi = P_{H(P)} \xi. \tag{2.9}$$

It is easy to check that

$$(\Gamma(P)\xi)(z) = ((P(z) - P(0))\xi)/z, \tag{2.10}$$

$$(\gamma(P)\xi)(z) = (I - P(z)P(0)^*)\xi \tag{2.11}$$

and

$$\gamma(P)^* f = f(0) \quad \text{for all } f \in H(P). \tag{2.12}$$

For the inner function P we define a map $\tau_P : L^2(N) \to L^2(\dot{N})$ by

$$\tau_P f = \bar{\chi}\tilde{P}(Jf), \tag{2.13}$$

that is,

$$(\tau_P f)(e^{it}) = e^{-it}\tilde{P}(e^{it}) f(e^{-it}). \tag{2.14}$$

It has been proved in [4] that τ_P is a unitary map of $L^2(N)$ for which

$$\tau_P(H(P)) = H(\tilde{P}) \tag{2.15}$$

and

$$\tau_P S(P)^* = S(\tilde{P})\tau_P. \tag{2.16}$$

The operators $\Gamma(P)$ and $\gamma(P)$ defined by (2.8) and (2.9) respectively are related by

$$\tau_P \Gamma(P) = \gamma(\tilde{P}) \tag{2.17}$$

and

$$\tau_P \gamma(P) = \Gamma(\tilde{P}). \tag{2.18}$$

These properties of the map τ_P make it extremely useful in the study of duality properties of restricted shift operators and systems.

Given a function A in $L^\infty(B(N, M))$ we let $M_A : N \to H^2(M)$ be defined by

$$M_A \eta = P_{H^2(M)}(A\eta); \tag{2.19}$$

thus $M_A = H_A | N$ where N is embedded in $H^2(N)$ in a natural way.

Suppose now that A is a strictly noncyclic function in $H^\infty(B(N, M))$ admitting the factorizations (2.5) on the unit circle which satisfy the relative primeness conditions (2.6).

We consider now the system

$$\Sigma_A = \{S(P)^*, M_A, \gamma(P)^*\} \tag{2.20}$$

in the state space $H(P)$. An elementary calculation shows that this system is a realization of A. Moreover, the controllability operator of this system coincides with H_A, whereas the observability operator $\mathcal{O} : H^2(M) \geqq H(P)$ is given by $\mathcal{O} = P_{H(P)}$ and hence by the characterization of exactly controllable and exactly observable systems obtained in [8] we infer that the system (2.20) is controllable and exactly observable. The system is exactly controllable if the condition $(P, C)_R = I_M$ is replaced by the stronger condition $[P, C]_R = I_M$. We shall refer to the realization (2.20) as the shift realization of A [14], [8]. If we assume $A \in L^\infty(B(N, M))$, then the shift realization constructed is a realization of the analytic part of A.

Together with the shift realization of A it is convenient to consider a related realization obtained as follows. Since $\tilde{A}$ is strictly noncyclic if and only if A is, then

$\tilde{A}$ has also a shift realization given by the system

$$\{S(\tilde{P}_1)^*, M_{\tilde{A}}, \gamma(P_1)^*\} \tag{2.21}$$

which acts in the state space $H(\tilde{P}_1)$. As before this realization is controllable and exactly observable. By passing to the adjoint system

$$\{S(\tilde{P}_1), \gamma(\tilde{P}_1), M_{\tilde{A}}^*\}, \tag{2.22}$$

we obtain a realization of A which is observable and exactly controllable. We apply now the unitary map $\tau_{\tilde{P}_1}: H(\tilde{P}_1) \to H(P_1)$ which was defined by (2.12). Since

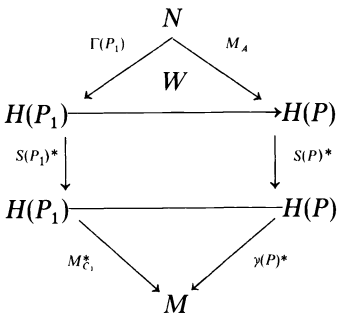
$$M_{\tilde{A}\tilde{P}_1}^* = M_{C_1}^* \tag{2.23}$$

and the relations (2.16) and (2.17) hold, it follows that the system

$$\Sigma_A' = \{S(P_1)^*, \Gamma(P_1), M_{C_1}^*\} \tag{2.24}$$

acting in the state space $H(P_1)$ is a realization of A which is observable and exactly controllable. We refer to the realization Σ_A' given by (2.24) as the *-shift realization of A. The relation between the shift and *-shift realizations of A is important. In fact from results of Moore [16] it follows that there exists a one-to-one operator W with dense range which makes the diagram in (2.25) commutative. This in turn

(2.25)



implies that the two shifts $S(P)^*$ and $S(P_1)^*$ are quasi-similar and, by the work of Moore and Nordgren [17], that the two inner functions P and P_1 are quasi-equivalent and thus have the same Jordan model associated with them. These considerations play a central role in the proof of Theorem 2.1.

Since the state space $H(P)$ in the shift realization of A is determined by the inner function P, there must be a relation between the dimension of $H(P)$ and P itself. In fact $p = \det P$ is a scalar inner function and P is a finite Blaschke product if and only if $H(P)$ is finite-dimensional. Moreover, dim $H(P)$ in that case is equal to the number of factors in p, multiplicities counted. The determinant of inner functions will be used from here on as a multiplicative substitute for dimension and it is the measure of size to which we referred in the Introduction. If P and R are two inner functions in $H^\infty(B(M, M))$, then there exists an inner function Q, unique up to a constant unitary factor on the right, for which $QH^2(M) = PH^2(M) \cap RH^2(M)$. The relation $\det Q | (\det P) \cdot (\det R)$ is always satisfied with equality if and only if $(P, R)_L = I_M$. This later condition is equivalent to $H(P) \cap H(R) = \{0\}$, or $H(Q)$ is equal to the span of the two subspaces $H(P)$ and $H(R)$. $H(Q)$ is actually equal to the, nonorthogonal, direct sum of $H(P)$ and $H(R)$ if and only if $[P, R]_L = I_M$. From this point it is clear that $\det Q = (\det P) \cdot (\det R)$ is really the multiplicative

equivalent in the context of left invariant subspaces of the additive relation $\dim(M_1 + M_2) = \dim M_1 + \dim M_2$ for the sum of two finite-dimensional subspaces for which $M_1 \cap M_2 = \{0\}$. From Theorem 1 it follows that the state spaces $H(P)$ and $H(P_1)$ of the shift and *-shift realizations satisfy $\det P = \det P_1$, i.e., an "equidimensionality" condition. The above discussion should be compared with a proof of the resultant theorem of which it is a generalization.

3. On the shift realization of sums and products of transfer functions. This section is devoted to the study of the shift realizations of sums and products of strictly noncyclic functions. Since all the state spaces appearing are left invariant subspaces which are associated with inner functions, we use the determinant of the corresponding inner function as a measure of the "dimensionality" of the space. We will be interested in: Under what conditions is the determinant of the inner function associated with the shift realization of a product (or a sum) equal to the product of the determinants of the inner functions associated with the shift realizations of the individual factors in the product? This is closely related to the series (or parallel) connection of systems and the representation of the state space as a direct sum of the state spaces of the component systems. This problem is deferred to §4. Intuitively the determinant condition referred to before fails to hold if and only if there are some "zero-pole" cancellations. In terms of the representation of rational functions as quotients of polynomials this is equivalent to some polynomials having nontrivial common divisors. In the multivariable finite-dimensional case the problem can be handled through the use of polynomial matrix factorizations of matrix rational functions [19]. For the details we refer to [2], [11]. In our approach the factorizations appearing in Theorem 2.1 will be used. The results of this section have been derived in [10] where the full details can be found. Since the state spaces appearing in shift realizations are range closures of Hankel operators we will study these. We begin with products.

Let L, M, N be finite-dimensional Hilbert spaces and let A belong to $H^\infty(B(N, M))$ and B to $H^\infty(B(L, N))$, and assume both are strictly noncyclic having the following factorizations on the unit circle:

$$A = \bar{\chi} P C^* = \bar{\chi} C_1^* P_1 \tag{3.1}$$

and

$$B = \bar{\chi} R D^* = \bar{\chi} D_1^* R_1, \tag{3.2}$$

where P, P_1, R and R_1 are inner functions and C, C_1, D and D_1 bounded analytic. Moreover, we shall assume that the primeness conditions

$$(P, C)_R = I_M, \qquad (P_1, C_1)_L = I_N \tag{3.3}$$

and

$$(R, D)_R = I_N, \qquad (R_1, D_1)_L = I_L \tag{3.4}$$

are satisfied. Since both A and B are noncyclic so is the function χAB. The extra χ appears for reasons which become clear in the next section. Applying Theorem 2.1 we see that the function χAB has the factorizations

$$\chi AB = \bar{\chi} Q H^* = \bar{\chi} H_1^* Q_1 \tag{3.5}$$

satisfying

$$(Q, H)_R = I_M, \qquad (Q_1, H_1)_L = I_L. \tag{3.6}$$

The analysis of the general case follows from the two special cases where $B = \bar{\chi}R$ or $B = \bar{\chi}D^*$ respectively. In general we have range $H_{AD^*} \subset$ range $H_A \subset$ range H_{AR}. Thus multiplication on the right by an inner function R, which extends meromorphically to the exterior of the unit disc and has no zeros there, increases the number of singularities and hence also the range of the corresponding Hankel operator. On the other hand, D^* extends analytically to the exterior of the unit disc and this tends to decrease the singularities of the product as well as the range of the Hankel operator. The precise conditions for the noncancellation of singularities are given below.

THEOREM 3.1. *Let A and B be strictly noncyclic functions in $H^\infty(B(N, M))$ and $H^\infty(B(L, N))$ respectively which have the factorizations* (3.1) *and* (3.2) *satisfying the primeness conditions* (3.3) *and* (3.4). *Let χAB have the factorization* (3.5) *satisfying* (3.6).

(a) *A necessary and sufficient condition for*

$$\det Q = (\det P) \cdot (\det R) \tag{3.7}$$

to hold is that

$$(C, R)_L = I_N \quad \text{and} \quad (P_1, D_1)_R = I_N \tag{3.8}$$

are satisfied.

(b) *Assume the factorizations* (3.1) *and* (3.2) *satisfy*

$$[P, C]_R = I_M, \qquad [P_1, C_1]_L = I_N \tag{3.9}$$

and

$$[R, D]_R = I_N, \qquad [R_1, D_1]_L = I_L \tag{3.10}$$

respectively. A necessary and sufficient condition for $H_{\chi AB}$ to have closed range $H(Q)$ with (3.7) *satisfied is that*

$$[C, R]_L = I_N \quad \text{and} \quad [P_1, D_1]_R = I_N \tag{3.11}$$

hold.

In terms of the shift realization described in the previous section the theorem can be rewritten in the following form.

THEOREM 3.2. *Let A and B be strictly noncyclic functions in $H^\infty(B(N, M))$ and $H^\infty(B(L, N))$ respectively which have the prime factorizations* (3.1) *and* (3.2). *The shift realization of χAB has a state space $H(Q)$ with* (3.7) *satisfied if and only if the primeness conditions* (3.8) *are satisfied. If the shift realizations of A and B are both exactly controllable and exactly observable then the shift realization of χAB is exactly controllable and exactly observable with* (3.7) *satisfied if and only if the conditions* (3.11) *hold.*

The additive results are presented next.

THEOREM 3.3. *Let A and B be strictly noncyclic functions in $H^\infty(B(N, M))$ having the prime factorizations* (3.1) *and* (3.2). *Let $A + B$, which is also strictly noncyclic, have the prime factorization*

$$A + B = \chi Q H^* = \bar{\chi} H_1^* Q_1. \tag{3.12}$$

A necessary and sufficient condition for (3.7) *to hold is that*

$$(P, R)_L = I_M \quad \text{and} \quad (P_1, R_1)_R = I_N \tag{3.13}$$

are satisfied.

Again this can be interpreted in terms of the shift realization.

THEOREM 3.4. *Let A and B be strictly noncyclic functions in* $H^\infty(B(N, M))$ *which have the prime factorizations* (3.1) *and* (3.2). *The shift realization of* $A + B$ *has a state space* $H(Q)$ *with* (3.7) *satisfied if and only if the primeness conditions* (3.14) *are satisfied. If the shift realizations of A and B are both exactly controllable and exactly observable, then the shift realization of* $A + B$ *is exactly controllable and exactly observable with* (3.7) *satisfied if and only if the conditions* (3.14) *hold.*

4. Series and parallel coupling of linear systems. We come now to the central topic of this paper, the study of the series and parallel coupling of the shift realizations of two transfer functions. This will be achieved by comparing the coupled systems with the shift realizations of the product and sum of the corresponding transfer functions.

We shall begin by introducing some new concept and deriving some simple results needed in the sequel.

DEFINITION 4.1. Let Σ and Σ_1 be two realizations given by $\{F, G, H\}$ and $\{F_1, G_1, H_1\}$ having the state space K and K_1 respectively. Let $X: K \to K_1$ be a bounded operator for which the relations

$$XF = F_1 X, \quad XG = G_1 \quad \text{and} \quad H = H_1 X \tag{4.1}$$

hold. In that case we say that X *intertwines* Σ and Σ_1. If only $XF = F_1 X$ and $XG = G_1$ we say that X *intertwines* $\{F, G\}$ and $\{F_1, G_1\}$ and similarly for intertwining the observability part.

A bounded operator $X: K \to K_1$ will be called a quasi-affinity if X is one-to-one and has dense range. This is a slight relaxation of the notion of invertibility.

DEFINITION 4.2. (a) The system Σ_1 is a *quasi-affine transform* of Σ if there exists a quasi-similarity X that intertwines Σ and Σ_1.

(b) Two systems Σ and Σ_1 are *quasi-similar* if each one is the quasi-affine transform of the other.

(c) Two systems Σ and Σ_1 are *similar* if there exists a boundedly invertible operator X that intertwines Σ and Σ_1. (Then X^{-1} intertwines Σ_1 and Σ).

Whereas the quasi-similarity of two operators does not imply their similarity [21] this is the case with controllable systems.

LEMMA 4.1. *Let* $\Sigma = \{F, G, H\}$ *and* $\Sigma_1 = \{F_1, G_1, H_1\}$ *be two controllable systems; then they are quasi-similar if and only if they are similar.*

Proof. Of course similarity implies quasi-similarity. Let $X: K \to K_1$ and $Y: K_1 \to K$ be the two quasi-affinities that intertwine the two systems. From relations (4.1) it follows that $XF^nG = F_1^n G_1$ and $YF_1^n G_1 = F^n G$ and hence $YXF^nG = F^nG$ and $XYF_1^n G_1 = F_1^n G_1$. Since the set of all vectors of the form $\Sigma F^n G u_n$ and $\Sigma F_1^n G_1 u_n$ are, by the controllability assumption, dense in K and K_1 respectively, then clearly $XY = I_{K_1}$ and $YX = I_K$ and hence the similarity. Of course the result holds if the assumption of controllability is replaced by observability.

The proof of the next two lemmas is equally simple but their use is central to all that follows:

LEMMA 4.2. *Let $X: K \to K_1$ intertwine the systems Σ and Σ_1.*

(a) *If X has dense range, then the controllability of Σ implies the controllability of Σ_1. If X is onto and Σ is exactly controllable, then Σ_1 is exactly controllable.*

(b) *If X is one-to-one, then the observability of Σ_1 implies the observability of Σ. If X^* is onto and Σ_1 is exactly observable, then so is Σ.*

Proof. (a) Let $z \in \bigcap_{n \geq 0} \ker G_1^* F_1^{*n}$; then for all n we have $G_1^* F_1^{*n} z = 0$. From (4.1) it follows that $G_1^* F_1^{*n} = G^* F^{*n} X^*$ and hence $X^* z \in \bigcap \ker G^* F^{*n} = \{0\}$. Since X has dense range, X^* is one-to-one and so $z = 0$ and $\bigcap \ker G_1^* F_1^{*n} = \{0\}$ which shows the controllability of Σ_1. If X is onto and C and C_1 are the controllability operators of Σ and Σ_1 respectively, then $C_1 = XC$; and hence if C is onto K, then C_1 is onto K_1 and Σ_1 is exactly controllable. Part (b) follows from (a) by duality.

LEMMA 4.3. *Given two systems $\Sigma = \{F, G, H\}$ and $\Sigma_1 = \{F_1, G_1, H_1\}$, let $X: K \to K_1$.*

(a) *If X intertwines $\{F, G\}$ and $\{F_1, G_1\}$ and Σ_1 is controllable, then X has dense range. If Σ_1 is exactly controllable, X is onto.*

(b) *If X intertwines $\{F, H\}$ and $\{F_1, H_1\}$ and Σ is observable, then X is one-to-one. If Σ is exactly observable, then X has a bounded left inverse.*

Proof. Part (a) follows from the definitions. To prove (b) we note that $H_1 F_1^n X x = HF^n x$ for each x in K. Thus $Xx = 0$ implies $x \in \bigcap_{n \geq 0} \ker HF^n$, and hence $x = 0$ by the observability assumption.

From the above lemmas it is clear that the explicit construction of intertwining maps between realizations enables us to study a system in terms of its relation to another system with known properties. Also knowledge of the systems sheds light on the properties of intertwining maps. We begin by studying the series coupling of two shift realizations.

Let L, N, M be three finite-dimensional complex Hilbert spaces. Let $A \in H^\infty(B(N, M))$ and $B \in H^\infty(B(L, N))$ be two strictly noncyclic functions, having the prime factorizations (3.1) and (3.2) respectively. Their shift realizations have state spaces $H(P)$ and $H(R)$ respectively and are given by

$$\Sigma_A = \{S(P)^*, M_A, \gamma(P)^*\} \tag{4.2}$$

and

$$\Sigma_B = \{S(R)^*, M_B, \gamma(R)^*\} \tag{4.3}$$

respectively, where Σ_A and Σ_B are defined as in (2.20).

Their series connection has $H(R) \oplus H(P)$ as state space: and the coupled system, which we will denote by $\Sigma_A \Sigma_B$, is given in terms of this direct sum by

$$\Sigma_A \Sigma_B = \left\{ \begin{pmatrix} S(R)^* & 0 \\ M_A \gamma(R)^* & S(P)^* \end{pmatrix}, \begin{pmatrix} M_B \\ 0 \end{pmatrix}, (0 \;\; \gamma(P)^*) \right\}, \tag{4.4}$$

and has χAB as its transfer function.

Assume first that $(C, R)_L = I_M$ which, by Theorem 2.1, implies the existence of a function $C' \in H^\infty(B(M, N))$ and an inner function $R' \in H^\infty(B(M, M))$ for which

$C^*R = R'C'^*$, $(R', C')_R = I_M$ and $\det R = \det R'$ are satisfied. Now from the factorizations (3.1) and (3.2) of A and B respectively it follows that $\chi AB = \bar{\chi}PC^*RD^* = \bar{\chi}PR'C'^*D^*$ and this enables us to produce a shift realization of χAB. In fact if we choose $H(PR')$ as the state space and consider the system $\Sigma_{\chi AB}$ given by

$$\Sigma_{\chi AB} = \{S(PR')^*, M_{\chi AB}, \gamma(PR')^*\}, \tag{4.5}$$

then we obtain a realization of χAB. This realization is clearly exactly observable but not necessarily controllable. Its controllability is equivalent to the equality $H(PR') = \overline{\text{range } H_{\chi AB}}$ which in turn is equivalent to the relative primeness condition $(P_1, D_1)_R = I_N$. We proceed now with a more detailed analysis of the realization (4.5).

By a lemma of Ahern and Clark [1] which has an immediate generalization to the vector-valued case the left invariant subspace $H(PR')$ has a direct sum decomposition of the form

$$H(PR') = H(P) \oplus PH(R'). \tag{4.6}$$

Hence we have an isometric isomorphism of $H(PR')$ onto $H(R') \oplus H(P)$ given by $f = g + Ph \to h \oplus g$. Here $g + Ph$ is the unique decomposition of f in $H(PR')$ with respect to the direct sum (4.6).

From the above representation of $f \in H(PR')$ we have

$$\begin{aligned}(f(z) - f(0))/z &= (g(z) - g(0))/z + (P(z)h(z) - P(0)h(0))/z \\ &= (g(z) - g(0))/z + P(z)(h(z) - h(0))/z + (P(z) - P(0))h(0)/z,\end{aligned}$$

and hence,

$$S(PR')f = S(P)^*g + PS(R')^*h + \Gamma(P)\gamma(R')^*h. \tag{4.7}$$

Next we have $f(0) = g(0) + P(0)h(0)$ which implies

$$\gamma(PR')^*f = \gamma(P)^*g + P(0)\gamma(R')^*h. \tag{4.8}$$

Finally for $\xi \in L$ let $M_{\chi AB}\xi = \chi AB\xi = g + Ph$, with $g \in H(P)$ and $h \in H(R')$. Then we have

$$h = P_{H^2(M)}C^*B\xi = P_{H^2(M)}C^*M_B\xi \tag{4.9}$$

and

$$g = M_{\chi AB_\xi} - P \cdot P_{H^2(M)}C^*M_B\xi. \tag{4.10}$$

In conclusion relations (4.7)–(4.10) imply that with respect to the direct sum $H(R') \oplus H(P)$ the shift realization of χAB is given by

$$\left\{\begin{pmatrix} S(R')^* & 0 \\ \Gamma(P)\gamma(R')^* & S(P) \end{pmatrix}, \begin{pmatrix} P_{H^2(M)}C^*M_B \\ M_{\chi AB} - P \cdot P_{H^2(M)}C^*M_B \end{pmatrix}, (P(0)\gamma(R')^*\gamma(P)^*)\right\}. \tag{4.11}$$

We shall construct now a map

$$X : H(R) \oplus H(P) \to H(R') \oplus H(P)$$

which intertwines $\Sigma_A\Sigma_B$ and $\Sigma_{\chi AB}$. A comparison of the generators in the system

given by (4.4) and (4.11), both in lower triangular form, indicates that an intertwining operator X may exist of the form

$$X = \begin{pmatrix} W & 0 \\ Z & I \end{pmatrix}, \tag{4.12}$$

where $W: H(R) \to H(R')$ and $Z: H(R) \to H(P)$ are bounded. For the operator W we have as a natural candidate the quasi-affinity that intertwines the *-shift and shift realizations of the analytic part of the function $E = \bar{\chi}C^*R = \bar{\chi}R'C'^*$. These two realizations, in the state spaces $H(R)$ and $H(R')$, are given by

$$\Sigma'_E = \{S(R)^*, \Gamma(R), M_C^*\} \tag{4.13}$$

and

$$\Sigma_E = \{S(R')^*, M_E, \gamma(R')^*\} \tag{4.14}$$

respectively. From the commutativity of a diagram analogous to (2.25) we have that

$$W\Gamma(R)\xi = WP_{H^2(N)}\bar{\chi}R\xi = P_{H^2(M)}\bar{\chi}C^*R\xi \tag{4.15}$$

and since

$$WS(R)^* = S(R')^*W \tag{4.16}$$

holds, it follows that for each $f \in H(R)$ we have

$$Wf = P_{H(R')}C^*f = P_{H^2(M)}C^*f \tag{4.17}$$

and we take (4.17) to be the definition of W. The primeness conditions $(R, C)_L = I_N$ and $(R', C')_R = I_M$ guarantee that W is indeed a quasi-affinity [5]. Clearly X defined by (4.12) is a quasi-affinity if and only if W is. X intertwines the two systems $\Sigma_A\Sigma_B$ and $\Sigma_{\chi AB}$ if and only if the following relations hold:

$$WS(R)^* = S(R')^*W, \tag{4.18}$$

$$ZS(R)^* + M_A\gamma(R)^* = \Gamma(P)\gamma(R')^*W + S(P)^*Z, \tag{4.19}$$

$$P(0)\gamma(R')^*W + \gamma(P)^*Z = 0, \tag{4.20}$$

$$WM_B = P_{H^2(M)}C^*M_B \tag{4.21}$$

and

$$ZM_B = M_{\chi AB} - P \cdot P_{H^2(M)}C^*M_B. \tag{4.22}$$

Obviously (4.18) and (4.21) follow from the definition of W as given by (4.17). We define $Z: H(R) \to H(P)$ by

$$Zf = P_{H(P)}PC^*f \quad \text{for} \quad f \in H(R) \tag{4.23}$$

which immediately implies (4.22).

Now $P(0)\gamma(R')^*Wf = (P \cdot P_{H(R')}C^*f)(0)$, and hence,

$$\begin{aligned}(P(0)\gamma(R')^*W + \gamma(P)^*Z)f &= (P \cdot P_{H(R')}C^*f + P_{H(P)}PC^*f)(0) \\ &= (PC^*f)(0) = (\chi Af)(0) = 0.\end{aligned}$$

This proves (4.20) and it remains to prove (4.19). Let $f \in H(R)$; then

$$(ZS(R)^* - S(P)^*Z)f = P_{H(P)}PC^*P_{H^2(N)}\bar{\chi}f - P_{H(P)}\bar{\chi}P_{H(P)}PC^*f.$$

As $A = PC^*$ is analytic in the unit disc PC^*f belongs to $H(PR')$ and

$$P_{H(P)}PC^*f = PC^*f - P \cdot P_{H(R')}C^*f = PC^*f - PWf.$$

Therefore it follows that

$$\begin{aligned} P_{H(P)}\bar{\chi}PWf &= P_{H(P)}P_{H^2(M)}\bar{\chi}PWf \\ &= P_{H(P)}\{PS(R')^*Wf + \Gamma(P)\gamma(R')^*f\} = \Gamma(P)\gamma(R')^*f. \end{aligned}$$

Moreover,

$$\begin{aligned} P_{H(P)}PC^*P_{H^2(N)}\bar{\chi}f - P_{H(P)}PC^*f &= P_{H(P)}\bar{\chi}PC^*(f - f(0)) - P_{H(P)}\bar{\chi}PC^*f \\ &= -P_{H(P)}PC^*f(0) = -Af(0) = -M_A\gamma(R)^*f \end{aligned}$$

which proves (4.19).

To prove the necessity of the condition $(R, C)_L = I_N$ for the observability of the coupled system $\Sigma_A\Sigma_B$ we state first a simple lemma, omitting the proof.

LEMMA 4.4. *Given two systems* $\Sigma = \{F, G, H\}$ *and* $\Sigma_1 = \{F_1, G_1, H_1\}$:

(a) *If* Σ *is controllable, then if there exists a bounded operator intertwining* Σ *and* Σ_1 *then it is unique.*

(b) *If* Σ_1 *is observable, then if there exists a bounded operator intertwining* Σ *and* Σ_1 *then it is unique.*

Assume now that the series coupling $\Sigma_A\Sigma_B$ is observable. The map X defined by (4.12), (4.17) and (4.23) intertwines $\Sigma_A\Sigma_B$ and $\Sigma_{\chi AB}$ and, since $\Sigma_{\chi AB}$ is observable, it follows from the previous lemma that this is the only intertwining map. Since $\Sigma_A\Sigma_B$ is assumed observable the intertwining map X, by Lemma 4.3(b), is necessarily a one-to-one map. Now because of its special triangular structure X is one-to-one if and only if W is one-to-one. Now for W as defined by (4.17) to be one-to-one it is necessary that $(R, C)_L = I_N$ holds. In fact if R and C have a nontrivial common left inner factor S, then $R = SR''$ and $H(R) = H(S) \oplus SH(R'')$. It is clear that $W|H(S) = 0$ and hence W in that case is not one-to-one.

The analysis carried out above can be translated, by way of duality considerations, to the series connection of the *-shift realization of A and B. Let us denote the *-shift realizations of A and B by Σ'_A and Σ'_B respectively. Since Σ'_A is, by the construction in § 2, unitarily equivalent to $\Sigma^*_{\tilde{A}}$, the adjoint of the shift realization of $\tilde{A}$ it follows that the series connection $\Sigma'_A\Sigma'_B$ of the *-shift realizations of A and B is unitarily equivalent to $(\Sigma_{\tilde{B}}\Sigma_{\tilde{A}})^*$, the adjoint of the series connection of the shift realizations of $\tilde{B}$ and $\tilde{A}$. Hence controllability properties of $\Sigma'_A\Sigma'_B$ are equivalent to observability properties of $\Sigma_{\tilde{B}}\Sigma_{\tilde{A}}$. The map X that intertwines $\Sigma_A\Sigma_B$ and $\Sigma_{\chi AB}$ has its analogue in a map X' that intertwines $\Sigma_{\chi AB}$ and $\Sigma'_A\Sigma'_B$. Moreover there is always a quasi-affinity intertwining Σ'_A and Σ_A and another intertwining Σ'_B and Σ_B. The direct sum of these quasi-affinities is a quasi-affinity Ξ which intertwines $\Sigma'_A\Sigma'_B$ and $\Sigma_A\Sigma_B$. Thus we have the following diagram:

$$\Sigma'_{\chi AB} \xrightarrow{X'} \Sigma'_A\Sigma'_B \xrightarrow{\Xi} \Sigma_A\Sigma_B \xrightarrow{X} \Sigma_{\chi AB}.$$

We note that X has dense range by construction and is one-to-one if $(R, C)_L = I_N$ holds. X' is one-to-one by construction and has dense range if $(P_1, D_1)_R = I_N$ holds. X is boundedly invertible if $(R, C)_L = I_N$ is replaced by $[R, C]_L = I_N$ and X' is boundedly invertible if $(P_1, D_1)_R = I_N$ is replaced by $[R, C]_L = I_N$ and X' is boundedly invertible if $(P_1, D_1)_R = I_N$ is replaced by $[P_1, D_1]_R = I_N$. The map Ξ becomes boundedly invertible if and only if the quasi-affinities intertwining the *-shift and the shift realizations of A and B respectively are actually boundedly invertible. This is equivalent to the exact controllability and exact observability of Σ_A, Σ'_A, Σ_B and Σ'_B. By results of [8] this is equivalent to replacing the primeness conditions (3.3) and (3.4) by (3.9) and (3.10) respectively.

We summarize the above analysis in the following theorem.

THEOREM 4.1. *Let L, M and N be finite-dimensional Hilbert spaces and let $A \in H^\infty(B(N, N))$ and $B \in H^\infty(B(L, N))$ be two strictly noncyclic functions having the prime factorization* (3.1) *and* (3.2), *respectively.*

(a_1) *The series coupling $\Sigma_A\Sigma_B$ of the shift realizations of A and B is observable if and only if $(R, C)_L = I_N$ holds, and exactly observable if and only if $[R, C]_L = I_N$ holds.*

(a_2) *The series coupling $\Sigma'_A\Sigma'_B$ of the *-shift realizations of A and B is controllable if and only if $(P_1, D_1)_R = I_N$ holds, and exactly controllable if and only if $[P_1, D_1]_R = I_N$ holds.*

(b_1) *A sufficient condition for the controllability of $\Sigma_A\Sigma_B$ is $(P_1, D_1)_R = I_N$. This condition is also necessary if Ξ is boundedly invertible. If Σ_A and Σ_B are both exactly controllable, then $\Sigma_A\Sigma_B$ is exactly controllable if and only if $[P_1, D_1]_R = I_N$.*

(b_2) *A sufficient condition for the observability of $\Sigma'_A\Sigma'_B$ is $(R, C)_L = I_N$. This condition is also necessary if Ξ is boundedly invertible. If Σ'_A and Σ'_B are both exactly observable, then $\Sigma'_A\Sigma'_B$ is exactly observable if and only if $[R, C]_L = I_N$.*

We note in passing that in case of finite-dimensional systems the map Ξ is always boundedly invertible and that by application of the state space isomorphism theorem the above results hold for the series connection of any two canonical finite-dimensional systems.

We pass now on to the analysis of the parallel connection of two shift realizations. This problem is easier to handle as the parallel coupling of two shift systems is also a shift system and hence the theorem in [8] characterizing analytically the controllability and observability properties of these systems can be applied directly. So let us assume now that A and B are two strictly noncyclic functions in $H^\infty(B(N, M))$ having the prime factorizations (3.1) and (3.2), respectively. Their shift realizations have state spaces $H(P)$ and $H(R)$ and are given by (4.2) and (4.3) respectively.

We denote by $\Sigma_A + \Sigma_B$ the parallel connection of the shift realizations Σ_A and Σ_B. By this we mean the system given by

$$\left\{\begin{pmatrix} S(P)^* & 0 \\ 0 & S(R)^* \end{pmatrix}, \begin{pmatrix} M_A \\ M_B \end{pmatrix}, (\gamma(P)^* \quad \gamma(R)^*)\right\} \tag{4.24}$$

acting in the state space $H(P) \oplus H(R)$ which is a left invariant subspace of $H^2(M \oplus M)$. The inner function associated with $H(P) \oplus H(R)$ has a natural matrix representation of the form

$$\begin{pmatrix} P & 0 \\ 0 & R \end{pmatrix}. \tag{4.25}$$

By Theorem 3.4 in [7] the system $\Sigma_A + \Sigma_B$ is observable if and only if

$$\left(\begin{pmatrix} P & 0 \\ 0 & R \end{pmatrix}, \begin{pmatrix} I_M \\ I_M \end{pmatrix} \right)_L = I_{M \oplus M}, \tag{4.26}$$

and it is exactly observable if and only if

$$\left[\begin{pmatrix} P & 0 \\ 0 & R \end{pmatrix}, \begin{pmatrix} I_M \\ I_M \end{pmatrix} \right]_L = I_{M \oplus M}. \tag{4.27}$$

We shall show now that condition (4.26) is equivalent to

$$(P, R)_L = I_M \tag{4.28}$$

whereas condition (4.27) is equivalent to

$$[P, R]_L = I_M. \tag{4.29}$$

Let us prove first a simple lemma.

LEMMA 4.5. *An inner function in* $H^\infty(B(M \oplus M, M \oplus M))$ *is a left factor of* $\begin{pmatrix} I_M \\ I_M \end{pmatrix}$ *if and only if it has up to a constant unitary factor on the right, the form*

$$\frac{1}{2} \begin{pmatrix} I_M + S & I_M - S \\ I_M - S & I_M + S \end{pmatrix} \tag{4.30}$$

for some inner function S *in* $H^\infty(B(M, M))$.

Proof. Let S be inner; then clearly the function given by (4.30) is also inner, and moreover, since

$$\begin{pmatrix} I_M \\ I_M \end{pmatrix} = \frac{1}{2} \begin{pmatrix} I_M + S & I_M - S \\ I_M - S & I_M + S \end{pmatrix} \begin{pmatrix} I_M \\ I_M \end{pmatrix} \tag{4.31}$$

it is a left inner factor of $\begin{pmatrix} I_M \\ I_M \end{pmatrix}$.

To prove the converse we consider the constant unitary operator in $M + M$ which is defined through its matrix representation

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} I_M & I_M \\ I_M & -I_M \end{pmatrix}. \tag{4.32}$$

U can be naturally extended to a unitary operator in $H^2(M + M)$ and we have

$$U \begin{pmatrix} I_M \\ I_M \end{pmatrix} = \sqrt{2} \begin{pmatrix} I_M \\ 0 \end{pmatrix}.$$

Thus an inner function Q in $H^\infty(B(M \oplus M, M \oplus M))$ is a left factor of $\begin{pmatrix} I_M \\ I_M \end{pmatrix}$

if and only if UQ is a left factor of $\begin{pmatrix} I_M \\ 0 \end{pmatrix}$. Now the left inner factors of $\begin{pmatrix} I_M \\ 0 \end{pmatrix}$ are those associated with right invariant subspaces of full range of $H^2(M \oplus M)$ which contain $H^2(M) \oplus \{0\}$. These subspaces are clearly of the form $H^2(M) \oplus SH^2(M)$ where S is inner in $H^\infty(B(M, M))$, and hence the corresponding inner functions have the representation

$$\begin{pmatrix} I_M & 0 \\ 0 & S \end{pmatrix}. \tag{4.33}$$

From here it follows that

$$Q = U^* \begin{pmatrix} I_M & 0 \\ 0 & S \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} I_M & S \\ I_M & -S \end{pmatrix}.$$

Since Q is unique up to multiplication on the right by a constant unitary matrix, by right multiplying with U we obtain the representation (4.30).

LEMMA 4.6. *The relative left primeness conditions* (4.26) *and* (4.27) *are equivalent to conditions* (4.28) *and* (4.29) *respectively.*

Proof. Let S be a common left inner factor of P and R. Thus $P = SP_1$ and $R = SR_1$. Since

$$\begin{aligned} \begin{pmatrix} P & 0 \\ 0 & R \end{pmatrix} &= \frac{1}{2} \begin{pmatrix} I_M + S & I_M - S \\ I_M - S & I_M + S \end{pmatrix} - \frac{1}{2} \begin{pmatrix} I_M + S & S - I_M \\ S - I_M & I_M + S \end{pmatrix} \begin{pmatrix} P_1 & 0 \\ 0 & R_1 \end{pmatrix} \\ &= \begin{pmatrix} SP_1 & 0 \\ 0 & SR_1 \end{pmatrix}, \end{aligned}$$

then together with (4.31) it follows that the inner function given by (4.30) is a nontrivial left inner factor of $\begin{pmatrix} P & 0 \\ 0 & R \end{pmatrix}$ and $\begin{pmatrix} I_M \\ I_M \end{pmatrix}$.

Conversely assume $\begin{pmatrix} P & 0 \\ 0 & R \end{pmatrix}$ and $\begin{pmatrix} I_M \\ I_M \end{pmatrix}$ have a common left factor. By Lemma 4.3 it must be of the form (4.30). Thus

$$\begin{pmatrix} P & 0 \\ 0 & R \end{pmatrix} = \frac{1}{2} \begin{pmatrix} I_M + S & I_M - S \\ I_M - S & I_M + S \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

and it follows that necessarily $P = S(A - C)$ and $R = S(D - B)$ which shows that S is a common left factor of P and R.

Next we prove the equivalence of the strong relative primeness conditions. By a generalization of the Carleson corona theorem [2] to the case of bounded matrix-valued analytic functions [4] the condition $[A, B]_L = I_N$ is equivalent to the existence of bounded analytic functions A_1 and B_1 for which $AA_1 + BB_1 = I_N$ holds. Thus if $[P, R]_L = I_M$ it follows that there exist Γ and Δ in $H^\infty(B(M, M))$ such that $P\Gamma + R\Delta = I_M$ holds. This in turn implies that

$$\begin{pmatrix} P & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} \Gamma & -\Gamma \\ -\Delta & \Delta \end{pmatrix} + \begin{pmatrix} I_M \\ I_M \end{pmatrix} \begin{pmatrix} R\Delta & P\Gamma \end{pmatrix} = \begin{pmatrix} I_M & 0 \\ 0 & I_M \end{pmatrix}$$

which proves (4.27).

Conversely assume (4.27) holds. Thus there exist bounded analytic functions

$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ and $(E \quad F)$ for which

$$\begin{pmatrix} P & 0 \\ 0 & R \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix} + \begin{pmatrix} I_M \\ I_M \end{pmatrix}(E \quad F) = \begin{pmatrix} I_M & 0 \\ 0 & I_M \end{pmatrix}.$$

This implies the equality $PA - RC = I_M$, and hence also, by another application of the corona result, that (4.29) holds.

Applying the preceding analysis to the parallel connection of shift systems we obtain the following.

THEOREM 4.2. *Let A and B be strictly noncyclic functions in $H^\infty(B(N, M))$ having the prime factorizations* (3.1) *and* (3.2) *respectively. The parallel connection $\Sigma_A + \Sigma_R$ of the two shift realizations Σ_A and Σ_B is observable if and only if the relative primeness condition*

$$(P, R)_L = I_M \tag{4.34}$$

holds, and it is exactly observable if and only if

$$[P, R]_L = I_M \tag{4.35}$$

holds.

Elementary duality considerations applied to Theorem 4.2 yield the next theorem as a direct corollary. We will denote again by Σ'_A the *-shift realization of A.

THEOREM 4.3. *Let A and B be strictly noncyclic functions in $H(B(N, M))$ having prime factorizations* (3.1) *and* (3.2), *respectively. The parallel connection $\Sigma'_A + \Sigma'_B$ of the two *-shift realizations of A and B is controllable if and only if*

$$(P_1, R_1)_R = I_N \tag{4.36}$$

holds, and it is exactly controllable if and only if

$$[P_1, R_1] = I_N \tag{4.37}$$

holds.

Let us denote now by X_A the map that intertwines the *-shift and shift realizations of A, i.e., the unique map that makes the diagram (2.25) commutative. Similarly we define X_B. Both X_A and X_B are quasi-affinities and hence the map $X_A \oplus X_B$ is quasi-affinity from $H(P_1) \oplus H(R_1)$ into $H(P) \oplus H(R)$ which intertwines $\Sigma'_A + \Sigma'_B$ and $\Sigma_A + \Sigma_B$. Thus the controllability of $\Sigma'_A + \Sigma'_B$ implies the controllability of $\Sigma_A + \Sigma_B$. If each of the systems Σ_A and Σ_B is exactly controllable, then X_A and X_B are boundedly invertible and so is $X_A \oplus X_B$. In this case $\Sigma'_A + \Sigma'_B$ is controllable or exactly controllable if and only if $\Sigma_A + \Sigma_B$ has these properties. Summarizing the above discussion in a theorem we get the following.

THEOREM 4.4. *Let A and B be strictly noncyclic in $H^\infty(B(N, M))$ having the prime factorizations* (3.1) *and* (3.2). *A sufficient condition for the controllability of $\Sigma_A + \Sigma_B$ is* (4.36): *If Σ_A and Σ_B are both exactly controllable, then condition* (4.36) *is also necessary.*

This leaves open the question of the necessity of (4.3) when the exact controllability assumption is violated. We note however that the finite-dimensional problem is completely solved as in that case controllability and exact controllability are equivalent.

REFERENCES

[1] P. R. AHERN AND D. N. CLARK, *On functions orthogonal to invariant subspaces*, Acta Math., 124 (1970), pp. 191–204.

[2] F. M. CALLIER AND C. D. NAHUM, *Necessary and sufficient conditions for the complete controllability and observability of systems in series using the coprime decomposition of a rational matrix*, IEEE. Trans. Circuits and Systems., to appear.

[3] L. CARLESON, *Interpolation of bounded analytic functions and the corona problem*, Ann. of Math., 76 (1962), pp. 547–559.

[4] P. A. FUHRMANN, *On the corona theorem and its applications to spectral problems in Hilbert space*, Trans. Amer. Math. Soc., 132 (1968), pp. 55–66.

[5] ———, *A functional calculus in Hilbert space based on operator valued analytic functions*, Israel J. Math., 6 (1968), pp. 267–278.

[6] ———, *On realization of linear systems and applications to some questions on stability*, Math. Systems Theor., 8 (1974), pp. 132–141.

[7] ———, *Exact controllability and observability and realization theory in Hilbert space*, J. Math. Anal. Appl., to appear.

[8] ———, *Realization theory in Hilbert space for a class of transfer functions*, J. Functional Anal., 18 (1975), pp. 338–349.

[9] ———, *Factorization theorems for a class of bounded measurable operator valued functions*, to appear.

[10] ———, *On Hankel operators induced by sums and products*, Israel J. Math., to appear.

[11] ———, *On controllability and observability of systems connected in parallel*, IEEE Trans. Circuits and Systems, 22 (1975), p. 57.

[12] ———, *On canonical realization of sums and products of non-rational transfer functions*, Proc. 8th Princeton Conference on Information Sciences and Systems, Princeton University, 1974, pp. 213–217.

[13] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.

[14] J. W. HELTON, *Discrete time systems, operator models and scattering theory*, J. Functional Anal., 16 (1974), pp. 15–38.

[15] R. E. KALMAN, *Lectures on Controllability and Observability*, C.I.M.E., Bologna, 1968.

[16] B. MOORE, III, *Canonical forms in linear systems*, Proc. 1973 Allerton Conference, University of Illinois, 1973, pp. 36–44.

[17] B. MOORE, III AND E. A. NORDGREN, *On quasi-equivalence and quasi-similarity*, Acta Sci. Math., 34 (1973), pp. 311–316.

[18] E. A. NORDGREN, *On quasi-equivalence of matrices over H^∞*, Ibid., 34 (1973), pp. 301–310.

[19] H. H. ROSENBROCK, *State Space and Multivariable Theory*, John Wiley, New York, 1970.

[20] B. SZ-NAGY AND C. FOIAS, *Sur les contractions de l'espace de Hilbert VIII, Fonctions caracteristiques. Modeles fonctionells*, Acta Sci. Math., 29 (1964), pp. 38–71.

[21] ———, *Operators sans multiplicité*, Ibid., 30 (1969), pp. 1–18.

[22] ———, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

SINGULAR REGIMES IN CERTAIN CLASSES OF RELAXED CONTROL PROBLEMS*

THOMAS E. CARTER†

Abstract. We present new necessary conditions for a relaxed minimum in optimal control problems defined by certain classes of ordinary differential equations. These necessary conditions may be helpful in computing singular extremal arcs, in determining when these arcs are "strictly relaxed", and in defining regions of the state space that contain nonsingular extremal arcs only.

1. Introduction. Let T be a closed interval $[t_0, t_1]$ in $\mathbb{R}$, m and n positive integers, R a convex compact set in $\mathbb{R}^m$ and V an open set in $\mathbb{R}^n$. We denote by $\mathcal{U}$ the collection of (Lebesgue) measurable functions $u: T \to R$ and by $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ the set of real $m \times n$ matrices. We assume given functions $g: T \times V \to \mathbb{R}^n$, $B: T \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, $\phi: T \times V \to \mathbb{R}$ with continuous second derivatives, a continuous $\psi: R \to \mathbb{R}$, and shall consider the optimal control problem of minimizing $y_0(t_1)$ subject to the condition that $u \in \mathcal{U}$, and y_0 and y are absolutely continuous solutions of the differential equations

$$\begin{aligned} \dot{y}_0(t) &= \phi(t, y(t)) + \psi(u(t)), \\ \dot{y}(t) &= g(t, y(t)) + B(t)u(t) \end{aligned} \tag{1.1}$$

a.e. in T and satisfy preassigned boundary restrictions.

As is well known, a problem of this kind (which we shall refer to as the *original problem*) does not always admit a minimizing solution, but under fairly general conditions, the corresponding relaxed problem does. We define this relaxed problem as does Warga in [3] and [5]. (His two definitions are equivalent for our problem). Specifically, we consider the space $RP(R)$ of regular Radon probability measures on R with the weak star topology of $C(R)^*$ (the dual of $C(R)$) and the set $\mathcal{S}$ of *relaxed control functions*, that is, of (Lebesgue) measurable functions $\sigma: T \to RP(R)$. The set $\mathcal{U}$ is embedded in $\mathcal{S}$ by identifying $u \in \mathcal{U}$ with the function $t \to \delta_{u(t)}$, where δ_r is the Dirac measure at r. For $t \in T$ and $\sigma \in \mathcal{S}$, we write $\psi(\sigma(t)) = \int \psi(r)\sigma(t)(dr)$ and $s(t) = \int r\sigma(t)(dr)$.

The *relaxed problem* is defined by replacing (1.1) by

$$\begin{aligned} \dot{y}_0(t) &= \phi(t, y(t)) + \psi(\sigma(t)), \\ \dot{y}(t) &= g(t, y(t)) + B(t)s(t) \end{aligned} \tag{1.2}$$

a.e. in T. Effectively, the inclusion of relaxed controls enlarges the set of admissible values of $(\dot{y}_0(t), \dot{y}(t))$ to the set

$$(\phi(t, y(t)), g(t, y(t))) + \overline{\text{co}}\,\{(\psi(r), B(t)r) | r \in R\},$$

where $\overline{\text{co}}$ denotes the closed convex hull. A triplet (y_0, y, σ) is a *minimizing relaxed*

* Received by the editors October 1, 1974, and in revised form March 9, 1975.

† Department of Mathematics, Nasson College, Springvale, Maine 04083. This paper is based on the author's Ph.D. dissertation, written under the supervision of Professor J. Warga and submitted to the Department of Mathematics, Northeastern University, Boston, Massachusetts.

solution if it yields the minimum of $y_0(t_1)$ subject to the specified boundary conditions.

It is our purpose in this paper to investigate the regime in which an extremal of (1.2) is singular, and this is the case, in particular, when it is strictly relaxed, that is, not locally equivalent to a solution of (1.1) which we refer to as original. For regimes of this kind, a necessary condition supplementing the Pontryagin maximum principle was derived by Warga in [4, Thm. 5.1, p. 138]. (See also [5, Thm. VI.2.5, pp. 365–366].) We shall study such regimes in greater detail for the more specialized problems here defined, and we shall derive additional necessary conditions. These conditions can be utilized to further specify the singular optimal controls and to determine regions of the y-space in which all the relaxed extremals are actually original.

As pointed out by the referee, some of the restrictions in our formulation are unnecessary. For example, it is sufficient to assume that ψ is lower semicontinuous and R compact. However, our somewhat stronger assumptions enable us to apply without modification Warga's results in [5] and thus eliminate the need for several technical lemmas.

Section 2 presents additional new necessary conditions and some examples, primarily in the case where

(i) $n = 2$,

but also with a much briefer discussion of the substantially simpler problems where

(ii) ψ is concave and R a polyhedron,

(iii) ψ is separable (i.e., $\psi(r) = \sum_{i=1}^{m} \psi_i(r^i)$ where $r = (r^1, \cdots, r^m) \in R$) and R is the Cartesian product of closed intervals. Section 3 contains the proofs. In § 4 we also briefly consider the relaxed control version of the simplest one-dimensional problem of the calculus of variations (to which (1.1) and (1.2) do not apply necessarily).

If k, l, p are positive integers, $A \subseteq \mathbb{R}^k$, and $h: A \to \mathbb{R}^l$, we define the derivatives $h'(a), h''(a), \cdots, h^{(p)}(a)$ in the sense of Fréchet but relative to A. Specifically, we say that a linear function $F: \mathbb{R}^k \to \mathbb{R}^l$ is the derivative of h at a (denoted by $h'(a)$) if a belongs to a convex subset of A with a nonempty interior and

$$\lim |x - a|^{-1} |h(x) - h(a) - F(x - a)| = 0 \quad \text{as } x \to a, \quad x \in A \sim \{a\}.$$

(See [5, II.3, p. 167ff. for details]). Since the functions that we encounter are not defined on open sets, this definition enables us to obtain more general results.

We shall denote (total) derivatives with respect to t by a dot, e.g., $\dot{z}(t)$ or $(z(t)^T B(t))^{\cdot}$, and partial derivatives with respect to a (one- or multidimensional) variable by displaying the latter as a subscript, e.g., ϕ_t or Ψ_{r^1}. In particular, the partial derivative with respect to the second argument of ϕ and g (which belongs to V) will be denoted by ϕ_v and g_v respectively. We denote the Borel–Lebesgue measure by μ and use the terms "measure" and "measurable" to mean "μ-measure" and "μ-measurable." The superscript T will denote a row vector or the transpose of a matrix, and components of a vector will be distinguished by superscripts, e.g., $r = (r^1, r^2)$, $\eta(t) = (\eta^1(t), \eta^2(t))$, etc.

We shall constantly use the concept of an extremal, which is closely related to the following necessary conditions for a minimizing solution of (1.2) that follow directly from [4, Thm. 6.1, pp. 142, 143].

LEMMA 1.1. *Let (y_0, y, σ) be a minimizing relaxed solution. Then there exist constants c and $z_0 \geqq 0$ and an absolutely continuous function $z: T \to \mathbb{R}^n$ such that $z_0 + |z(t)| > 0$ for each $t \in T$, and for almost all $t \in T$,*

$$\dot{z}(t)^T = -z_0\phi_v(t, y(t)) - z(t)^T g_v(t, y(t)), \tag{1.1.1}$$

$$z_0\psi(\sigma(t)) + z(t)^T B(t)s(t) = \min_{\rho \in R} (z_0\psi(\rho) + z(t)^T B(t)\rho), \tag{1.1.2}$$

$$z_0[\phi(t, y(t)) + \psi(\sigma(t))] + z(t)^T[g(t, y(t)) + B(t)s(t)] + \int_t^{t_1} [z_0\phi_t(\tau, y(\tau)) + z(\tau)^T(g_t(\tau, y(\tau)) + \dot{B}(\tau)s(\tau))]\, d\tau = c, \tag{1.1.3}$$

$$[z(t)^T\dot{B}(t) - (\phi_v(t, y(t)) + z(t)^T g_v(t, y(t))B(t))](r - s(t)) = 0 \tag{1.1.4}$$

for all $r \in J_0(t) \overset{\text{def}}{=} \{\rho \in R | z_0\psi(\rho) + z(t)^T B(t)\rho = \min_{\rho \in R} (z_0\psi(\rho) + z(t)^T B(t)\rho)\}$.

We shall refer to (y_0, y, σ, z_0, z) as a *relaxed extremal* if (1.2), (1.1.1) and (1.1.2) are satisfied a.e. on T, whether (y_0, y, σ) is a minimizing relaxed solution or not. It is shown in the proof of [4, Thm. 5.1, p. 138] that every relaxed extremal also satisfies relations (1.1.3) and (1.1.4) a.e. on T.

A relaxed extremal is *abnormal* if $z_0 = 0$. The study of abnormal extremals of equations (1.2) is considerably simpler than that of other extremals, and it resembles the study of the case where ψ is linear. For this reason we shall consider only *normal* relaxed extremals (i.e., with $z_0 \neq 0$), and in this case we may assume that $z_0 = 1$ (because the relations (1.1.1)–(1.1.4) are homogeneous in (z_0, z, c)). The term *extremal* will be used to mean a "normal relaxed extremal" for the remainder of this work. For a given extremal $(y_0, y, \sigma, 1, z)$ of the relaxed problem, we shall define c and $J_0(t)$ as in Lemma 1.1.

Let

$$\operatorname{gr}\psi = \{(\psi(r), r) \in \mathbb{R}^{m+1} | r \in R\},$$

$$\Psi(r) = \min\{r^0 | (r^0, r) \in \overline{\operatorname{co}}(\operatorname{gr}\psi)\}.$$

The function Ψ is the *convex envelope of* ψ.[1] It can be verified that a function Ψ is the convex envelope of ψ if and only if Ψ is a convex function, $\Psi(r) \leqq \psi(r)$ for each $r \in R$ and, if $\mathscr{X}$ is any other convex function having this property, then $\Psi(r) \geqq \mathscr{X}(r)$ for each $r \in R$. We shall denote by $\Psi^{(p)}$ the pth derivative of Ψ (relative to R), and set

$$R^{(p)} = \{r \in R | \Psi^{(p)} \text{ exists and is continuous in some neighborhood of } r \text{ relative to } R\}.$$

We define the *closed tread* associated with $w \in \mathbb{R}^n$ as the convex set

$$\bar{G}(w) = \{r \in R | \Psi(r) + w^T r = \min_{\rho \in R} (\Psi(\rho) + w^T\rho)\}.^2$$

[1] As shown by Kruskal [1], the function Ψ need not be continuous even though ψ is continuous. I thank the referee for drawing my attention to this fact.

[2] In the notation of "Convex Analysis" (Rockafellar [2]) $\bar{G}(w) = \partial\psi^*(-w)$ where $\psi^*(w) = \max_{r \in R}\{w \cdot r - \psi(r)\}$ and $\partial\psi^*$ is the subgradient mapping associated with ψ^*. (Here ψ is defined for all $r \in \mathbb{R}^m$ but takes the value $+\infty$ outside of R). Furthermore, in this notation the function Ψ is actually ψ^{**} and $\Psi^* = \psi^*$.

Any closed tread can be viewed as a projection of a closed face of $\operatorname{gr}\Psi$ on R. The term *open tread* or simply *tread* will be used to refer to the relative interior of a closed tread $\bar{G}(w)$ in its affine hull and will be denoted $G(w)$. The *dimension of* $\bar{G}(w)$ or $G(w)$ (written $\dim G(w)$) is the dimension of its affine hull, and $|G(w)|$ denotes the diameter of the set $G(w)$. We shall say that a closed or open tread is *nontrivial* if its diameter (or, equivalently, its dimension) is nonzero.

For a given extremal $(y_0, y, \sigma, 1, z)$, we define the set-valued mappings $\bar{J}$ and J by

$$\bar{J}(t) = \bar{G}(B(t)^T z(t)) = \overline{\operatorname{co}} J_0(t), \qquad J(t) = G(B(t)^T z(t)).$$

We refer the reader to [5, I.7, p. 146] for results pertaining to set-valued mappings that we use here. The function $\bar{J}$ maps T into the set $\mathscr{K}$ of nonempty closed subsets of R with the topology of the Hausdorff metric, and J maps T into the set $\mathscr{P}'$ of nonempty subsets of R with the topology of the Hausdorff semimetric [5, I.7, p. 146]. We say that $\bar{J}$ (respectively J) is measurable if the set $\bar{J}^{-1}(A)$ (respectively $J^{-1}(A)$) is measurable in T for each open A in $\mathscr{K}$ (respectively $\mathscr{P}'$). A function $\hat{r}: S \to R$ is a *selection* of $\bar{J}$ (respectively J) if $S \subseteq T$ and $\hat{r}(t) \in \bar{J}(t)$ (respectively $J(t)$) for each $t \in S$.

We shall refer to an extremal $(y_0, y, \sigma, 1, z)$ and the corresponding relaxed control function σ as *singular at* $t \in T$ if $|J(t)| > 0$, as *nonsingular at* t if it is not singular at t, as *original at* t if $(\dot{y}_0(t), \dot{y}(t))$ exists and

$$(\dot{y}_0(t), \dot{y}(t)) \in (\phi(t, y(t)), g(t, y(t))) + \{(\psi(r), B(t)r) | r \in R\},$$

and as *strictly relaxed at* t if $(\dot{y}_0(t), \dot{y}(t))$ exists and it is not original at t. Clearly σ is strictly relaxed a.e. in the set where it is not original. (Observe that this definition of the term "original" does not mean that $\sigma(t)$ is a Dirac measure but only that it can be replaced by one.) We refer to this extremal as *nonsingular* (respectively *original*) if it is nonsingular (respectively original) at t a.e. in T. For any $S \subseteq T$, we define the *singular* and *strictly relaxed regimes* on S of this extremal as, respectively, the sets

$$A(S) = \{t \in S | \, |J(t)| > 0\} = \{t \in S | \, \dim J(t) > 0\},$$

and

$$A'(S) = \{t \in S | s(t) \notin J_0(t)\}.$$

It is easy to verify, using relations (1.2) and (1.1.2), that for almost all $t \in T$, the extremal is singular (respectively strictly relaxed) at t if and only if $(\dot{y}_0(t), \dot{y}(t))$ exists and $t \in A(T)$ (respectively $t \in A'(T)$).

We now consider the optimal control problem that is defined by replacing ψ with Ψ in (1.2). We shall refer to it as the *related problem.* It is known that, because of the convexity of Ψ, for every relaxed extremal of the related problem there exists an equivalent original extremal. In fact, we shall prove in § 3 the following Lemma in which $s(t)$ is defined as in (1.2).

LEMMA 1.2. *The related problem admits* $(y_0, y, s, 1, z)$ *as a normal original extremal if the relaxed problem admits* $(y_0, y, \sigma, 1, z)$ *as a normal extremal.*

It is clear that our extremal of the relaxed problem and the corresponding original extremal of the related problem which is specified by Lemma 1.2 have identical singular regimes on T because both extremals define the same set-valued mapping $\bar{J}$. Furthermore if, for the relaxed problem, σ is strictly relaxed at a point $t \in T$ then, for the related problem, s is singular at t and

$$s(t) \in \bar{J}(t) \sim J_0(t) = \{r \in R | \psi(r) > \Psi(r)\} \cap \bar{J}(t) \subseteq J(t).$$

We shall henceforth assume that $(y_0, y, \sigma, 1, z)$ is a specific extremal for the relaxed problem and shall focus our attention on it.

2. Necessary conditions and examples. We now present new necessary conditions for a minimizing solution for several cases of the related problem. These necessary conditions provide additional information about a normal extremal in its singular regime on T and therefore apply also to the relaxed problem. The information about a corresponding extremal of the relaxed problem in its strictly relaxed regime on T is provided by the additional necessary condition for strict relaxation, namely,

$$s(t) \in \bar{J}(t) \sim J_0(t), \qquad t \in A'(T).$$

These results can also be used to define regions in the state space that can only contain nonsingular extremal arcs (respectively, original extremal arcs) for the relaxed problem. Such information will be presented in the form of corollaries to the theorems which present the necessary conditions.

In the work to follow, we shall denote the columns of the matrix $B(t)$ by $b_1(t), \cdots, b_m(t) \in \mathbb{R}^n$ and, for all $(t, v, w) \in T \times V \times \mathbb{R}^n$, we set

$$\gamma_1(t, v, w) = \phi_v(t, v) + w^T g_v(t, v),$$
$$\gamma_2(t, v, w) = \phi_{vv}(t, v) + w^T g_{vv}(t, v),$$
$$\gamma_3(t, v, w) = \gamma_{1_t}(t, v, w) - \gamma_2(t, v, w) g(t, v) - \gamma_{1_t}(t, v, w).$$

We might mention, for the sake of clarity, that a derivative such as $\phi_v(t, v)$ is represented by a row vector. Similarly, a second derivative such as $\phi_{vv}(t, v)$ or $w^T g_{vv}(t, v) = (w^T g(t, v))_{vv}$ is an operator such that $\phi_{vv}(t, v)x$ is a row vector for $x \in \mathbb{R}^n$. Computationally, if $\phi_{vv}(t, v)$ is represented by a square matrix M, then $\phi_{vv}(t, v)x$ is represented by $x^T M$. For an extremal $(y_0, y, s, 1, z)$ of the related problem corresponding to the extremal $(y_0, y, \sigma, 1, z)$ of the relaxed problem, we shall write

$$\tilde{\gamma}_i(t) = \gamma_i(t, y(t), z(t)), \qquad i = 1, 2, 3,$$
$$\eta(t) = (\eta^1(t), \cdots, \eta^m(t))^T = z(t)^T \dot{B}(t) - \tilde{\gamma}_1(t) B(t) = (z(t)^T B(t))^{\cdot}, \qquad t \in T_1,$$

and

$$\mathscr{D}_i = \{t \in T | J(t) \subseteq R^{(i)}\}, \qquad i = 1, 2, 3.$$

Arbitrary closed and open treads are denoted by $\bar{G}$ and G, respectively. We set

$$G_0 = \{r \in \bar{G} | \psi(r) = \Psi(r)\},$$

and

$$\mathscr{J}_i = \{\bar{G} \subseteq R \mid |G| > 0 \text{ and } G \not\subseteq R^{(i)}\}, \qquad i = 1, 2, 3.$$

Thus $J(t) \notin \mathscr{J}_i$ if $t \in \mathscr{D}_i$ and $|\bar{J}(t)| > 0$.

Our basic results (except for the one stated in § 4) are presented below, and are followed by illustrative examples.

THEOREM 2.1. *Assume that R is a convex polyhedron in $\mathbb{R}^m$, and that $-\psi$ is continuous and convex. Let $(y_0, y, s, 1, z)$ be a normal extremal of the related problem. Then, for almost all $t \in T$, the expressions $\Psi(r) + z(t)^T B(t)r$, $\eta(t)r$ and $\dot{\eta}(t)r$ are constant in r over the set $\bar{J}(t)$.*

COROLLARY. *Assume that R is a convex polyhedron in $\mathbb{R}^m$, and that $-\psi$ is continuous and convex. Let $S \subseteq T$ and $Y \subseteq V$ be such that for each choice of $(t, v) \in S \times Y$, a closed tread $\bar{G}$ in R, $r_1, r_2 \in \bar{G}$, $r \in G$ and $s_0 \in \bar{G} \sim G_0 \subseteq G$, the system of equations*

$$\Psi(r_2) - \Psi(r_1) + w^T B(t)(r_2 - r_1) = 0,$$

$$[w^T \dot{B}(t) - \gamma_1(t, v, w)B(t)](r_2 - r_1) = 0,$$

$$[w^T \ddot{B}(t) - 2\gamma_1(t, v, w)\dot{B}(t) + \gamma_3(t, v, w) - \gamma_2(t, v, w)B(t)s_0](r_2 - r_1) = 0$$

has no solution $w \in \mathbb{R}^n$. Then any normal extremal $(y_0, y, \sigma, 1, z)$ is original a.e. in $\{t \in S \mid y(t) \in Y\}$. If the above is valid also for each $s_0 \in \bar{G}$, then this extremal is also nonsingular a.e. in $\{t \in S \mid y(t) \in Y\}$.

Remark. Note that if $-\psi$ is strictly convex then $\bar{G} \sim G_0 = G$.

THEOREM 2.2. *Assume that $R = I^1 \times \cdots \times I^m$ where each I^i is a compact interval in $\mathbb{R}$, $\psi_i : I^i \to \mathbb{R}$ is continuous, and $\psi(r) = \sum_{i=1}^m \psi_i(r^i)$ for each $r = (r^1, \cdots, r^m) \in R$. Then all the closed treads in R are rectangles and the conclusion of Theorem 2.1 remains valid.*

COROLLARY. *Assume that $R = I^1 \times \cdots \times I^m$, and for each $i = 1, \cdots, m$, the set I^i is a compact interval in $\mathbb{R}$, $\psi_i : I^i \to \mathbb{R}$ is continuous, $\psi(r) = \sum_{i=1}^m \psi_i(r^i)$ for each $r = (r^1, \cdots, r^m) \in R$, and $C^{H^i} = (\Psi_i(r_2^i) - \Psi_i(r_1^i))/(r_2^i - r_1^i)$ for distinct $r_1^i, r_2^i \in \bar{H}^i$, where H^i is any component of $\{r^i \in I^i \mid \psi(r^i) > \Psi(r^i)\}$. Let $Y \subseteq V$, and $S \subseteq T$ be measurable. If for each $v \in Y$, $i = 1, 2, \cdots, m$, $s_0 \in I^1 \times \cdots \times H^i \times \cdots \times I^m$, every component H^i, and almost all $t \in S$, the system of equations*

$$w^T b_i(t) + C^{H^i} = 0,$$

$$\gamma_1(t, v, w)b_i(t) + w^T \dot{b}_i(t) = 0,$$

$$[\gamma_2(t, v, w)B(t)s_0][b_i(t)] - \gamma_3(t, v, w)b_i(t) + 2\gamma_1(t, v, w)\dot{b}_i(t) - w^T \ddot{b}_i(t) = 0$$

has no solution w, it follows that any extremal arc in Y is original a.e. on S.

THEOREM 2.3. *Let $V, R \subseteq \mathbb{R}^2$, and for each point $r = (r^1, r^2) \in R^{(3)}$ such that $\Psi_{r^1 r^1}(r) \neq 0$, $\Psi_{r^2 r^2}(r) \neq 0$, let*

$$K(r) = [\Psi_{r^2 r^2}(r)/\Psi_{r^1 r^1}(r)]_{r^2}/\Psi_{r^2 r^2}(r).$$

If $(y_0, y, s, 1, z)$ is an extremal of the related problem corresponding to the extremal, $(y_0, y, \sigma, 1, z)$ of the relaxed problem, then for almost all $t \in T$ the expressions $\Psi(r) + z(t)^T B(t)r$ and $\eta(t)r$ are constant in r over $\bar{J}(t)$. Furthermore, for almost all $t \in \mathscr{D}_1$, either the expression $\dot{\eta}(t)r$ is constant in r over $\bar{J}(t)$ or else

$$\Psi'(r) + z(t)^T B(t) = 0, \qquad r \in J(t); \tag{2.3.1}$$

and, if $t \in \mathscr{D}_3$,

$$\eta^1(t)\dot{\eta}^2(t) - \eta^2(t)\dot{\eta}^1(t) + \tfrac{1}{2}K(r)\eta^1(t)^3 = 0, \qquad r \in J(t). \tag{2.3.2}$$

If the sets $\mathscr{J}_1$ *(respectively* $\mathscr{J}_3$*) are finite or denumerable, then* $\dot{\eta}(t)r$ *is constant in* r *over* $\bar{J}(t)$ *for almost all* $t \notin \mathscr{D}_1$ *(respectively* $\mathscr{D}_3$*).*

If $\dim R \leqq 1$, *then* $\dot{\eta}(t)r$ *is constant in* r *over* $\bar{J}(t)$ *for almost all* $t \in T$ *without any assumptions about* $\mathscr{J}_1$.

Finally, the set $A(T)$ *(of all points* $t \in T$ *such that this extremal is singular at* t*), and the set* $A'(T)$ *(of all points* $t \in T$ *such that the corresponding extremal of the relaxed problem is strictly relaxed at* t*), are both measurable.*

Remark 1. It is well known (and easily verified) that relation (2.3.1) is valid for all $t \in \mathscr{D}_1$ for which $J(t)$ is in the interior of R. However, our alternative remains valid for all $t \in \mathscr{D}_1$ because we define $\Psi'(r)$ relative to R, and this derivative may therefore exist for some points r on the boundary of R. Furthermore, if $\mathscr{J}_1$ is finite or denumerable, then Theorem 2.3 asserts that $\dot{\eta}(t)r$ is constant in r over $\bar{J}(t)$ whenever Ψ is not continuously differentiable on $J(t)$.

Remark 2. Relation (2.3.2) of Theorem 2.3 may suggest the conjecture that $K(\cdot)$ is constant on every one-dimensional tread. The following counterexample shows that this conjecture is false.

Let $R = [0, \frac{1}{2}]^2 \subset \mathbb{R}^2$, and $\psi(r) = \Psi(r) = (r^1 - r^2)^2\, e^{(1/2)(r^1+r^2)^2}$. Along the closed tread that has the equation $r^1 = r^2$ we have $K(r) = 2r^1/e^{2(r^1)^2}$. This clearly is not constant.

COROLLARY. *Assume that, for the relaxed problem, the set* $\mathscr{J}_1$ *(respectively* $\mathscr{J}_3$*) is finite or denumerable. Let* $S \subseteq T$ *and* $Y \subseteq V$ *be such that for each choice of* $(t, v) \in S \times Y$, *a nontrivial closed tread* $\bar{G} \subseteq R, r_1, r_2 \in \bar{G}, r \in G$, *and* $s_0 \in \bar{G} \sim G_0 \subseteq G$, *neither the system*

$$\begin{aligned} &\Psi(r_2) - \Psi(r_1) + w^T B(t)(r_2 - r_1) = 0, \\ &[w^T \dot{B}(t) - \gamma_1(t, v, w)B(t)](r_2 - r_1) = 0, \\ &[w^T \ddot{B}(t) - 2\gamma_1(t, v, w)\dot{B}(t) + \gamma_3(t, v, w) \\ &\qquad - \gamma_2(t, v, w)B(t)s_0](r_2 - r_1) = 0 \end{aligned} \tag{2.3.3}$$

nor the system

$$\begin{aligned} &\Psi_{r^i}(r) + w^T b_i(t) = 0, \qquad i = 1, 2, \\ &[w^T \dot{B}(t) - \gamma_1(t, v, w)B(t)](r_2 - r_1) = 0 \end{aligned} \tag{2.3.4}$$

$$\begin{aligned} &(\textit{respectively}, \quad [w^T \dot{b}_1(t) - \gamma_1(t, v, w)b_1(t)] \cdot [w^T \ddot{b}_2(t) - 2\gamma_1(t, v, w)\dot{b}_2(t) \\ &+ (\gamma_3(t, v, w) - \gamma_2(t, v, w)B(t)s_0)b_2(t)] + [w^T \dot{b}_2(t) - \gamma_1(t, v, w)b_2(t)] \\ &\cdot [w^T \ddot{b}_1(t) - 2\gamma_1(t, v, w)\dot{b}_1(t) + (\gamma_3(t, v, w) - \gamma_2(t, v, w)B(t)s_0)b_1(t)] \\ &+ \tfrac{1}{2}K(r)[w^T \dot{b}_1(t) - \gamma_1(t, v, w)b_1(t)]^3 = 0) \end{aligned}$$

has a solution $w \in \mathbb{R}^2$. *Then any extremal* $(y_0, y, \sigma, 1, z)$ *is original a.e. in*

$\{t \in S | y(t) \in Y\}$. *If the above assumptions are also valid for each* $s_0 \in \bar{G}$, *then this extremal is both original and nonsingular a.e. in* $\{t \in S | y(t) \in Y\}$.

If $\dim R \leqq 1$, *then the assumption that* (2.3.3) *has no solution* $w \in \mathbb{R}^2$ *is sufficient without any assumptions about* $\mathscr{J}_1$, $\mathscr{J}_3$ or (2.3.4).

Examples.

Example 1. Let $R = [0, 1]^n, q^1, \cdots, q^n \in \mathbb{R}, q^i \neq -1, (i = 1, \cdots, n)$ and

$$\dot{y}_0(t) = \sum_{i=1}^{n} q^i(y^i(t))^2 - \sum_{i=1}^{m} (s^i(t))^2,$$

$$\dot{y}^1(t) = y^1(t)^2 + s^1(t),$$

$$\cdots \qquad\qquad \cdots$$

$$\dot{y}^n(t) = y^n(t)^2 + s^n(t) \qquad \text{a.e. in } T.$$

For each $i = 1, \cdots, m$, $H^i = (0, 1)$ and $C^{H^i} = -1$. The system of equations from the corollary of Theorem 2.2 is

$$w^i - 1 = 0,$$

$$2q^i v^i + 2v^i w^i = 0,$$

$$(2s_0^i - 2(v^i)^2)w^i + 2q^i(s_0^i - (v^i)^2) = 0, \qquad i = 1, \cdots, m.$$

Clearly there is no solution for this system (because $q^i \neq -1$ for each i), therefore the region Y of original extremal arcs is all of V. Note that if the set R is enlarged to contain the origin as an interior point, then a nonoriginal extremal trajectory (namely, $y(t) = 0$ for almost all $t \in T$) is possible.

Example 2. Let R be the compact circular disc centered at (1.1) with radius 1, $V = \mathbb{R}^2$, $\psi(r) = -f([(r^1 - 1)^2 + (r^2 - 1)^2]^{1/2})$ where $f:[0, 1] \to \mathbb{R}$, $f(0) = 0$, $f(1) = -1$, and f is strictly convex, $\phi(t, v) = (v^1)^2 + (v^2)^2$, $g(t, v) = 0$, and $B(t)$ is the identity matrix I on R^2.

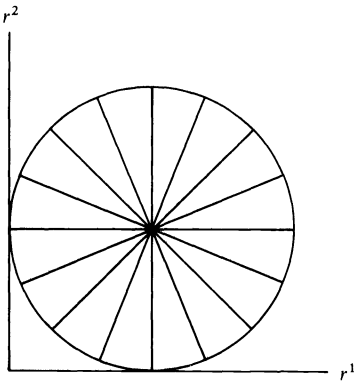


FIG. 1. *The region R in Example* 2 $(r^1 - 1)^2 + (r^2 - 1)^2 \leqq 1$

The convex envelope of ψ is

$$\Psi(r) = [(r^1 - 1)^2 + (r^2 - 1)^2]^{1/2};$$

therefore the sets $R^{(1)}$ and $R^{(3)}$ are the deleted disc $R \sim \{(1,1)\}$. The nontrivial closed treads are the closed rays from the center to the boundary and are clearly one-dimensional. (See Fig. 1.)

For this problem $\gamma_1(t,v,w) = 2v$, the matrix of $\gamma_2(t,v,w)$ is $2I$ and $\gamma_3(t,v,w) = 0$. For each $r \neq (1,1)$, $\Psi_{r^1}(r) = (r^1 - 1)/\Psi(r)$, $\Psi_{r^2}(r) = (r^2 - 1)/\Psi(r)$, and $K(r) = -2[\Psi(r)/(r^2 - 1)]^3$.

We shall use Theorem 2.3 to describe all possible singular control values at almost all $t \in T$ for a corresponding point $y(t) \in \mathbb{R}^2$ on an extremal trajectory of the related problem. We denote by S the set of points in T where a given extremal is singular. It follows from Theorem 2.3 that S is measurable, and for almost all $t \in S$, $r_1 = (1,1) \in \bar{J}(t)$, and $r_2 = r = (r^1, r^2) \in \bar{J}(t) \sim \{(1,1)\}$,

(i) $\Psi(r) + z^1(t)(r^1 - 1) + z^2(t)(r^2 - 1) = 0$,
(ii) $y^1(t)(r^1 - 1) + y^2(t)(r^2 - 1) = 0$ and either
(iii) $(r^1 - 1)s^1(t) + (r^2 - 1)s^2(t) = 0$, or else
(iv) $(r^1 - 1)/\Psi(r) + z^1(t) = 0, (r^2 - 1)/\Psi(r) + z^2(t) = 0$,
(v) $y^2(t)s^1(t) - y^1(t)s^2(t) + K(r)(y^1(t))^3 = 0$,

and, in addition to the above, we may set $r = s(t)$ where

(vi) $(s^1(t) - 1)^2 + (s^2(t) - 1)^2 \leqq 1$

because $s(t) \in \bar{J}(t)$.

Since $r \neq (1,1)$ the first alternative provides the simultaneous equations

$$y^1(t)s^1(t) + y^2(t)s^2(t) = y^1(t) + y^2(t),$$
$$y^2(t)s^1(t) - y^1(t)s^2(t) = 0,$$

which yield the solution

(vii) $$s^1(t) = \frac{(y^1(t)^2 + y^1(t)y^2(t))}{y^1(t)^2 + y^2(t)^2}, \qquad s^2(t) = \frac{y^1(t)y^2(t) + y^2(t)^2}{y^1(t)^2 + y^2(t)^2}.$$

We now consider the second alternative. It follows from (ii) that

$$K(r) = \pm 2[(y^1(t)^2 + y^2(t)^2)^{1/2}/y^1(t)]^3;$$

therefore we have the simultaneous equations

$$y^1(t)s^1(t) + y^2(t)s^2(t) = y^1(t) + y^2(t),$$
$$y^2(t)s^1(t) - y^1(t)s^2(t) = \mp 2[y^1(t)^2 + y^2(t)^2]^{3/2}$$

which yield the solution

(viii)
$$s^1(t) = \frac{y^1(t)^2 + y^1(t)y^2(t) \mp 2y^2(t)[y^1(t)^2 + y^2(t)^2]^{3/2}}{y^1(t)^2 + y^2(t)^2},$$
$$s^2(t) = \frac{y^1(t)y^2(t) + y^2(t)^2 \pm 2y^1(t)[y^1(t)^2 + y^2(t)^2]^{3/2}}{y^1(t)^2 + y^2(t)^2}.$$

Our singular extremal for this related problem has its control function defined a.e. on S by either (vii) or (viii). For the relaxed problem, a corresponding

extremal is strictly relaxed at $t \in S$ if $s(t) \in \bar{J}(t) \sim J_0(t)$; therefore the extremal controls represented by (vii) and (viii) are strictly relaxed if

(ix) $$0 < (s^1(t) - 1)^2 + (s^2(t) - 1)^2 < 1.$$

For this example the corollary of Theorem 2.3 can be used to define a set $Y \subseteq \mathbb{R}^2$ which can contain only nonsingular extremal arcs. It can be verified that for each nontrivial closed tread $\bar{G} \subset R$ and $s_0 \in \bar{G}$ (i.e., $(s_0^1 - 1)^2 + (s_0^2 - 1)^2 \leqq 1$), there are no solutions to the equations (2.3.3) if v is in the complement of the set

$$Y_1 = \left\{ v \in \mathbb{R}^2 \middle| \left[\frac{(v^1)^2 + v^1 v^2}{(v^1)^2 + (v^2)^2} - 1 \right]^2 + \left[\frac{v^1 v^2 + (v^2)^2}{(v^1)^2 + (v^2)^2} \right]^2 \leqq 1 \right\}$$

$$= \left\{ v \in \mathbb{R}^2 \middle| \frac{(v^2 - v^1)^2}{(v^1)^2 + (v^2)^2} \leqq 1 \right\} = \left\{ v \in \mathbb{R}^2 | v^1 v^2 \geqq 0 \right\},$$

and there are no solutions to the equations (2.3.4) if v is in the complement of the set

$$Y_2 = \left\{ v \in \mathbb{R}^2 \middle| \left[\frac{(v^1)^2 + v^1 v^2 \mp 2v^2[(v^1)^2 + (v^2)^2]^{3/2}}{(v^1)^2 + (v^2)^2} - 1 \right]^2 \right.$$

$$\left. + \left[\frac{v^1 v^2 + (v^2)^2 \pm 2v^1[(v^1)^2 + (v^2)^2]^{3/2}}{(v^1)^2 + (v^2)^2} - 1 \right]^2 \leqq 1 \right\}.$$

$$= \left\{ v \in \mathbb{R}^2 \middle| \frac{[v^1 - v^2 \mp 2[(v^1)^2 + (v^2)^2]^{3/2}]^2}{(v^1)^2 + (v^2)^2} \leqq 1 \right\}$$

$$= \left\{ v \in \mathbb{R}^2 \middle| \left| \frac{v^1 - v^2}{[(v^1)^2 + (v^2)^2]^{1/2}} \mp 2[(v^1)^2 + (v^2)^2] \right| \leqq 1 \right\}.$$

Therefore the set Y is the complement of the set $Y_1 \cup Y_2$.

The corollary of Theorem 2.3 can also be used to define a set $Y' \subseteq \mathbb{R}^2$ which can contain only original extremal arcs. This is the set of all points in $\mathbb{R}^2$ such that for each nontrivial closed tread $\bar{G} \subset R$ and $s_0 \in \bar{G} \sim G_0$ (i.e., $0 < (s_0^1 - 1)^2 + (s_0^2 - 1)^2 < 1$), there are no solutions to equations (2.3.3) or (2.3.4). Thus Y' is defined as the complement of the set $Y_1' \cup Y_2'$ where

$$Y_1' = \left\{ v \in \mathbb{R}^2 \middle| 0 < \frac{(v^2 - v^1)^2}{(v^1)^2 + (v^2)^2} < 1 \right\} = \{ v \in \mathbb{R}^2 | v^1 \neq v^2 \text{ and } v^1 v^2 > 0 \}$$

and

$$Y_2' = \left\{ v \in \mathbb{R}^2 \middle| 0 < \frac{[v^1 - v^2 \mp 2((v^1)^2 + (v^2)^2)^{3/2}]^2}{(v^1)^2 + (v^2)^2} < 1 \right\}$$

$$= \left\{ v \in \mathbb{R}^2 \middle| 0 < \left| \frac{v^1 - v^2}{((v^1)^2 + (v^2)^2)^{1/2}} \mp 2((v^1)^2 + (v^2)^2) \right| < 1 \right\}.$$

The sets Y and Y' are presented respectively in Figs. 2 and 3. It should be noted in Fig. 3 that the curve $\mathscr{F}$ whose equation is

$$\frac{[v^1 - v^2 \mp 2((v^1)^2 + (v^2)^2)^{3/2}]^2}{(v^1)^2 + (v^2)^2} = 0$$

is part of the set Y' which is obtained through the corollary of Theorem 2.3. This curve corresponds to those singular points of an extremal where the extremal control is supported at (1, 1). We shall verify that there is no loss in deleting this curve from the figure.

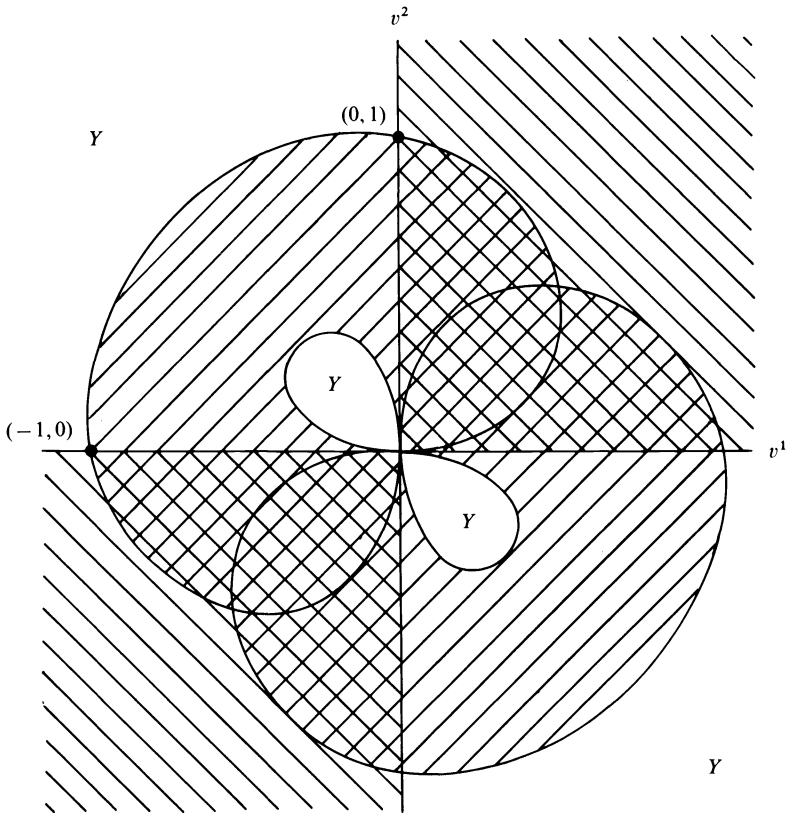


FIG. 2. *The region Y of only nonsingular arcs in Example* 2

For any extremal $(y_0, y, \sigma, 1, z)$, let $F = \{t \in A(T) | s(t) = (1, 1), y(t) \in \mathscr{F}\}$. By differentiating $y(\cdot)$ a.e. on F we see that either $y(t) = 0$ or $y(t) = (\pm\sqrt[4]{2}/2, \mp\sqrt[4]{2}/2)$ a.e. on F. Consequently $\mu(F) = 0$ because $s(t) = (1, 1)$ on F.

Example 3. Let R be the square $[-1, 1]^2$, $V = \mathbb{R}^2$,

$$R_L = \{r \in R | 2r^1 + 1 < r^2 < 1\},$$

$$R_R = \{r \in R | -2r^1 + 1 < r^2 < 1\},$$

$$R_M = \{r \in R | -1 < r^2 < \min(2r^1 + 1, -2r^1 + 1)\},$$

and

$$R_B = \{r \in R | r^2 = -1\}.$$

Let $\psi(r) = (r^1)^2 r^2$, $\phi(t, v) = (v^1)^2 + (v^2)^2$, $g(t, v) = v$, and $B(t)$ be the 2×2 identity matrix I.

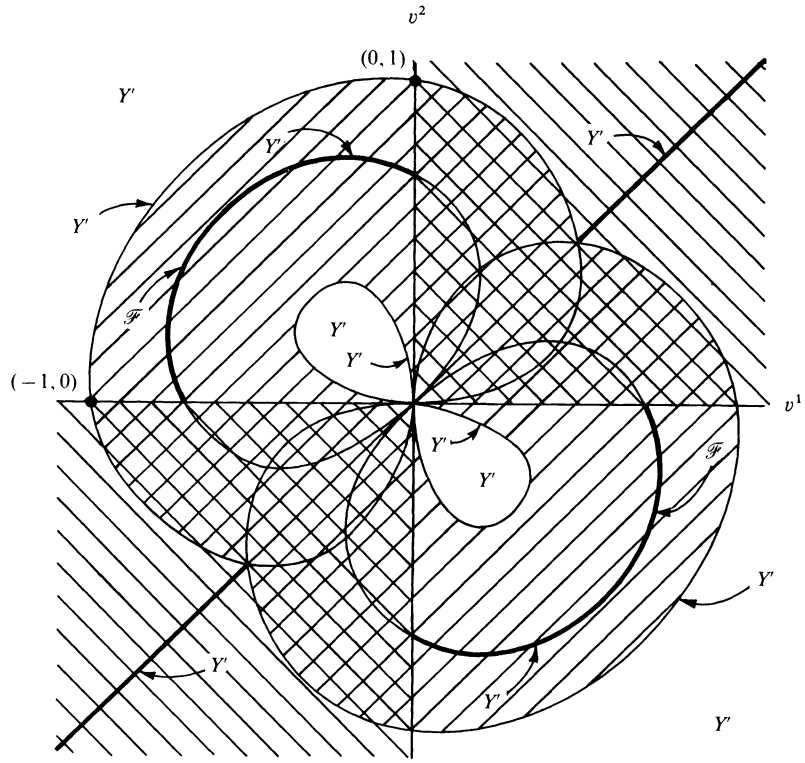


FIG. 3. *The region Y′ of only original arcs in Example* 2

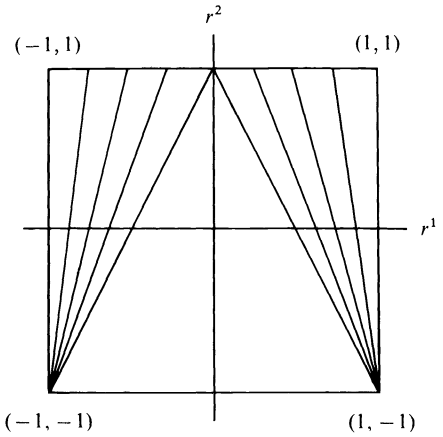


FIG. 4. *The region R in Example* 3

The convex envelope of ψ is defined by

$$\Psi(r) = \begin{cases} -1 + \frac{1}{2}(r^2 + 1)\left\{\left[2\left(\dfrac{r^1 + 1}{r^2 + 1}\right) - 1\right]^2 + 1\right\}, & r \in \bar{R}_L, \\ \frac{1}{2}(r^2 - 1), & r \in \bar{R}_M, \\ -1 + \frac{1}{2}(r^2 + 1)\left\{\left[2\left(\dfrac{r^1 - 1}{r^2 + 1}\right) + 1\right]^2 + 1\right\}, & r \in \bar{R}_R, \end{cases}$$

where 0/0 is defined as 1. It is easily verified that $R^{(1)} = R$ and $R^{(3)}$ is the set R with the exception of the two lines

$$r^2 = \pm 2r^1 + 1.$$

The nontrivial closed treads in R are the one-dimensional treads in $\bar{R}_L$ and $\bar{R}_R$ which emanate from the two lower corners of R, the lower edge R_B of R and the two-dimensional closed tread $\bar{R}_M$. Clearly $\mathscr{J}_3$ is finite consisting of the two one-dimensional closed treads in $R \sim R^{(3)}$. (See Fig. 4.)

For this problem $\gamma_1(t, v, w) = (2v + w)^T$, the matrix of $\gamma_2(t, v, w)$ is $2I$ and $\gamma_3(t, v, w) = w^T$. For each $r \in \bar{R}_L$, the number

$$\alpha = \frac{r^1 + 1}{r^2 + 1}$$

is the reciprocal of the slope of the tread containing r. For each $r \in \bar{R}_L$, $\Psi_{r^1}(r) = 4\alpha - 2$, $\Psi_{r^2}(r) = -2\alpha^2 + 1$ and $K(r) = -\frac{1}{2}$. Similarly, for each $r \in \bar{R}_R$, we define the number

$$\beta = \frac{r^1 - 1}{r^2 + 1},$$

and, in this region, $\Psi_{r^1}(r) = 4\beta + 2$, $\Psi_{r^2}(r) = -2\beta^2 + 1$ and again $K(r) = -\frac{1}{2}$.

We shall use Theorem 2.3 to describe all possible singular control values at almost all $t \in T$ for a corresponding point $y(t) \in \mathbb{R}^2$ on an extremal trajectory of the related problem. We denote by S_L the set of points in T where an extremal control of the related problem (also of the relaxed problem) is singular and supported on $\bar{R}_L$. It follows from Theorem 2.3 that S_L is measurable and, for almost all $t \in S_L$ and $r \in J(t) \subset \bar{R}_L$,

(i) $2\alpha^2 - 2\alpha + 1 + \alpha z^1(t) + z^2(t) = 0$,
(ii) $2\alpha y^1(t) + 2y^2(t) + \alpha z^1(t) + z^2(t) = 0$ and either
(iii) $-2\alpha s^1(t) - 2s^2(t) + \alpha z^1(t) + z^2(t) = 0$ or else
(iv) $4\alpha - 2 + z^1(t) = 0$, $-2\alpha^2 + 1 + z^2(t) = 0$,
(v) $(2y^1(t) + z^1(t))(z^2(t) - 2s^2(t)) - (2y^2(t) + z^2(t))(z^1(t) - 2s^1(t))$
$- \frac{1}{4}(2y^1(t) + z^1(t))^3 = 0$,

and in addition to the above,

(vi) $s^1(t) - \alpha s^2(t) - \alpha + 1 = 0$, $-1 \leqq s^2(t) \leqq 1$ because $s(t) \in \bar{J}(t)$, and
(vii) $0 \leqq \alpha \leqq \frac{1}{2}$ because $\bar{J}(t) \subset \bar{R}_L$.

From relations (i) and (ii), we have for almost all $t \in S_L$,

$$\alpha^2 - (1 + y^1(t))\alpha + (\tfrac{1}{2} - y^2(t)) = 0.$$

The roots of this equation are

(viii) $\alpha = \frac{1}{2}(y^1(t) + 1) \pm \frac{1}{2}[(y^1(t) + 1)^2 + 4y^2(t) - 2]^{1/2}$ a.e. on S_L. Solving the system (i), (ii), (iii) and (vi), and again solving the system (i), (ii), (iv), (v) and (vi) for $s(t)$ subject to the restrictions (vii) and (viii), we see that Theorem 2.3 implies that either

$$\text{(ix)} \qquad s^1(t) = \frac{-\alpha^3 + \alpha^2 + \frac{1}{2}\alpha - 1}{\alpha^2 + 1}, \qquad s^2(t) = \frac{-2\alpha^2 + 2\alpha - \frac{1}{2}}{\alpha^2 + 1}$$

or else

$$\text{(x)} \qquad \begin{aligned} s^1(t) &= \frac{-\frac{1}{2}(y^1(t) - 2\alpha + 1)^2\alpha - \alpha^3 + \alpha^2 + \frac{1}{2}\alpha - 1}{\alpha^2 + 1}, \\ s^2(t) &= \frac{-\frac{1}{2}(y^1(t) - 2\alpha + 1)^2 - 2\alpha^2 + 2\alpha - \frac{1}{2}}{\alpha^2 + 1} \end{aligned}$$

for almost all $t \in S_L$ where $-1 \leqq s^2(t) \leqq 1$ and $0 \leqq \alpha \leqq \frac{1}{2}$.

Similarly if S_R is the set of all points in T where an extremal control is singular and supported on $\bar{R}_R$, then for almost all $t \in S_R$ we have

$$\text{(xi)} \qquad \beta = \tfrac{1}{2}(y^1(t) - 1) \pm \tfrac{1}{2}[(y^1(t) - 1)^2 + 4y^2(t) - 2]^{1/2}$$

and either

$$\text{(xii)} \qquad s^1(t) = \frac{-\beta^3 - \beta^2 + \frac{1}{2}\beta + 1}{\beta^2 + 1}, \qquad s^2(t) = \frac{-2\beta^2 - 2\beta - \frac{1}{2}}{\beta^2 + 1}$$

or else

$$\text{(xiii)} \qquad \begin{aligned} s^1(t) &= \frac{-\beta^3 - \beta^2 + \frac{1}{2}\beta + 1 + \beta(y^1(t) - \beta - 1)^2}{\beta^2 + 1}, \\ s^2(t) &= \frac{-2\beta^2 - 2\beta - \frac{1}{2} + (y^1(t) - \beta - 1)^2}{\beta^2 + 1}, \end{aligned}$$

where $-1 \leqq s^2(t) \leqq 1$ and $-\frac{1}{2} \leqq \beta \leqq 0$.

Next we consider the case where S_B is the set of all points in T where an extremal control is singular and supported on R_B. The fact that $\Psi(r) + z(t)^T B(t)r$ and $\eta(t)r$ are constant in $r \in R_B$ yields

$$z^1(t) = 0, \qquad y^1(t) = 0 \quad \text{a.e. on } S_B,$$

and therefore

$$s^1(t) = 0 \quad \text{a.e. on } S_B.$$

The additional fact that $s(t) \in R_B$ for each $t \in S_B$ yields

(xiv) $s^1(t) = 0$, $s^2(t) = -1$ a.e. on S_B.

Finally, if S_M is the set of all points in $T \sim (S_L \cup S_R \cup S_B)$ where an extremal control is singular and supported on $\bar{R}_M$, then relations $\Psi(r_2) - \Psi(r_1) + z(t)^T B(t) \cdot (r_2 - r_1) = 0$, $\eta(t)(r_2 - r_1) = 0$ are valid for two linearly independent values of $r_2 - r_1$; consequently,

$$z^1(t) = 0, \quad z^2(t) = -1, \quad y^1(t) = 0, \quad y^2(t) = \tfrac{1}{2} \qquad \text{a.e. on } S_M,$$

and therefore

(xv) $$s^1(t) = 0, \qquad s^2(t) = -\tfrac{1}{2} \qquad \text{a.e. on } S_M.$$

The relations (viii)–(xv) provide a means of computing all possible values of a singular extremal control at a point $t \in T$ corresponding to a point $y(t)$ of the trajectory for the related problem. For the relaxed problem, an extremal control is strictly relaxed at t if $s(t) \in \bar{J}(t) \sim J_0(t)$. Therefore the extremal controls represented by (xiv) and (xv) are strictly relaxed, those represented by (ix) and (x) are strictly relaxed if

$$0 < \alpha \leqq \tfrac{1}{2} \quad \text{and} \quad -1 < s^2(t) < 1,$$

and those represented by (xii) and (xiii) are strictly relaxed if

$$-\tfrac{1}{2} \leqq \beta < 0 \quad \text{and} \quad -1 < s^2(t) < 1.$$

For this example the corollary of Theorem 2.3 can easily be applied, without computing $s(t)$, to define a set $Y \subset \mathbb{R}^2$ which can contain only nonsingular extremals. It is easily verified that the first two equations of (2.3.3), and the first three equations of (2.3.4) have no solutions for w if the relations

$$0 \leqq \alpha \leqq \tfrac{1}{2}, \quad -\tfrac{1}{2} \leqq \beta \leqq 0, \quad v^1 = 0$$

are not satisfied. These restrictions define the following set which can contain only nonsingular extremal arcs:

$$Y = \{v \in \mathbb{R}^2 | \tfrac{1}{2}(v^1 + 1) \pm \tfrac{1}{2}[(v^1 + 1)^2 + 4v^2 - 2]^{1/2} \notin [0, \tfrac{1}{2}], \tfrac{1}{2}(v^1 - 1) \pm \tfrac{1}{2}[(v^1 - 1)^2 + 4v^2 - 2]^{1/2} \notin [-\tfrac{1}{2}, 0], \text{ and } v^1 \neq 0\}.$$

To compute a set $Y' \subset \mathbb{R}^2$ which can contain only original extremals, we find the set of all points in $\mathbb{R}^2$ such that the first two equations of (2.3.3) and the first three equations of (2.3.4) have no solution if the relations

$$0 < \alpha \leqq \tfrac{1}{2}, \quad -\tfrac{1}{2} \leqq \beta < 0, \quad v^1 = 0$$

are not satisfied. This defines the following set which can contain only original extremal arcs:

$$Y' = \{v \in \mathbb{R}^2 | \tfrac{1}{2}(v^1 + 1) \pm \tfrac{1}{2}[(v^1 + 1)^2 + 4v^2 - 2]^{1/2} \notin (0, \tfrac{1}{2}], \tfrac{1}{2}(v^1 + 1) \pm \tfrac{1}{2}[(v^1 - 1)^2 + 4v^2 - 2]^{1/2} \notin [-\tfrac{1}{2}, 0), \text{ and } v^1 \neq 0\}.$$

The regions Y and Y' are presented in Fig. 5. The two regions are identical except for the line $v^2 = \frac{1}{2}$. This line is the set $Y' \sim Y$. It represents a possible singular extremal whose control is supported at the right or left edge of R.

3. Auxiliary lemmas and proofs of theorems. We begin with a proof of the Lemma 1.2 which was stated in the Introduction.

Proof of Lemma 1.2. Let $t \in T$ and relations (1.2) and (1.1.1)–(1.1.4) be satisfied at t. We define the function $h: R \to \mathbb{R}$ by

$$h(r) = \psi(r) + z(t)^T B(t) r,$$

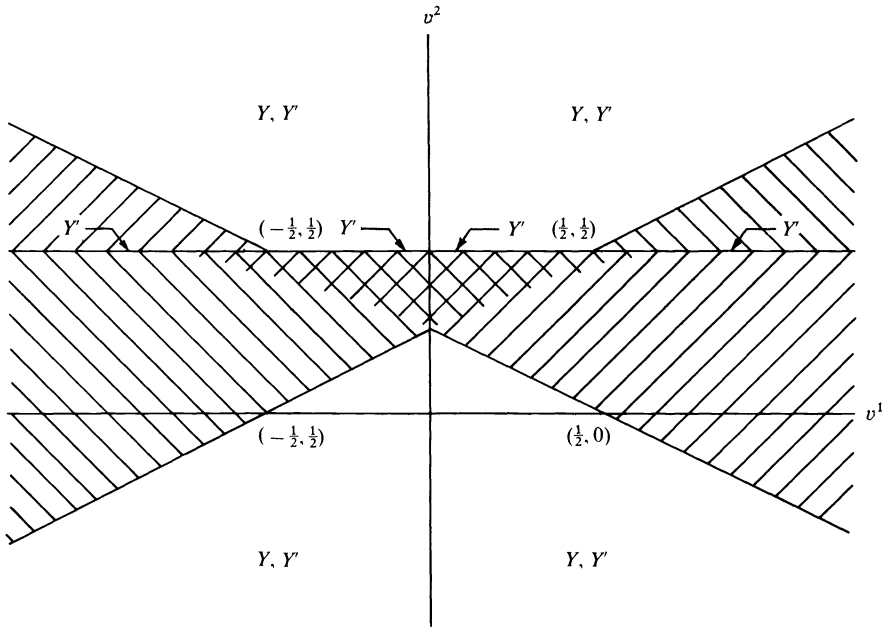


FIG. 5. *The region Y of only nonsingular arcs and the region Y' of only original arcs in Example* 3

and denote by H the convex envelope of h. As it is well known,

$$\int (h(r), r)\sigma(t)(dr) \in \overline{\text{co}}\,(\text{gr } h),$$

and if we let $\alpha(t) = \int h(r)\sigma(t)(dr)$, then $(\alpha(t), s(t)) \in \overline{\text{co}}\,(\text{gr } h)$ and therefore $\alpha(t) \geqq H(s(t))$. Since $\alpha(t) = \min_{r\in R} h(r)$ by (1.1.2), and $\min_{r\in R} h(r) = \min_{r\in R} H(r)$, we have $\alpha(t) \leqq H(s(t))$. It follows that $\alpha(t) = H(s(t))$; consequently,

$$\Psi(s(t)) + z(t)^T B(t)s(t) = \min_{r\in R} (\Psi(r) + z(t)^T B(t)r),$$

and this relation corresponds to (1.1.2) with $z_0 = 1$ for the relaxed problem. It is clear that (1.1.1) is also true with $z_0 = 1$ for the related problem. Furthermore, s is original at t for the related problem, and (1.2) is replaced by

$$\dot{y}_0(t) = \phi(t, y(t)) + \Psi(s(t)),$$
$$\dot{y}(t) = g(t, y(t)) + B(t)s(t).$$

Therefore $(y_0, y, s, 1, z)$ is a normal original extremal of the related problem. Q.E.D.

Let T_1 denote the set of the points of density 1 in the collection of all points in T where the relations (1.2), (2.1.1)–(2.1.4) are satisfied. It is known that $\mu(T_1) = \mu(T)$ and it is clear that the original extremal $(y_0, y, s, 1, z)$ of the related problem satisfies the following relations for each $t \in T_1$:

$$\dot{y}_0(t) = \phi(t, y(t)) + \Psi(s(t)),$$
$$\dot{y}(t) = g(t, y(t)) + B(t)s(t), \tag{1.2'}$$

$$\dot{z}(t)^T = -\phi_v(t, y(t)) - z(t)^T g_v(t, y(t)), \tag{1.1.1'}$$

$$\Psi(s(t)) + z(t)^T B(t)s(t) = \min_{\rho \in R} (z_0\psi(\rho) + z(t)^T B(t)\rho), \tag{1.1.2'}$$

$$\phi(t, y(t)) + \Psi(s(t)) + z(t)^T(g(t, y(t)) + B(t)s(t)) + \int_t^{t_1} [\phi_t(\tau, y(\tau)) + z(\tau)^T(g_t(\tau, y(\tau)) + \dot{B}(\tau)s(\tau))]\, d\tau = c, \tag{1.1.3'}$$

$$[z(t)^T\dot{B}(t) - (\phi_v(t, y(t)) + z(t)^T g_v(t, y(t))B(t))](r - s(t)) = 0, \qquad r \in \bar{J}(t). \tag{1.1.4'}$$

Henceforth we shall consider the restrictions of the maps $\bar{J}, J, s$, etc., to T_1 without changing the notation. For example, we write $\bar{J}: T_1 \to \mathscr{K}$ instead of $\bar{J}|T_1 : T_1 \to \mathscr{K}$. Whenever the domain of a map is not specifically stated, it should be understood that it is the set T_1.

For each $t \in T$, let

$$\alpha(t) = c - \phi(t, y(t)) - z(t)^T g(t, y(t)) - \int_t^{t_1} [\phi_t(\tau, y(\tau)) + z(\tau)^T(g_t(\tau, y(\tau)) + \dot{B}(\tau)s(\tau))]\, d\tau.$$

LEMMA 3.1. *For each $t \in T_1$, $r_1, r_2, r \in \bar{J}(t)$, and every selection $\hat{r}$ of $\bar{J}$, we have*

$$\Psi(r) + z(t)^T B(t)r = \alpha(t), \tag{3.1.1}$$

$$\Psi(r_2) - \Psi(r_1) + z(t)^T B(t)(r_2 - r_1) = 0, \tag{3.1.2}$$

$$(z(t)^T\dot{B}(t) - [\phi_v(t, y(t)) + z(t)^T g_v(t, y(t))]B(t))(r_2 - r_1) = 0, \tag{3.1.3}$$

$$\lim_{\substack{\tau \to t \\ \tau \in T_1}} [\Psi(\hat{r}(\tau)) + z(\tau)^T B(\tau)\hat{r}(\tau)] = \Psi(r) + z(t)^T B(t)r. \tag{3.1.4}$$

Proof. Relation (3.1.1) follows from (1.1.2'), (1.1.3'), and the definitions of $\bar{J}(t)$ and $\alpha(t)$. Relation (3.1.2) is obtained by subtracting (3.1.1) evaluated at two points r_1 and r_2 on $\bar{J}(t)$. Similarly relation (3.1.3) follows by subtracting (1.1.4') evaluated at two points r_1 and r_2 on $\bar{J}(t)$. Finally relation (3.1.4) is obtained from (3.1.1) and the fact that α is continuous. Q.E.D.

LEMMA 3.2. *The mapping $\bar{J}: T_1 \to \mathscr{K}$ is upper semicontinuous.*

Proof. We first show that the set-valued mapping $J_0 : T_1 \to \mathscr{P}'$ which is defined in (1.1.4) of Lemma 1.1 is upper semicontinuous.

Let $\{t_i\}$ be any sequence in T_1 that converges to $t \in T_1$, $\rho_i \in J_0(t_i)$ for each $i = 1, 2, \cdots$, and ρ_0 any limit point of the sequence $\{\rho_i\}$. There is a subsequence $\{t_{i_k}\} \subseteq \{t_i\}$ such that $\rho_0 = \lim_{k \to \infty} \rho_{i_k}$ and, by the continuity of ψ, z, and B,

$$\lim_{k \to \infty} [\psi(\rho_{i_k}) + z(t_{i_k})^T B(t_{i_k})\rho_{i_k}] = \psi(\rho_0) + z(t)^T B(t)\rho_0.$$

Since for any selection $\hat{r}$ of J_0 the mapping

$$t \mapsto \psi(\hat{r}(t)) + z(t)^T B(t)\hat{r}(t) = \min_{\rho \in R} (\psi(\rho) + z(t)^T B(t)\rho) = \alpha(t)$$

is continuous on T_1, it follows that $\rho_0 \in J_0(t)$. This shows that J_0 is upper semicontinuous.

We know that $\bar{J}(t) = \overline{\text{co}}J_0(t)$. It remains to show that for any $\varepsilon > 0$ there exists $\eta > 0$ such that $|\tau - t| < \eta$ implies $\bar{J}(\tau) \subset S(J(t), \varepsilon)$, where $S(A, \varepsilon) \overset{\text{def}}{=} \{r \in R \mid |r, A| < \varepsilon\}$ for $A \in \mathscr{P}'$ and $|r, A| \overset{\text{def}}{=} \inf_{a \in A} |r - a|$.

Given $\varepsilon > 0$ there exist $\varepsilon' > 0$ and $\eta > 0$ such that $J_0(\tau) \subset S(J_0(t), \varepsilon') \subset S(\bar{J}(t), \varepsilon/2)$ for $|\tau - t| < \eta$ because J_0 is upper semicontinuous and $J_0(t) \subseteq \bar{J}(t)$. Since $\bar{J}(t)$ is convex, $S(\bar{J}(t), \varepsilon/2)$ is also convex, therefore

$$\bar{J}(\tau) = \overline{\text{co}}J_0(\tau) \subset \overline{\text{co}}(J_0(t), \varepsilon') \subset \overline{S(\bar{J}(t), \varepsilon/2)} \subset S(\bar{J}(t), \varepsilon)$$

for $|t - \tau| < \eta$. Q.E.D.

It is well known [3, Lemma I.7.5, p. 150] that $\bar{J}$, as an upper semicontinuous mapping, must be μ-measurable. Since $\bar{J}(t)$ is the closure of $J(t)$ for all $t \in T_1$, it follows that J is also μ-measurable (since the Hausdorff pseudodistance of $\bar{J}(t)$ and $J(t)$ is zero).

LEMMA 3.3. *For each $t \in T_1$ and every selection $\hat{r}$ of $\bar{J}$, we have*

$$\lim_{\substack{\tau \to t \\ \tau \in T_1}} \left[\frac{\Psi(\hat{r}(\tau)) - \Psi(\hat{r}(t))}{\tau - t} + z(t)^T B(t) \frac{(\hat{r}(\tau) - \hat{r}(t))}{\tau - t} \right] = 0.$$

Proof. Because relations (1.2′) and (1.1.1′) are true for each $t \in T_1$, it follows that $\dot{\alpha}(t)$ exists for each $t \in T_1$. Therefore relation (3.1.1) of Lemma 3.1 implies that

$$\lim_{\substack{\tau \to t \\ \tau \in T_1}} \left[\frac{\Psi(\hat{r}(\tau)) - \Psi(\hat{r}(t))}{\tau - t} + \frac{z(\tau)^T B(\tau)\hat{r}(\tau) - z(t)^T B(t)\hat{r}(t)}{\tau - t} \right] = \dot{\alpha}(t).$$

Since R is bounded, this relation yields

$$\lim_{\substack{\tau \to t \\ \tau \in T_1}} \left[\frac{\Psi(\hat{r}(\tau)) - \Psi(\hat{r}(t))}{\tau - t} + z(t)^T B(t) \frac{(\hat{r}(\tau) - \hat{r}(t))}{\tau - t} + (\dot{z}(t)^T B(t) + z(t)^T \dot{B}(t))\hat{r}(\tau) - \dot{\alpha}(t) \right] = 0.$$

We differentiate the expression for $\alpha(t)$ using relations (1.1.1′) and (1.2′), and the above relation becomes

$$\lim_{\substack{\tau \to t \\ \tau \in T_1}} \left[\frac{\Psi(\hat{r}(\tau)) - \Psi(\hat{r}(t))}{\tau - t} + z(t)^T B(t) \frac{(\hat{r}(\tau) - \hat{r}(t))}{\tau - t} + [(\phi_v(t, y(t)) + z(t)^T g_v(t, y(t))B(t)) - z(t)^T \dot{B}(t)](s(t) - \hat{r}(\tau)) \right] = 0.$$

Our conclusion now follows from Lemma 3.2 and relation (1.1.4′). Q.E.D.

LEMMA 3.4. *Let $\{t_i\}$ be any sequence in T_1 converging to a point $t \in T_1$, and $\{\rho_i\}$ a sequence converging to $r \in \bar{J}(t) \cap R^{(1)}$ and such that $\rho_i \in \bar{J}(t_i)$ for each $i = 1, 2, \cdots$. Then*

$$\lim_{t_i \to t} \left([\Psi'(r) + z(t)^T B(t)] \left[\frac{\rho_i - r}{t_i - t} \right] \right) = 0.$$

Proof. This lemma is an immediate consequence of Lemma 3.3 and the assumption that $r \in R^{(1)}$. Q.E.D.

LEMMA 3.5. *Let S be a measurable subset of T_1. Then the sets $A(S)$ and $A'(S)$ are measurable.*

Proof. The functions s and J are measurable. Therefore it follows from Lusin's theorem that for every $\varepsilon > 0$ there exists a closed set $S_\varepsilon \subseteq S$ such that $\mu(S \sim S_\varepsilon) \leqq \varepsilon$ and both $J|S_\varepsilon$ and $s|S_\varepsilon$ are continuous. Let

$$D_\varepsilon^i = \{t \in S_\varepsilon | \dim J(t) \geqq i\}, \qquad i = 0, 1, \cdots.$$

If $t \in D_\varepsilon^i$, then $J(t)$ contains an i-simplex, and therefore, by the continuity of $J|S_\varepsilon$, $J(\tau)$ will also contain an i-simplex for all τ in some neighborhood of t in S_ε. Thus each D_ε^i is relatively open in S_ε, hence measurable. It follows that the sets

$$E_\varepsilon^i = \{t \in S_\varepsilon | \dim J(t) = i\} = D_\varepsilon^i \sim D_\varepsilon^{i+1}$$

are also measurable for $i = 0, 1, 2, \cdots, m$. Since $J|E_\varepsilon^i$ and $s|E_\varepsilon^i$ are continuous, it follows that for each i the set

$$G_\varepsilon^i = \{t \in E_\varepsilon^i | s(t) \in J(t)\}$$

is relatively open in E_ε^i and so is the set

$$G_\varepsilon = \{t \in S_\varepsilon | s(t) \in J(t)\} = \bigcup_{i=0}^{m} G_\varepsilon^i.$$

We now choose $\varepsilon = 1, \frac{1}{2}, \frac{1}{3}, \cdots$, and conclude that there exist sets Z and Z' of measure zero such that

$$A(S) = \left(\bigcup_{j=1}^{\infty} D_{1/j}^1 \right) \cup Z$$

and

$$A'(S) = \left(\bigcup_{j=1}^{\infty} G_{1/j} \right) \cup Z'.$$

Thus both $A(S)$ and $A'(S)$ are measurable. Q.E.D.

LEMMA 3.6. *Let $\mathscr{J} = \{\bar{G}_1, \bar{G}_2, \cdots\}$ be a finite or denumerable collection of nontrivial closed treads in R, and $E = \{t \in A(T_1) | \bar{J}(t) \in \mathscr{J}\}$. Then E is measurable and, for almost all $t \in E$ and each $r_1, r_2 \in \bar{J}(t)$, we have*

$$(z(t)^T B(t))^{\cdot}(r_2 - r_1) = \{z(t)^T \dot{B}(t) - [\phi_v(t, y(t)) + z(t)^T g_v(t, y(t))] B(t)\}(r_2 - r_1) = 0.$$

Proof. Let $E_i = \{t \in E | \bar{J}(t) = \bar{G}_i\}$ for each $i = 1, 2, \cdots$. Since the singleton $\{\bar{G}_i\}$ is a closed subset of $\mathscr{K}$, its inverse image E_i under the measurable mapping $\bar{J}$ is measurable. The set E is measurable because $E = \bigcup_{i=1}^{\infty} E_i$.

Now let j be an arbitrary positive integer and t an arbitrary limit point of E_j. Our conclusion follows by choosing two fixed points $r_1, r_2 \in \bar{G}_j$ and differentiating relation (3.1.3) at the point t. Q.E.D.

Proof of Theorem 2.1. The statement that the first two expressions are constant is a restatement of (3.1.2) and (3.1.3) of Lemma 3.1. We may, therefore, restrict our attention to the third expression.

Let $w \in \mathbb{R}^m$, $k = \min_{\rho \in R} (\psi(\rho) + w^T\rho)$, $G_0(w) = \{r \in R | \psi(r) + w^T r = k\}$, and $\mathscr{V}$ be the smallest set of vertices of R such that $G_0(w) \subseteq \overline{\text{co}}\, \mathscr{V}$. The function $r \to \psi(r) + w^T r : R \to \mathbb{R}$ is concave; therefore, it has a minimum at every point of $\mathscr{V}$. The convex envelope of this function takes the value k at each point on $\overline{\text{co}}\, \mathscr{V}$, and is strictly greater than k elsewhere. Therefore $\overline{\text{co}}\, \mathscr{V}$ is a closed tread. Since there are finitely many vertices in R, there are therefore finitely many closed treads in R. Thus the conclusion follows from Lemma 3.6. Q.E.D.

Proof of Theorem 2.2. Let $w = (w^1, \cdots, w^m) \in \mathbb{R}^m$, and for each $i = 1, \cdots, m$ define

$$\bar{G}^i(w) = \left\{ r^i \in I^i | \Psi_i(r^i) + w^i r^i = \min_{\rho^i \in I^i} (\Psi_i(\rho^i) + w^i \rho^i) \right\}$$

which is clearly a compact subinterval of I^i. Since $\psi(r) = \sum_{i=1}^m \psi_i(r^i)$, it follows also that $\Psi(r) = \sum_{i=1}^m \Psi_i(r^i)$, and consequently each closed tread $\bar{G}(w)$ in R is of the form $\bar{G}^1(w) \times \cdots \times \bar{G}^m(w)$ and therefore is rectangular.

The statement that the first two expressions are constant is a restatement of (3.1.2) and (3.1.3) of Lemma 3.1. It follows from (3.1.3) that $\eta(t)(r_2 - r_1) = 0$ for each $r_1, r_2 \in \bar{J}(t)$. Let $e_1, \cdots, e_m$ be standard unit vectors in $\mathbb{R}^m$ and $S_i \overset{\text{def}}{=} \{t \in T | \eta(t)e_i = 0\}$ for each $i = 1, \cdots, m$. It is easily verified that $\dot{\eta}(t)e_i = 0$ a.e. in S_i. Since all treads are rectangular, for each $r_1, r_2 \in \bar{J}(t)$, we may write $r_2 - r_1$ as a linear combination of standard unit vectors parallel to $\bar{J}(t)$. Consequently, $\dot{\eta}(t)(r_2 - r_1) = 0$ a.e. on the set where $\eta(t)(r_2 - r_1) = 0$. Q.E.D.

The lemmas which follow are leading to the proof of Theorem 2.3. They all refer to the extremal $(y_0, y, s, 1, z)$ of Theorem 2.3. When referring to points in $\mathbb{R}^2$, we shall consider the second coordinate as vertical and the first as horizontal. The slopes of nonvertical lines are defined accordingly.

LEMMA 3.7. *Assume that* $\dim R = 2$, *and let* $t \in A(T_1)$ *and* $\eta(t) = (z(t)^T B(t))^{\cdot} \neq 0$. *Then* $J(t)$ *is one-dimensional. If* $\eta^2(t) = (z(t)^T b_2(t))^{\cdot} \neq 0$, *then* $J(t)$ *is nonvertical and the slope* $m(t)$ *of* $J(t)$ *is given by*

$$m(t) = -\eta^1(t)/\eta^2(t) = -\frac{(z(t)^T b_1(t))^{\cdot}}{(z(t)^T b_2(t))^{\cdot}}.$$

Proof. If, for any $t \in A(T_1)$, $J(t)$ is two-dimensional, then there exist $r_1, r_2, r_3 \in \bar{J}(t)$ such that $r_2 - r_1$ and $r_3 - r_1$ are linearly independent, and relation (3.1.3) implies that $\eta(t)$ is orthogonal to both of these and is therefore zero, contradicting the assumption. Since $J(t)$ is nontrivial (because $t \in A(T_1)$), it must be one-dimensional.

If $\eta^2(t) \neq 0$, then $\eta(t) \neq 0$ and, as we have just seen, $J(t)$ is one-dimensional and has the direction of $r_2 - r_1$ for any two distinct points $r_1, r_2 \in \bar{J}(t)$. From relation (3.1.3) we have

$$\eta(t)(r_2 - r_1) = \eta^1(t)(r_2^1 - r_1^1) + \eta^2(t)(r_2^2 - r_1^2) = 0,$$

which shows that $J(t)$ is nonvertical and the slope of $J(t)$ is given by the formula in the statement of the lemma. Q.E.D.

LEMMA 3.8. *Assume that* $\dim R = 2$ *and that* S *is a subset of* T_1 *of positive measure such that* $\eta^1(t) \neq 0$, $\eta^2(t) \neq 0$ *and* $J(t) \subset R^{(1)}$ *for each* $t \in A(S)$. *Let* $Z_0 = \{t \in A(S) | \dot{\eta}(t)(r_2 - r_1) = 0$ *for each* $r_1, r_2 \in \bar{J}(t)\}$, $Z_i = (t \in A(S)|$ *relation* (2.3.i) *is valid for each* $r \in J(t)\}$ $(i = 1, 2)$. *Then* Z_0 *and* Z_1 *are measurable and, if* $J(t) \subset R^{(3)}$ *for each* $t \in A(S)$, Z_2 *is also measurable.*

Proof. We can replace the vector $r_2 - r_1$ by the proportional vector $(1, m(t))$. Since the slope $m(t)$ of $J(t)$, as evaluated in Lemma 3.7, is a continuous function of t on S, it follows that Z_0 is measurable.

By Castaing's theorem [5, Thm. I.7.8, p. 152], there exists a denumerable set $\{\hat{r}_1, \hat{r}_2, \cdots\}$ of measurable selections of $\bar{J}$ such that the set $\{\hat{r}_1(t), \hat{r}_2(t), \cdots\}$ is dense in $\bar{J}(t)$ for almost all $t \in T_1$. The set

$$P_i = \{t \in A(S) | \Psi'(\hat{r}_i(t)) = -B(t)^T z(t)\}$$

is measurable for each $i = 1, 2, \cdots$ because Ψ' is continuous on $R^{(1)}$ and $\hat{r}_i(t) \in R^{(1)}$ for each $t \in A(S)$ and $i = 1, 2, \cdots$. If $J(t) \subset R^{(3)}$ for each $t \in A(S)$, then the set

$$Q_i = \{t \in A(S) | \eta^1(t)\dot{\eta}^2(t) - \eta^2(t)\dot{\eta}^1(t) + \tfrac{1}{2}K(\hat{r}_i(t))\eta^1(t)^3 = 0\}$$

is measurable for each $i = 1, 2, \cdots$ because K is continuous on its open domain of definition. By continuity of Ψ' and K,

$$Z_1 = \bigcap_{i=1}^{\infty} P_i \quad \text{and} \quad Z_2 = \bigcap_{i=1}^{\infty} Q_i. \qquad \text{Q.E.D.}$$

LEMMA 3.9. *Let* $\dim R = 2$, *the sets* S, Z_0 *and* Z_1 *be as described in Lemma* 3.8, *and assume that*

$$\mu(A(S) \sim (Z_0 \cup Z_1)) > 0.$$

Then there exist a closed set $M \subset A(S) \sim (Z_0 \cup Z_1)$ *and an open rectangle* $P \subset \mathbb{R}^2$ *such that* $J(t) \cap P \neq \varnothing$ *for each* $t \in M$, $\mu(M) > 0$, *and either*

$$\Psi_{r^1}(r) + b_1(t)^T z(t) \neq 0, \qquad t \in M, \quad r \in J(t) \cap P,$$

or

$$\Psi_{r^2}(r) + b_2(t)^T z(t) \neq 0, \qquad t \in M, \quad r \in J(t) \cap P.$$

Proof. By Lusin's theorem, there exists a closed subset N of $A(S) \sim (Z_0 \cup Z_1)$ such that $\mu(N) > 0$ and $J|N$ is continuous. Let τ be a density point of N (i.e., a point of density 1). Since $\tau \notin Z_1$, there exists $\rho \in J(\tau)$ such that

$$\Psi'(\rho) + z(\tau)^T B(\tau) \neq 0.$$

It follows that we have

$$\Psi_{r^i}(\rho) + z(\tau)^T b_i(\tau) \neq 0$$

for some $i \in \{1, 2\}$. Since Ψ_{r^i} is continuous in some neighborhood of ρ relative to R (because $\rho \in R^{(1)}$) and both B and z are continuous, there exist an open rectangle $P \subset \mathbb{R}^2$ and a closed neighborhood M of τ in N such that

$$\Psi_{r^i}(r) + z(t)^T b_i(t) \neq 0, \qquad r \in P \cap R, \quad t \in M.$$

We have $\mu(M) > 0$ because τ is a density point of N and, since $J|N$ is continuous, we may choose M small enough so that $J(t) \cap P \neq \varnothing$ for each $t \in M$. Q.E.D.

LEMMA 3.10. *Assume that* $\dim R = 2$, *and let* S, Z_0 *and* Z_1 *be as described in Lemma* 3.8. *Then*

$$\mu(A(S)) = \mu(Z_0 \cup Z_1).$$

Proof. Assume that this statement is false. Then, by Lemma 3.9, there exist a closed set M of positive measure and a closed rectangle $Q \subset \mathbb{R}^2$ such that $\dim Q = 2$ and either

$$\text{I.} \qquad \Psi_{r^1}(r) + b_1(t)^T z(t) \neq 0, \qquad t \in M, \quad r \in J(t) \cap Q,$$

or

$$\text{II.} \qquad \Psi_{r^2}(r) + b_2(t)^T z(t) \neq 0, \qquad t \in M, \quad r \in J(t) \cap Q.$$

It follows also from our assumptions about S and from Lemma 3.7 that the tread $J(t)$ is one-dimensional and neither horizontal nor vertical for each $t \in M$. Let us assume, for the sake of definiteness, that II holds. Our subsequent arguments would be analogous in the other case.

Let V be any vertical line in $\mathbb{R}^2$ such that $V \cap Q \neq \varnothing$, and define

$$B(V) = \{t \in M | \bar{J}(t) \cap V \cap Q \neq \varnothing\}.$$

The set $B(V)$ is closed because, by Lemma 3.2, $\bar{J}$ is upper semicontinuous. Let $\mathscr{V} = \{V_1, V_2, \cdots\}$ denote a countable collection of vertical lines in $\mathbb{R}^2$ such that the set $\{V_1 \cap Q, V_2 \cap Q, \cdots\}$ is dense in Q. By Lemma 3.7, the tread $J(t)$ is, for each $t \in M$, one-dimensional and nonvertical; therefore $J(t) \cap V_j \cap Q \neq \varnothing$ for some $V_j \in \mathscr{V}$ and consequently $M \subseteq \bigcup_{i=1}^{\infty} B(V_i)$. Since $\mu(M) > 0$, there exists $V \in \mathscr{V}$ such that $\mu(B(V)) > 0$.

Let $t_0 \in B(V)$, $c = (c^1, c^2)$ denote the point of intersection of $\bar{J}(t_0)$ and V, $\delta > 0$,

$$C_R(\delta) = \{t \in B(V) | r^1 - c^1 \geqq \delta \text{ for some } r \in \bar{J}(t)\},$$

and

$$C_L(\delta) = \{t \in B(V) | c^1 - r^1 \geqq \delta \text{ for some } r \in \bar{J}(t)\}.$$

The sets $C_R(\delta)$ and $C_L(\delta)$ are closed for each $\delta > 0$ as a consequence of $B(V)$ being closed and $\bar{J}$ upper semicontinuous. By Lemma 3.7, the slope function $m(\cdot)$ is continuous and therefore bounded on the compact set M. Thus it is possible to pick $\delta > 0$ sufficiently small so that, for each $t \in B(V)$, there exists a point $r \in \bar{J}(t)$ such that either $r^1 - c^1 \geqq \delta$ or $c^1 - r^1 \geqq \delta$. Consequently $B(V) \subseteq C_L(\delta) \cup C_R(\delta)$, and therefore at least one of the sets $C_L(\delta)$ or $C_R(\delta)$ has positive measure.

Let M' be the set of all limit points of either the closed set $C_L(\delta)$ or $C_R(\delta)$ (whichever has positive measure), and let Q' be the corresponding closed rectangle

$$\{r \in Q | c^1 - \delta \leqq r^1 \leqq c^1\} \quad \text{or } \{r \in Q | c^1 \leqq r^1 \leqq c^1 + \delta\}.$$

It follows that each vertical line $V \subset \mathbb{R}^2$ that intersects Q' also intersects $\bar{J}(t)$ for each $t \in M'$. We next determine a closed rectangle $P' \subset Q'$ of width $\delta/2$ with neither of its vertical sides coinciding with a side of Q'. This ensures that, for each $t \in M'$, every vertical line $V \subset R^2$ that intersects P' also intersects the nonvertical open

tread $J(t)$ at a unique point $r_V(t)$. The function $r_V: M' \to V$ thus defined is continuous because J is upper semicontinuous.

It follows from Lemma 3.4 that

$$\lim_{\substack{\tau \to t \\ \tau \in M'}} \left[(\Psi'(r_V(t) + B(t)^T z(t))) \left(\frac{r_V(\tau) - r_V(t)}{\tau - t} \right) \right] = 0$$

for each $t \in M'$ and each vertical line V that intersects P'; hence, by II,

III. $$\lim_{\substack{\tau \to t \\ \tau \in M'}} \left[\frac{r_V(\tau) - r_V(t)}{\tau - t} \right] = 0$$

for each vertical line V that intersects P' and each $t \in M'$.

Now let V_1 and V_2 be two distinct vertical lines that intersect P'. It follows from relation (3.1.3) of Lemma 3.1 that

IV. $$\eta(t)(r_{V_2}(t) - r_{V_1}(t)) = 0, \qquad t \in M'.$$

Since ϕ_v and g_v are differentiable, and B, $\dot{B}$, y and z are differentiable on M', the function $\eta(\cdot)$ is also differentiable on M'. Therefore, we may write

V. $$[\eta(t) + (\tau - t)(\dot{\eta}(t) + \varepsilon(\tau))](r_{V_2}(\tau) - r_{V_1}(\tau)) = 0$$

for each τ, $t \in M'$, where $\lim_{\tau \to t, t \in M'} |\varepsilon(\tau)| = 0$. We subtract IV from V and divide by $\tau - t$ to obtain

$$\eta(t) \left[\frac{(r_{V_2}(\tau) - r_{V_2}(t))}{\tau - t} - \frac{(r_{V_1}(\tau) - r_{V_1}(t))}{\tau - t} \right] + (\dot{\eta}(t) + \varepsilon(\tau))(r_{V_2}(\tau) - r_{V_1}(\tau)) = 0$$

for distinct τ, $t \in M'$. We let τ approach t in M' and apply III for $V = V_1, V_2$. This yields

VI. $$\dot{\eta}(t)(r_{V_2}(t) - r_{V_1}(t)) = 0, \qquad t \in M'.$$

Now let $t \in M'$ and r_1, r_2 be distinct points of $\bar{J}(t)$. There is a nonzero real number $\beta(t)$ such that

$$r_{V_2}(t) - r_{V_1}(t) = \beta(t)(r_2 - r_1);$$

therefore VI implies that

$$\dot{\eta}(t)(r_2 - r_1) = 0.$$

This shows that $M' \subseteq Z_0$. This contradicts our original assumption and the conclusion follows. Q.E.D.

LEMMA 3.11. *Assume that* $\dim R = 2$. *If the slope m of any one-dimensional tread $G \subset R^{(2)}$ is finite, then, for each $r \in G$, either*

$$m^2 = \Psi_{r^1 r^1}(r)/\Psi_{r^2 r^2}(r) \quad \text{or} \quad \Psi''(r) = 0.$$

Proof. The function Ψ is convex on R and its restriction to a closed tread $\bar{G}$ is an affine function. We designate the endpoints of the one-dimensional closed tread $\bar{G}$ by r_0 and r_1. Then any point $r \in \bar{G}$ can be represented by $r = r_0 + \alpha(r_1 - r_0)$

where $0 \leqq \alpha \leqq 1$. Since $G \subset R^{(2)}$ and Ψ restricted to G is affine, it follows that

$$(r_1 - r_0)^T \Psi''(r)(r_1 - r_0) = \frac{d^2\Psi}{d\alpha^2}(r_0 + \alpha(r_1 - r_0)) = 0$$

for each $r \in G$. Since $r_1 - r_0$ is proportional to the vector $(1, m)$, the above relation can be written as

$$\text{I.} \qquad \Psi_{r^2r^2}(r)m^2 + 2\Psi_{r^1r^2}(r)m + \Psi_{r^1r^1}(r) = 0$$

for each $r \in G$. Since m is real, the discriminant of this quadratic equation is nonnegative, but the convexity of Ψ implies that $\Psi''(r)$ must be positive semidefinite, so this discriminant is also nonpositive; consequently,

$$\text{II.} \qquad \Psi_{r^1r^2}(r)^2 - \Psi_{r^1r^1}(r)\Psi_{r^2r^2}(r) = 0$$

for each $r \in G$. If $\Psi_{r^2r^2}(r) \neq 0$, then the root of the quadratic equation I is

$$m = -\Psi_{r^1r^2}(r)/\Psi_{r^2r^2}(r).$$

This relation and II can be combined to produce the first alternative. If $\Psi_{r^2r^2}(r) = 0$, then II yields $\Psi_{r^1r^2}(r) = 0$ which, together with I, implies that

$$\Psi_{r^1r^1}(r) = 0. \qquad \text{Q.E.D.}$$

LEMMA 3.12. *Assume that* $\dim R = 2$, *and let* S, Z_0, Z_1 *and* Z_2 *be as defined in Lemma* 3.8. *If* $J(t) \subset R^{(3)}$ *for each* $t \in A(S)$, *then*

$$\mu(A(S)) = \mu(Z_0 \cup (Z_1 \cap Z_2)).$$

Proof. We shall assume that the conclusion is false and argue by contradiction. Since $\mu(A(S)) = \mu(Z_0 \cup Z_1)$ by Lemma 3.8, our assumption implies that $\mu(Z) > 0$ for $Z = Z_1 \sim (Z_0 \cup Z_2)$. Because of Lusin's theorem there exists a closed set $M \subseteq Z$ such that $\mu(M) > 0$ and $J|M$ is continuous; furthermore, we may assume that each point in M is a density point of M. Let $t \in M$, $r \in J(t)$ and V be the vertical line through r. Since $J|M$ is continuous, there exists a relatively open interval I in M about t such that $J(\tau)$ intersects V at a unique point $r_V(\tau)$ for each $\tau \in I$. It follows from the definition of Z_1 that

$$\Psi_{r^2}(r_V(t)) + b_2(t)^T z(t) = 0.$$

This relation and the assumption (about $t \in S$) that

$$(z(t)^T b_2(t))^{\cdot} = \eta^2(t) \neq 0$$

imply that $\Psi_{r^2r^2}(r_V(t)) \neq 0$ (hence $\Psi''(r_V(t)) \neq 0$) and $|r_V(\tau) - r_V(t)| \neq 0$ for all τ in some relatively open interval $I' \subseteq I$ that contains t.

Thus it follows from Lemma 3.11 that

$$m(t)^2 = \Psi_{r^1r^1}(r_V(t))/\Psi_{r^2r^2}(r_V(t)),$$

where $m(t)$ represents the slope of $J(t)$. It follows from Lemma 3.7 that $m(t) \neq 0$

because $\dot{\eta}(t) \neq 0$; consequently, $\Psi_{r^1r^1}(r_V(t)) \neq 0$, and we have

$$\lim_{\substack{\tau \to t \\ \tau \in I'}} \left[\frac{\dfrac{\Psi_{r^2r^2}(r_V(\tau))}{\Psi_{r^1r^1}(r_V(\tau))} - \dfrac{\Psi_{r^2r^2}(r_V(t))}{\Psi_{r^1r^1}(r_V(t))}}{\Psi_{r^2}(r_V(\tau)) - \Psi_{r^2}(r_V(t))} \right] = -\lim_{\substack{\tau \to t \\ \tau \in I'}} \left[\frac{\dfrac{m(\tau)^{-2} - m(t)^{-2}}{\tau - t}}{\dfrac{z(\tau)^T b_2(\tau) - z(t)^T b_2(t)}{\tau - t}} \right]$$

$$= -\frac{(m(t)^{-2})^{\cdot}}{(z(t)^T b_2(t))^{\cdot}} = \frac{2(\eta^2(t)/\eta^1(t))^{\cdot}}{\eta^2(t)} = \frac{2(\eta^2(t)\dot{\eta}^1(t) - \eta^1(t)\dot{\eta}^2(t))}{\eta^1(t)^3}.$$

Since $r_V^2(\tau) - r_V^2(t) \neq 0$ for $\tau \in I'$, we may divide the numerator and the denominator of the expression on the left by this term before passing to the limit as $\tau \to t$, $\tau \in I'$. The expression on the left is therefore $K(r)$ because $r_V(t) = r \in J(t) \subset R^{(3)}$. Thus relation (2.3.2) holds for any point $t \in M$ and $r \in J(t)$, so $M \subseteq Z_2$. This provides the contradiction. Q.E.D.

Proof of Theorem 2.3. The assertions of the theorem are trivially true for all $t \in T_1 \sim A(T_1)$. We shall therefore restrict our attention to the set $A(T_1)$.

The fact that $\Psi(r) + z(t)^T B(t) r$ and $\eta(t) r$ are constant in r over $\bar{J}(t)$ follows from relations (3.1.2) and (3.1.3) of Lemma 3.1.

Next we assume that $\dim R = 2$, and first consider the case where $\mathscr{J}_1$ (respectively $\mathscr{J}_3$) is finite or denumerable. Let

$$E_i = \{t \in A(T_1) | \bar{J}(t) \in \mathscr{J}_i\} = A(T_1) \sim \mathscr{D}_i, \qquad i = 1, 3.$$

It follows from Lemma 3.6 that E_i is measurable and $\dot{\eta}(t)(r_2 - r_1) = 0$ for almost all $t \in E_i$ and each $r_1, r_2 \in \bar{J}(t)$ if $\mathscr{J}_i$ is finite or denumerable ($i = 1, 3$). Thus, for $i = 1, 3$, the set $\mathscr{D}_i$ is measurable and $\dot{\eta}(t) r$ is constant in r over $\bar{J}(t)$ for almost all $t \notin \mathscr{D}_i$ if $\mathscr{J}_i$ is finite or denumerable.

To show that the theorem is valid on $\mathscr{D}_1 \cap A(T_1)$, respectively, $\mathscr{D}_3 \cap A(T_1)$, whether or not $\mathscr{J}_1$ or $\mathscr{J}_3$ is finite or denumerable, we consider the following subsets of $\mathscr{D}_1 \cap A(T_1)$:

$$S_0 = \{t \in \mathscr{D}_1 \cap A(T_1) | \eta(t) = 0\},$$

$$S_j = \{t \in \mathscr{D}_1 \cap A(T_1) \sim S_0 | \eta^j(t) = 0\}, \qquad j = 1, 2,$$

$$S = \{t \in \mathscr{D}_1 \cap A(T_1) | \eta^1(t) \neq 0, \eta^2(t) \neq 0\}.$$

These sets are clearly measurable and form a partition of $\mathscr{D}_1 \cap A(T_1)$.

The theorem is true a.e. on S_0 because, for each limit point t of S_0, $\dot{\eta}(t) = 0$, which implies that $\dot{\eta}(t) r$ is constant in r over $\bar{J}(t)$.

The theorem is also true a.e. on S_1 because, for each limit point t of S_1, Lemma 3.7 implies that $r_2^2 - r_1^2 = 0$ for each $r_1, r_2 \in \bar{J}(t)$, and therefore

$$\dot{\eta}(t)(r_2 - r_1) = \dot{\eta}^1(t)(r_2^1 - r_1^1) = 0,$$

which shows that $\dot{\eta}(t) r$ is constant in r over $\bar{J}(t)$. A similar argument shows that the theorem is also true a.e. on S_2.

It follows from Lemma 3.10 that, for almost all $t \in S$, either $\dot{\eta}(t)(r_2 - r_1) = 0$ for each $r_1, r_2 \in \bar{J}(t)$ or else relation (2.3.1) is valid for each $r \in J(t)$. Furthermore, Lemma 3.12 implies that, for almost all $t \in S \cap \mathscr{D}_3 \cap A(T_1)$, either $\dot{\eta}(t) r$ is constant in r over $\bar{J}(t)$ or else both relations (2.3.1) and (2.3.2) are valid for each $r \in J(t)$.

Now we assume that dim $R \leqq 1$. In that case there is at most a countable set of nontrivial treads in R. Then Lemma 3.6 implies that $\dot{\eta}(t)r$ is constant in r over $\bar{J}(t)$ for almost all $t \in T$.

Finally, the last statement of the theorem follows from Lemma 3.5. Q.E.D.

4. The simplest problem of the calculus of variations. The relaxed control version of the simplest type of problem of the calculus of variations can be defined as follows: Let $T = [t_0, t_1]$ and $R = [\alpha, \beta]$ be closed intervals in $\mathbb{R}$, V an open subset of $\mathbb{R}$, $f: T \times V \times R \to \mathbb{R}$, and $\mathscr{S}$ the set of relaxed control functions which map T into $RP(R)$. For $(t, v, \sigma) \in T \times V \times \mathscr{S}$, we write $f(t, v, \sigma(t)) = \int f(t, v, r)\sigma(t)(dr)$ and $s(t) = \int r\sigma(t)(dr)$. We shall assume that $f(t, v, \cdot)$ is continuous for each $(t, v) \in T \times V$ and that $f(\cdot, \cdot, r)$ is continuously differentiable for each $r \in R$. The problem is to minimize $y_0(t_1)$ subject to the condition that $\sigma \in \mathscr{S}$, and y_0 and y are absolutely continuous solutions of the differential equations

$$\begin{aligned} \dot{y}_0(t) &= f(t, y(t), \sigma(t)), \\ \dot{y}(t) &= s(t) \end{aligned} \tag{4.1}$$

a.e. in T and satisfy preassigned boundary restrictions.

It follows from [4, Thm. 6.1, pp. 142, 143] that a relaxed normal extremal $(y_0, y, \sigma, 1, z)$ of this problem must satisfy the following conditions.

There exists a constant C such that, a.e. in T,

$$\dot{z}(t) = -f_v(t, y(t), \sigma(t)), \tag{4.2}$$

$$f(t, y(t), \sigma(t)) + z(t)s(t) = l(t) \stackrel{\text{def}}{=} \min_{\rho \in R} (f(t, y(t), \rho) + z(t)\rho), \tag{4.3}$$

$$f(t, y(t), \sigma(t)) + z(t)s(t) + \int_t^{t_1} f_t(\tau, y(\tau), \sigma(\tau))\, d\tau = C, \tag{4.4}$$

$$f_t(t, y(t), r) - f_t(t, y(t), \sigma(t)) + f_v(t, y(t), r)s(t) - f_v(t, y(t), \sigma(t))r = 0 \tag{4.5}$$

for each $r \in J_0(t) \stackrel{\text{def}}{=} \{r \in R \mid f(t, y(t), r) + z(t)r = l(t)\}$.

We say that the above extremal is *singular at a point* $t \in T$ if $J_0(t)$ contains more than a single point, *strictly relaxed* at t if $s(t) \in \overline{\text{co}}\, J_0(t) \sim J_0(t)$, *nonsingular* at t if it is not singular at t and *original at* t if it is not strictly relaxed at t. We define the *singular regime* and *strictly relaxed regime* on a set $S \subseteq T$ as in § 1 and refer to these as $A(S)$ and $A'(S)$, respectively.

THEOREM 4.1. *Let* $(y_0, y, \sigma, 1, z)$ *be a normal extremal of the relaxed problem defined by* (4.1). *Then for almost all* $t \in A(T)$ *and for distinct* $r_1, r_2 \in J_0(t)$, *we have*

$$f_v(t, y(t), \sigma(t)) = \left[\frac{f_v(t, y(t), r_2) - f_v(t, y(t), r_1)}{r_2 - r_1}\right] s(t) + \frac{f_t(t, y(t), r_2) - f_t(t, y(t), r_1)}{r_2 - r_1}.$$

Proof. We subtract the relation (4.5) evaluated at the distinct points r_1 and r_2, then divide by $r_2 - r_1$ and solve for $f_v(t, y(t), \sigma(t))$. Q.E.D.

COROLLARY. *If* $f(t, v, r)$ *is independent of its first argument,* (i.e., $f(t, v, r) = f(v, r)$), *then a normal relaxed extremal* $(y_0, y, \sigma, 1, z)$ *is singular at* t *only if the*

graph of the function $f_v(y(t), \cdot)$ is supported at two or more points by a line through the origin.

Examples.

Example 1. Let $f(t, v, r) = \phi(v) + \psi(r)$. This example is also a special case of the relaxed problem defined by (1.2). It follows from either Theorem 2.2, from relation (3.1.3) of Lemma 3.1, or from Theorem 4.1 that

$$\phi'(y(t)) = 0 \quad \text{a.e. in } A(T).$$

If the roots of the equation

$$\text{I.} \qquad \phi'(v) = 0$$

are isolated, then the absolute continuity of $y(\cdot)$ on T implies that

$$s(t) = 0 \quad \text{a.e. on } A(T).$$

Conversely, if a region Y does not contain the roots of the equation I or if the roots are isolated and $0 \notin R$, then any extremal trajectory in Y is nonsingular.

Example 2. Let $f(t, v, r) = a(v)r^2 + b(v)r + c(v)$. We note first that, for $t \in A(T)$, the function $r \to f(t, y(t), r) + z(t)r$ has at least two minima on R; hence

$$a(y(t)) \leqq 0.$$

We can also see that $J_0(t)$ is either $\{\alpha, \beta\}$ or $[\alpha, \beta]$, and specifically that $J_0(t) = \{\alpha, \beta\}$ if $t \in A'(T)$. It follows, therefore, from Theorem 4.1, setting $r_1 = \alpha$, $r_2 = \beta$, that

$$a'(y(t)) \int r^2 \sigma(t)(dr) + b'(y(t))s(t) + c'(y(t)) = [(\alpha + \beta)a'(y(t)) + b'(y(t))]s(t)$$

a.e. on $A(T)$. Furthermore, for $t \in A'(T)$, $\sigma(t)$ is of the form

$$\sigma(t) = \theta(t)\,\delta_\beta + (1 - \theta(t))\,\delta_\alpha,$$

where $0 < \theta(t) < 1$ and δ_r is the Dirac measure at r. Therefore, a.e. in $A'(T)$, we have

$$a'(y(t))[\alpha^2 + \theta(t)(\beta^2 - \alpha^2)] + c'(y(t)) = a'(y(t))(\alpha + \beta)[\alpha + \theta(t)(\beta - \alpha)],$$

where

$$0 < \theta(t) < 1.$$

Consequently,

$$c'(y(t)) - \alpha\beta a'(y(t)) = 0 \quad \text{a.e. on } A'(T).$$

We also have

$$a(y)(t)) = 0 \quad \text{a.e. on } A(T) \sim A'(T).$$

We may therefore draw the following conclusions.

If all the roots of the equation

$$\text{I.} \qquad c'(v) - \alpha\beta a'(v) = 0$$

that are in the set $\{v \in V | a(v) < 0\}$ are isolated, then the absolute continuity of $y(\cdot)$ on T implies that

$$s(t) = 0 \quad \text{a.e. on } A'(T).$$

If, moreover, all the roots of the equation

II. $$a(v) = 0$$

are isolated, then also

$$s(t) = 0 \quad \text{a.e. on } A(T).$$

Conversely, if a region Y does not intersect the set $\{v \in V | a(v) \leqq 0\}$ or if it does not contain any roots of I or II, or if these roots are isolated and $0 \notin R$, then any extremal trajectory in Y is nonsingular. Furthermore, if a region Y' does not intersect the set $\{v \in V | a(v) < 0\}$ or if it contains no root of I or if these roots are isolated and $0 \notin R$, then any extremal trajectory in Y' is original.

Acknowledgment. The author is grateful to Prof. Jack Warga for suggesting this area of research, and for providing much guidance and insight in all phases of the work. He also thanks the referee for constructive criticism and helpful advice in improving the exposition.

REFERENCES

[1] J. B. KRUSKAL, *Two convex counterexamples: A discontinuous envelope function and a nondifferentiable nearest-point mapping*, Proc. Amer. Math. Soc., 23 (1969), pp. 697–703.

[2] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[3] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–129.

[4] ———, *Necessary conditions for minimum in relaxed variational problems*, Ibid., 4 (1962), pp. 111–120.

[5] ———, *Optimal Control of Differential and Functional Equations*, Academic Press, New York and London, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

ERRATUM: CONTROLABILITE DES SYSTEMES NON LINEAIRES*

C. LOBRY†

P. Stefan pointed out in [1] that Lemma 1.2.1 in the paper [2] of the author was false. In [2], Lemma 1.2.1 was attributed to R. Hermann, but actually it was a modification of Hermann's result [3]. We refer the reader to [1] for a counter-example and other comments.

However, Proposition 1.2.1 of [2], which is important for applications to control theory, is true. A proof can be found in a paper of Nagano [4] which was not known to the author at that time.

The results contained in [2] have been considerably improved, and we refer the reader interested in those topics to [1], [5], [6], [7], [8].

REFERENCES

[1] P. STEFAN, Ph.D. thesis, Warwick Univ., 1973.
[2] C. LOBRY, *Contrôlabilité des systèmes non linéaires*, this Journal, 8 (1970), pp. 573–605.
[3] R. HERMANN, *The differential geometry of foliations. II*, Math. Mech., 11 (1962), pp. 305–315.
[4] T. NAGANO, *Linear differential systems with singularities and application to transitive lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.
[5] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
[6] H. SUSSMANN, *Orbits of families of vector fields and integrability of systems with singularities*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–188.
[7] A. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems*, this Journal, 12 (1974), pp. 43–52.
[8] D. Q. MAYNE AND R. W. BROCKETT, eds., *Geometric Method in System Theory*, D. Reidel, Dordrecht (Holland) and Boston.

* This Journal, 8(1970), pp. 573–605. Received by the editors December 18, 1974.

† Université de Bordeaux I, U.E.R. de Mathématiques et d'Informatique, 33405 Talence, France.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

MODULE STRUCTURE OF INFINITE-DIMENSIONAL SYSTEMS WITH APPLICATIONS TO CONTROLLABILITY*

EDWARD W. KAMEN†

Abstract. A theory of infinite-dimensional time-invariant continuous-time systems is developed in terms of modules defined over a convolution ring of generalized functions. In particular, input/output operators are formulated as module homomorphisms between free modules over the convolution ring, and systems are defined in terms of a state module. Results are presented on causality and the problem of realization. The module framework is then utilized to study the reachability and controllability of states and outputs. New results are obtained on the smoothness of controls, bounded-time controls, and minimal-time controls.

1. Introduction. The existing theory of infinite-dimensional systems is based primarily on the elements of topology and analysis (e.g., Banach spaces, Hilbert spaces, etc.). In contrast, in this paper the emphasis is on the application of modern algebra to the study of infinite-dimensional time-invariant systems. The objective here is to formulate a theory in terms of rings and modules which yield new results as a consequence of finiteness properties enjoyed by these algebraic structures.

Here the rings and modules are convolution structures that come into play as a result of the additional assumption of time invariance. In particular, as discussed in § 2, linear time-invariant input/output (i/o) operators can be formulated as module homomorphisms between finitely-generated modules defined over a convolution ring of functions. Although the convolution structure of these i/o operators is well known, very little attention has been devoted to the relationship between the i/o module framework and the internal system structure defined in terms of the concept of state.

The first major work on the role of the convolution structure in a state space setting was Kalman's $K[z]$-module description of finite-dimensional discrete-time systems [1]. Kalman was also the first one to consider a module structure over a convolution ring of functions in the state space theory of continuous-time systems (see Kalman and Hautus [2]). However, the theory of [2], which centers on the problem of realization, does not apply to a very large class of infinite-dimensional systems since it is assumed that for any positive integer n, the output response resulting from the nth derivative of the Dirac distribution at $\{0\}$ is infinitely differentiable on $(0, \infty)$. For example, this constraint prevents consideration of systems having time delays. The extension of Kalman's module framework to a suitably large class of infinite-dimensional systems is carried out here.

The convolution structure of the i/o description can be reflected in the internal system structure in two ways, depending on the type of internal model

* Received by the editors August 17, 1972, and in revised form November 18, 1974.

† School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332. This research was supported by the U.S. Army Research Office, Durham, under Grant DA-ARO-D-31-124-73-G171.

used. On the one hand, the dynamical equations can be given by operational-differential equations defined by convolution operators belonging to a Noetherian ring. This approach is developed in [3] and will not be considered here. In the second approach, which will be pursued here, systems are given in operational form by module homomorphisms with the state space also possessing a (topological) module structure over the convolution ring of functions.

The module structure on the state space provides a new approach to the study of dynamical properties. For instance, as revealed in §§ 5, 6 and 7, the concept of the annihilator of a module plays a central role in controllability. Using this concept, in § 6 we obtain the surprising result that if the reachable states of a system are controllable (to the zero state), then every reachable state can be controlled within some fixed time period (bounded-time controllability). Examples of systems in which the reachable states are always controllable are given in § 7.

In the following development, it is crucial that we work with a convolution ring of functions which contains the identity δ_0 = Dirac distribution at $\{0\}$. In other words, we need to consider a convolution ring of distributions (generalized functions). Then since we want the input function space and the output function space to be modules over the convolution ring, these spaces must also be spaces of distributions. The requirement that the ring contain δ_0 is mainly for algebraic reasons. For example, it is then possible to consider the operation of inversion which, as we shall see, leads to the construction of control signals.

2. Input/output operators. Let $\mathbb{R}$ denote the field of real numbers with the usual topology. Let $\mathscr{D}$ (resp. $\mathscr{D}_-$) denote the linear space of $\mathbb{R}$-valued infinitely differentiable functions defined on $\mathbb{R}$ with compact supports (resp., with supports bounded on the right). With the Schwartz topology [4], $\mathscr{D}$ and $\mathscr{D}_-$ are Hausdorff locally convex linear topological spaces. Let $\mathscr{D}'$ (resp. $\mathscr{D}'_+$) denote the dual of $\mathscr{D}$ (resp. $\mathscr{D}_-$) with the strong topology. Then $\mathscr{D}'_+$ is the space of $\mathbb{R}$-valued distributions on $\mathbb{R}$ with support bounded on the left. The canonical injections $\mathscr{D} \to \mathscr{D}'_+ \to \mathscr{D}'$ are continuous and $\mathscr{D}$ is dense in $\mathscr{D}'_+$ (see [4]).

From the results of Schwartz [4], with the operations of addition and convolution $\mathscr{D}'_+$ is a commutative (topological) ring with no divisors of zero. Given $u, v \in \mathscr{D}'_+$, the convolution of u and v, denoted by $u * v$, is defined by

$$\langle u * v, \varphi \rangle = \langle u, \langle v, \varphi(t+\tau) \rangle \rangle, \quad \text{all } \varphi \in \mathscr{D}. \tag{2.1}$$

It is easily verified that, if $u, v \neq 0$, then $\operatorname{supp}(u * v) \subseteq (\operatorname{supp} u) + (\operatorname{supp} v)$ where supp denotes the support. The identity of the ring $\mathscr{D}'_+$ is the Dirac distribution δ_0. We also note that the linear structure on $\mathscr{D}'_+$ is compatible with the ring structure in that $\mathscr{D}'_+$ is a convolution algebra over $\mathbb{R}$. To simplify the notation, from here on we let V denote the ring $\mathscr{D}'_+$.

For any fixed positive integer n, let V^n denote the n-fold direct sum of V with the elements of V^n written as column vectors. Then V^n is a free n-dimensional topological module over the ring V. Given $v \in V^n$ and $\alpha \in V$, we let $\alpha * v$ denote the operation of α on v in the V-module structure of V^n.

DEFINITION 2.1. Given fixed positive integers m and k, an *input/output* (i/o) *operator* f is a $\mathbb{R}$-linear continuous map $f : V^m \to V^k$.

As usual, an i/o operator $f: V^m \to V^k$ characterizes the correspondence between input functions (in V^m) and the resulting output functions (in V^k) for some m-input terminal k-output terminal linear continuous-time system. There are two main reasons for taking $V = \mathscr{D}'_+$ as the space of "admissible signals" appearing at the input and output terminals. First, V is a convolution ring containing δ_0 which, as mentioned in the Introduction, is necessary for the algebraic constructions that follow. Second, the class of systems describable by an i/o operator $f: V^m \to V^k$ is extremely large, including, for example, distributed-parameter devices such as LC and RC transmission lines.

Unfortunately, the topology on V is not normable, and in some applications it may be highly desirable to work with a convolution ring (with δ_0) having a nice topological structure, such as a Banach convolution algebra (Bensoussan and Kamen [5]). However, most of the results that follow can be carried over to these other rings.

In this paper, we restrict attention to i/o operators f having the property that $f(\delta_\tau * v) = \delta_\tau * f(v)$, all $\tau \in \mathbb{R}$, $v \in V^m$; that is, f commutes with the shift operator δ_τ. Such i/o operators are said to be time invariant or constant.

Letting $V^{k\times m}$ denote the V-module of $k \times m$ matrices over V, we have the following result on the representation of i/o operators.

THEOREM 2.1. *For each time-invariant i/o operator $f: V^m \to V^k$, there exists a unique $W \in V^{k\times m}$ such that $f(v) = W * v$ for all $v \in V^m$. Conversely, given $W \in V^{k\times m}$, the operator $V^m \to V^k : v \mapsto W * v$ is a time-invariant i/o operator.*

Proof. The proof follows from the Schwartz kernel theorem [4] using the fact that the canonical injections $\mathscr{D} \to V \to \mathscr{D}'$ are continuous and $\mathscr{D}$ is dense in V.

COROLLARY 2.1. *With respect to the topological V-module structure on V^m and V^k, every time-invariant i/o operator $f: V^m \to V^k$ is a (topological) V-module homomorphism.*

COROLLARY 2.2. *For fixed positive integers m and k, the V-module consisting of all time-invariant i/o operators $f: V^m \to V^k$ is isomorphic to $V^{k\times m}$.*

The matrix W whose existence is asserted in Theorem 2.1, is usually referred to as the impulse response matrix. A major point here is that the existence of W is directly connected to the fact that the i/o operator is a V-module homomorphism. The basic idea of this work is to exploit the module structure. But before we begin to do this, we need to consider the notion of causality in the space V.

DEFINITION 2.2. An i/o operator $f: V \to V : v \mapsto w * v$, $w \in V$, is *causal* if whenever $u|_{(-\infty,\tau)} = v|_{(-\infty,\tau)}$, $u, v \in V$, $\tau \in \mathbb{R}$, then $f(u)|_{(-\infty,\tau)} = f(v)|_{(-\infty,\tau)}$, where $|_{(-\infty,\tau)}$ denotes restriction to the open interval $(-\infty, \tau)$ in the sense of distributions.

PROPOSITION 2.1. *Given $f: V \to V : v \mapsto w * v$, the following are equivalent:*

(i) *f is causal,*

(ii) *if* $\operatorname{supp} v \subseteq [\tau, \infty)$, $v \in V$, $\tau \in \mathbb{R}$, *then* $\operatorname{supp} f(v) \subseteq [\tau, \infty)$,

(iii) $\operatorname{supp} w \subseteq [0, \infty)$.

Proof. (i)$\Rightarrow$(ii). Let $v \in V$ with $\operatorname{supp} v \subseteq [\tau, \infty)$. Then since $v|_{(-\infty,\tau)} = 0$ and f is causal, $f(v)|_{(-\infty,\tau)} = f(0)|_{(-\infty,\tau)} = 0$. Thus $\operatorname{supp} f(v) \subseteq [\tau, \infty)$. (ii)$\Rightarrow$(iii). Since $\operatorname{supp} \delta_0 = \{0\}$, by (ii) $\operatorname{supp} f(\delta_0) \subseteq [0, \infty)$. But $f(\delta_0) = w * \delta_0 = w$. (iii)$\Rightarrow$(i). Suppose that $u|_{(-\infty,\tau)} = v|_{(-\infty,\tau)}$. Then $\operatorname{supp}(u - v) \subseteq [\tau, \infty)$ and since $\operatorname{supp} w \subseteq [0, \infty)$, $\operatorname{supp}[w * (u - v)] \subseteq [0, \infty) + [\tau, \infty) = [\tau, \infty)$. Therefore, $\operatorname{supp} f(u - v) \subseteq [\tau, \infty)$ which implies that $f(u)|_{(-\infty,\tau)} = f(v)|_{(-\infty,\tau)}$ since f is additive.

Even though supp $w \subseteq [0, \infty)$ for a causal operator, in general it is not possible to construct the impulse response w from the restriction $w|_{(0,\infty)}$. This situation can occur when w is not regular on any neighborhood of the origin. (A distribution $v \in V$ is regular on an open set U if $v|_U$ can be generated in the usual manner from a locally integrable function on U.) For example, if $w = \delta_0 + e^{-t}H(t)$, $H(t) =$ Heaviside function, then $w|_{(0,\infty)} = e^{-t}|_{(0,\infty)}$ which does not contain any knowledge of the singular component δ_0.

Many system problems, such as the problem of realization, involve the restriction $w|_{(0,\infty)}$ of the impulse response w, assuming that w can be determined uniquely from $w|_{(0,\infty)}$. A sufficient condition for the determination of w from $w|_{(0,\infty)}$ is given in the following.

PROPOSITION 2.2. *Let $w \in V$ with* supp $w \subseteq [0, \infty)$. *If there exists an open neighborhood U of the origin such that $w|_U$ is a regular distribution, then w can be completely and uniquely determined from $w|_{(0,\infty)}$.*

Proof. Suppose that w satisfies the hypothesis of the proposition. Let a be a positive number belonging to U and write $w_+ = w|_{(0,\infty)}$. Since w_+ is regular on $(0, a)$, from w_+ we can construct the following regular distribution on $(-\infty, a)$:

$$w_a(t) = \begin{cases} w_+(t), & 0 < t < a, \\ 0, & t \leqq 0. \end{cases}$$

Then since $w_+ = w_a$ on $(0, a)$, by the theorem on "piecing together distributions" (Zemanian [6, p. 34]), from w_+ and w_a it is possible to construct one and only one distribution θ on $\mathbb{R}$ such that $\theta|_{(-\infty, a)} = w_a$ and $\theta|_{(0,\infty)} = w_+$. Further, θ is clearly independent of the value chosen for a. Now by construction, $\theta = w$ on $(-\infty, 0) \cup (0, \infty)$. Hence $\theta = w$ on $\mathbb{R}$ since the Lebesgue measure of $\{0\}$ is zero and both w and θ are regular on the open neighborhood U.

DEFINITION 2.3. A causal i/o operator f with impulse response w is said to be *strictly causal* if $w|_U$ is regular for some open neighborhood U of $\{0\}$.

The term strictly causal is taken from the work of Saeks [7]. Although Saek's formulation of causality is developed in terms of an abstract Hilbert space rather than a space of distributions, his definition of strictly causal is similar to that given here.

In many cases the impulse response w is an ordinary function (i.e., a regular distribution) with supp $w \subseteq [0, \infty)$, and thus the i/o operator is strictly causal as defined above. On the other hand, there exist important examples of systems whose impulse responses are not regular and yet the corresponding i/o operators are strictly causal. These systems are necessarily infinite-dimensional; that is, the Laplace transform of the impulse response is not rational. A simple example is the ideal delay line with impulse response δ_τ, $\tau > 0$.

An interesting class of causal operators which are not strictly causal is the class of operators having supp $w = \{0\}$. By a well-known theorem of distribution theory (Zemanian [6, p. 98]), supp $w = \{0\}$ if and only if w is a finite $\mathbb{R}$-linear combination of δ_0 and its derivatives. If we let $\delta_0^{(n)}$ denote the nth derivative of δ_0, then since $\delta_0^{(n)} * v = v^{(n)} = n$th derivative of $v \in V$, for an i/o operator f with supp $w = \{0\}$ the response $f(v)$ is a finite linear combination of the input v and its derivatives.

Most causal operators of interest can be decomposed uniquely into the sum of a strictly causal operator and an operator with impulse response concentrated at the origin.

PROPOSITION 2.3. *Given a causal operator* $f : V \to V : v \mapsto w * v$, *if there exists an* $a > 0$ *such that* $w|_{(0,a)}$ *is regular, then* f *can be decomposed uniquely into the sum* $f = f_{sc} + f_0$, *where* f_{sc} *is strictly causal and* $f_0 : v \mapsto w_0 * v$ *with* $w_0 = 0$ *or* supp $w_0 = \{0\}$.

Proof. Let w satisfy the hypothesis. As in the proof of Proposition 2.2, from $w|_{(0,\infty)}$ we can construct a distribution θ on $\mathbb{R}$ such that $\theta = w$ on $\mathbb{R} - \{0\}$ and the operator $f_{sc} : v \mapsto \theta * v$ is strictly causal. Now define $w_0 = -\theta + w$. Then $w_0 = 0$ or supp $w_0 = \{0\}$ and $f = f_{sc} + f_0$, where $f_0 : v \mapsto w_0 * v$.

Uniqueness. Suppose that $f = \bar{f}_{sc} + \bar{f}_0$, where $\bar{f}_{sc}$ is strictly causal and $\bar{f}_0(v) = \bar{w}_0 * v$, $\bar{w}_0 \neq w_0$, $\bar{w}_0 = 0$ or supp $\bar{w}_0 = \{0\}$. Then since $f_{sc} + f_0 = \bar{f}_{sc} + \bar{f}_0$, $f_0 - \bar{f}_0 = \bar{f}_{sc} - f_{sc}$. But this is impossible since the operator $\bar{f}_{sc} - f_{sc}$ is strictly causal and supp $(w_0 - \bar{w}_0) = \{0\}$.

The above results are easily extendable to the multi-input multi-output case. In particular, the i/o operator $f : V^m \to V^k : v \mapsto W * v$, $W = (w_{ij}) \in V^{k \times m}$, is strictly causal if for each i, j, there exists an open neighborhood U_{ij} of $\{0\}$ such that $w_{ij}|_{U_{ij}}$ is regular. In the remainder of this paper, we limit our study to strictly causal i/o operators.

3. State in a module framework. In this section we formulate a definition of systems which reflects the convolution module structure of the i/o representation. In order to express the concept of state in terms of the convolution structure, we need to define another type of i/o operator which is a module homomorphism between modules defined over a proper subring of $V = \mathscr{D}'_+$.

Let Ω denote the subring of V consisting of all distributions having compact support contained in $(-\infty, 0]$. With the induced topology, Ω is a topological subring of V, and the m-fold direct sum Ω^m is a free m-dimensional topological module over the ring Ω. (Throughout this paper it is understood that the topology of all modules considered is Hausdorff and locally convex.)

Let Γ denote the set $\{v|_{(0,\infty)} : v \in V\}$. With the induced operations, Γ is a linear subspace of $\mathscr{D}'(0, \infty)$, the space of all distributions defined on $(0, \infty)$. Further, it follows from the discussion given by Treves [8, p. 246] that Γ is a proper subspace of $\mathscr{D}'(0, \infty)$. We give Γ the strongest topology such that the map

$$\rho : V \to \Gamma : v \mapsto v|_{(0,\infty)} \tag{3.1}$$

is continuous. Note that ρ is also an open mapping since a set U_1 is a neighborhood of zero in Γ if and only if there is a neighborhood U_2 of zero in V such that $\rho(U_2) = U_1$.

PROPOSITION 3.1. Γ *is a topological module with multiplication*

$$\Omega \times \Gamma \to \Gamma : (\omega, \gamma) \mapsto \omega\gamma \triangleq (\omega * \bar{\gamma})|_{(0,\infty)}, \tag{3.2}$$

where $\bar{\gamma} \in V$ *is any extension of* γ *to* V (*i.e.*, $\bar{\gamma}|_{(0,\infty)} = \gamma$).

Proof. Multiplication (3.2) is independent of the extension considered. Let $\gamma \in \Gamma$. Then by definition of Γ, γ has at least one extension $\bar{\gamma} \in V$. Suppose that $\bar{\gamma}$ and $\bar{\gamma}'$ are two extensions of γ and let $\omega \in \Omega$. Then

$$\langle \bar{\gamma}, \varphi \rangle = \langle \bar{\gamma}', \varphi \rangle, \quad \text{all } \varphi \in \mathscr{D} : \text{supp } \varphi \subset (0, \infty). \tag{3.3}$$

Now since supp $\omega \subset (-\infty, 0]$, for every $\tau \leqq 0$,

$$\langle \omega, \varphi(t+\tau)\rangle = 0, \quad \text{all } \varphi : \text{supp } \varphi \subset (0, \infty).$$

Thus viewed as a function of τ, $\langle \omega, \varphi(t+\tau)\rangle$ is an element of $\mathscr{D}$ with support contained in $(0, \infty)$. Then from (3.3), we have

$$\langle \bar{\gamma}, \langle \omega, \varphi(t+\tau)\rangle\rangle = \langle \bar{\gamma}', \langle \omega, \varphi(t+\tau)\rangle\rangle, \quad \text{all } \varphi : \text{supp } \varphi \subset (0, \infty).$$

From the definition of convolution (2.1), we get

$$\langle \omega * \bar{\gamma}, \varphi\rangle = \langle \omega * \bar{\gamma}', \varphi\rangle, \quad \text{all } \varphi : \text{supp } \varphi \subset (0, \infty).$$

Thus $(\omega * \bar{\gamma})|_{(0,\infty)} = (\omega * \bar{\gamma}')|_{(0,\infty)}$, showing that multiplication (3.2) is properly defined. The proof that Γ with (3.2) is a topological module follows from the fact that ρ, given by (3.1), is open and continuous. The straightforward details are omitted.

COROLLARY 3.1. *The k-fold direct sum* Γ^k *is a (nonfinite) topological module over the ring* Ω.

Let $I : \Omega^m \to V^m$ denote the inclusion map and define the map $P : V^k \to \Gamma^k : (v_1, \cdots, v_k)^{\text{TR}} \mapsto (\rho(v_1), \cdots, \rho(v_k))^{\text{TR}}$, where TR denotes the transpose.

THEOREM 3.1. *Given a strictly causal i/o operator* $f : V^m \to V^k : v \mapsto W * v$, *let* f^* *denote the composition PfI. Then*

(i) f^* *is a (topological)* Ω*-module homomorphism,*

(ii) f^* *is completely and uniquely determined by* $W|_{(0,\infty)}$ *and vice versa,*

(iii) f *can be completely and uniquely constructed from* f^*.

Proof. (i) By definition of Ω^m, the inclusion map $I : \Omega^m \to V^m$ is an Ω-module homomorphism with V^m viewed as an Ω-module. It is also clear that $P : V^k \to \Gamma^k$ is an Ω-module homomorphism with V^k viewed as an Ω-module. Hence the composition $PfI = f^* : \Omega^m \to \Gamma^k$ is an Ω-module homomorphism.

(ii) Let $\omega \in \Omega^m$. Then $\omega = \sum_i \omega_i * e_i$, where

$$e_i = (0 \;\; 0 \cdots \underset{\substack{\uparrow \\ i\text{th place}}}{\delta_0} \;\; 0 \cdots 0)^{\text{TR}}$$

Since f^* is an Ω-module homomorphism, $f^*(\omega) = \sum_i \omega_i f^*(e_i)$, and by definition of f^*, $f^*(e_i) = (W * e_i)|_{(0,\infty)}$, where γ_i is the ith column of W. Hence $W|_{(0,\infty)}$ determines f^* uniquely and conversely.

(iii) This follows from (ii) and Proposition 2.2.

The operator $f^* = PfI$ characterizes the input/output behavior relative to the time reference $t = 0$. As will be done shortly, we can define the state space to be some space through which f^* is factored. The module structure comes into play by requiring that the factorization consist of Ω-module homomorphisms.

Let l denote the map

$$l : V \to \mathbb{R} : v \mapsto l(v) = \begin{cases} \inf_t \{t \in \text{supp } v\}, & v \neq 0, \\ 0, & v = 0. \end{cases} \tag{3.4}$$

Since $\operatorname{supp}(u * v) \subseteq (\operatorname{supp} u) + (\operatorname{supp} v)$, $u, v \in V$, $u, v \neq 0$,

$$l(u * v) \geqq l(u) + l(v), \quad \text{all } u, v \neq 0. \tag{3.5}$$

Note that $l(v) \leqq 0$ for any $v \in \Omega$. Finally, l can be extended to V^m by defining

$$l : (v_1, \cdots, v_m)^{\mathrm{TR}} \mapsto \min_i \{l(v_i)\}. \tag{3.6}$$

DEFINITION 3.1. An *m-input k-output strictly causal linear time-invariant system* Σ is a sextuple $\Sigma = (\Omega^m, X, \Gamma^k, \mu, \eta, \psi)$, where

(i) X is a topological Ω-module with multiplication denoted by $\pi \cdot x$, $\pi \in \Omega$, $x \in X$;

(ii) $\mu : \Omega^m \to X$ and $\eta : X \to \Gamma^k$ are (topological) Ω-module homomorphisms with the composition $\eta\mu$ equal to PfI for some strictly causal time-invariant i/o operator f;

(iii) ψ is a map defined by

$$\psi : \Omega^m \times X \times \mathbb{R}^- \to X : (\omega, x, a) \mapsto \delta_{a+l(\omega)} \cdot x + \mu(\omega),$$

where $\mathbb{R}^- = \{a \in \mathbb{R} : a < 0\}$.

In this definition, X is the module of states, $\mu(\omega)$ is the state at time $t = 0$ due to input $\omega \in \Omega^m$, and $\eta(x)$ is the output response on $(0, \infty)$ resulting from state x at $t = 0$. The map ψ is a state transition operator: $\psi(\omega, x, a)$ is the state at $t = 0$ due to input ω and initial state x at time $t = a + l(\omega)$ prior to the application of ω. The parameter a in the definition of ψ cannot be zero, in general, because the input ω may contain Dirac distributions at $\{l(\omega)\}$. Since the input (function) module Ω^m and the output (function) module Γ^k are fixed, we shall usually write $\Sigma = (X, \mu, \eta, \psi)$.

Note that since the composition $\eta\mu$ equals PfI for some strictly causal operator f, by Theorem 3.1, knowledge of $\eta\mu$ is equivalent to knowledge of f. Therefore f can be (and will be) taken as *the* i/o operator of the system $\Sigma = (X, \mu, \eta, \psi)$.

Although the definition of a system Σ is specified with respect to the time reference $t = 0$, this does not result in any special restrictions, other than those already given, since Σ is time invariant. The time invariance of Σ is a consequence of the fact that μ and η are Ω-module homomorphisms.

The requirement that the state set X admit a module structure over the convolution ring Ω is actually a very natural condition since we are considering systems whose input/output behavior is given by an Ω-module homomorphism. Furthermore, as shown in the next section, every strictly causal i/o operator can be realized by a system having an Ω-module structure.

One final point here is that since $\mathbb{R}$ can be viewed as a subring of Ω under the embedding $\mathbb{R} \to \Omega : a \mapsto a\delta_0$, X is also a linear space over $\mathbb{R}$. Thus the module structure on X "contains" the usual linear space structure. A system Σ is infinite-dimensional in the usual sense if X is infinite-dimensional as a linear space over $\mathbb{R}$.

4. Realization of input/output operators. Following the standard definitions, we say that a system $\Sigma = (X, \mu, \eta, \psi)$ is completely reachable (resp. completely observable) if μ is surjective (resp. η is injective). In the first part of this section it

is proved that every strictly causal i/o operator can be realized by a system that is completely reachable and observable. Then we consider realizations given by differential equations in the sense of distributions.

DEFINITION 4.1. A *realization of a strictly causal i/o operator* $f : V^m \to V^k$ is a system $\Sigma = (X, \mu, \eta, \psi)$ with $\eta\mu = PfI$. A realization is said to be canonical if it is completely reachable and observable.

THEOREM 4.1. *Every strictly causal time-invariant i/o operator f has a canonical realization.*

Proof. Given f, let $f^* = PfI$; since Γ^k is a Hausdorff space, $\{0\}$ is a closed set in Γ^k, and by the continuity of f^*, $\ker f^* \triangleq \{\omega \in \Omega^m : f^*(\omega) = 0\}$ is a closed set in Ω^m. Hence the quotient space $\Omega^m/\ker f^* \triangleq \{[\omega] \triangleq \omega + \ker f^* : \omega \in \Omega^m\}$ with the quotient topology is a Hausdorff locally convex linear topological space. Further, it is easily checked that $\Omega^m/\ker f^*$ is a topological Ω-module with multiplication $\pi \cdot [\omega] \triangleq [\pi * \omega]$, $\pi \in \Omega$, $\omega \in \Omega^m$. Now take $X_f \triangleq \Omega^m/\ker f^*$ to be the state module, and define the following Ω-module homomorphisms:

$$\mu_f : \Omega^m \to X_f : \omega \mapsto [\omega],$$

$$\eta_f : X_f \to \Gamma^k : [\omega] \mapsto f^*(\omega).$$

Clearly, μ_f is surjective and η_f is injective. Given $\omega \in \Omega^m$, $\mu_f(\omega) = [\omega]$ is defined to be the state at time $t = 0$ due to input ω, and for every $\tau \leqq 0$, $\mu_f(\delta_\tau * \omega) = \delta_\tau \cdot [\omega]$ is the state at time $t = 0$ due to state $[\omega]$ at time τ. Therefore, if the input $\omega \in \Omega^m$ is applied with initial state $x = [\beta]$ at time $a + l(\omega)$, $a < 0$, the state $\psi_f(\omega, x, a)$ at $t = 0$ is given by $\psi_f(\omega, x, a) = \delta_{a+l(\omega)} \cdot x + \mu_f(\omega)$. Finally, since $f^* = \eta_f\mu_f$, $(X_f, \mu_f, \eta_f, \psi_f)$ is a canonical realization of f.

Regarding the uniqueness of canonical realizations, we have the following.

PROPOSITION 4.1. *If* (X, μ, η, ψ) *and* $(\hat{X}, \hat{\mu}, \hat{\eta}, \hat{\psi})$ *are two canonical realizations of an i/o operator f, then with respect to the algebraic structure there exists a unique module isomorphism* $\xi : X \to \hat{X}$ *with* $\xi\mu = \hat{\mu}$ *and* $\hat{\eta}\xi = \eta$.

Proof. The proof follows from a standard isomorphism theorem.

COROLLARY 4.1. *If the composition PfI is an open mapping,* ξ *is a topological* Ω*-module isomorphism; that is,* ξ *is also a homeomorphism.*

Proof. Suppose that $f^* = PfI$ is open and let U be an open set in $\hat{X}$. Then $(f^*\hat{\mu}^{-1})(U)$ is open in Γ^k since $\hat{\mu}$ is continuous. Since η is injective and $\hat{\mu}$ is surjective and $f^* = \hat{\eta}\hat{\mu}$, $(\eta^{-1}f^*\hat{\mu}^{-1})(U) = \xi^{-1}(U)$ which is open in X because η is continuous. Hence ξ is continuous. A similar proof shows that ξ^{-1} is continuous.

In many applications it is desirable to have a realization given by dynamical differential equations. For example, with such a realization it would be possible to apply the theory of differential equations to the study of optimal control. As we now show, i/o operators can be realized by differential equations in the sense of distributions.

Given the i/o operator $f : V^m \to V^k$, let $\Sigma = (X, \mu, \eta, \psi)$ denote the canonical realization of f constructed in the proof of Theorem 4.1. Following Kalman and Hautus [2], define the truncation operator $\mathcal{S} : \mathcal{D}^m \to \Omega^m : \alpha \mapsto \mathcal{S}\alpha$, where $(\mathcal{S}\alpha)(t) = 0$, $t > 0$, and $(\mathcal{S}\alpha)(t) = \alpha(t)$, $t \leqq 0$. For every $\omega \in \Omega^m$, define

$$x_\omega : \mathcal{D} \to X : \varphi \mapsto [\mathcal{S}(\check{\varphi} * \omega)], \qquad \check{\varphi}(t) = \varphi(-t).$$

Note that since $\pi * \varphi \in \mathscr{D}$, all $\pi \in \Omega$, $\varphi \in \mathscr{D}$ (see [6]), x_ω is properly defined. As proved in [2], x_ω is an X-valued distribution. The interpretation of x_ω is that it is the generalized state trajectory resulting from the application of the input ω.

Now define

$$F : X \to X : [\omega] \mapsto [\delta_0^{(1)} * \omega],$$

$$G : \mathbb{R}^m \to X : (a_1, \cdots, a_m)^{\mathrm{TR}} \mapsto [(a_1\delta_0, \cdots, a_m\delta_0)^{\mathrm{TR}}].$$

Then for all $\varphi \in \mathscr{D}$, x_ω satisfies the differential equation

$$\frac{dx_\omega(\varphi)}{dt} = Fx_\omega(\varphi) + G\omega(\varphi). \tag{4.1}$$

The proof follows from [2].

Hence we have an internal differential equation describing the realization. However we do not have an output equation as constructed in [2] because here the output response on $(0, \infty)$ may not be an ordinary function. Nevertheless, in most cases it is possible to formulate an output equation as follows.

Let $\bar{X} = \{[\sigma] : \sigma \in \mathscr{S}(\mathscr{D}^m)\}$ which is a linear subspace of X viewed as a linear space over $\mathbb{R}$. Suppose that for each $\sigma \in \mathscr{S}(\mathscr{D}^m)$, $f(\sigma)$ is continuous on some neighborhood of zero. Then since $f(\beta) = f(\sigma)$ on $(0, \infty)$ for every $\beta \in [\sigma]$, we can define the operator

$$H : \bar{X} \to \mathbb{R}^k : [\sigma] \mapsto f(\sigma)(0) = \lim_{t \downarrow 0} f(\sigma)(t).$$

Let $\omega \in \Omega^m$. Then for every $\varphi \in \mathscr{D}$, we have that

$$\begin{aligned}
&\langle f(\omega), \varphi\rangle = (\check{\varphi} * f(\omega))(0) \\
\Rightarrow &\langle f(\omega), \varphi\rangle = (f(\check{\varphi} * \omega))(0) && \text{since } f \text{ is a } V\text{-module homomorphism} \\
\Rightarrow &\langle f(\omega), \varphi\rangle = (f(\mathscr{S}(\check{\varphi} * \omega)))(0) && \text{since } f \text{ is strictly causal} \\
\Rightarrow &\langle f(\omega), \varphi\rangle = H[\mathscr{S}(\check{\varphi} * \omega)] && \text{by definition of } H.
\end{aligned}$$

Thus

$$\langle f(\omega), \varphi\rangle = Hx_\omega(\varphi), \qquad \omega \in \Omega^m, \quad \varphi \in \mathscr{D}. \tag{4.2}$$

Hence we have proved the following.

THEOREM 4.2. *Given the* i/o *operator* $f : V^m \to V^k : v \mapsto W * v$, *if for each* $\sigma \in \mathscr{S}(\mathscr{D}^m)$, $f(\sigma)$ *is continuous on some neighborhood of zero, then* f *has a canonical realization which can be described by dynamical differential equations* (*given by* (4.1–4.2)).

Instead of working with dynamical differential equations, in the remainder of this paper we consider only the operational form of a system Σ as given in Definition 3.1. The objective is to study dynamical properties by using the module structure on $\Sigma = (X, \mu, \eta, \psi)$.

5. Controllability in a module framework. In terms of the $\mathbb{R}$-linear structure, few algebraic results exist on the controllability of infinite systems simply because

the state space is infinite-dimensional as a linear space. However, as a consequence of finiteness properties of the module structure, it is possible to study control from an algebraic standpoint. We shall do this here, setting up the theory in terms of a general framework that includes state and output function controllability. In the following development, the topological structure is not considered.

Let M be an Ω-module, and let $\lambda : \Omega^m \to M$ be an Ω-module homomorphism.

DEFINITION 5.1. An element $x \in M$ is *reachable* if there exists an $\omega \in \Omega^m$ such that $\lambda(\omega) = x$. An element $x \in M$ is *controllable* if there exist $\tau < 0$ and $\omega \in \Omega^m$ with $l(\omega) > \tau$, such that $\delta_\tau \cdot x + \lambda(\omega) = 0$. The element ω is called *a control* for x and $-\tau$ is a *control time*.

Given a system $\Sigma = (X, \mu, \eta, \psi)$, state controllability and a type of output function controllability are particular cases of the above definition.

1. *State controllability.* Take $M = X$ and $\lambda = \mu$. Then $x \in X$ is reachable if there exists an input in Ω^m which sets up (from the zero state) the state x at time $t = 0$, and x is controllable if there exists an input in Ω^m which drives the system to the zero state at $t = 0$ starting from state x at some time τ prior to the application of the input.

2. *Output function controllability.* Take $M = \Gamma^k$ and $\lambda = PfI = f^*$, where f is the i/o operator of the system Σ. Then an output function $\gamma \in \Gamma^k$ is reachable if there exists an input in Ω^m which produces this response with zero initial state prior to the application of the input. An element $\gamma \in \Gamma^k$ is controllable if there exist $\tau < 0$, $\omega \in \Omega^m$, $l(\omega) > \tau$, such that $\delta_\tau \gamma + f^*(\omega) = 0$ which implies that

$$\gamma|_{(-\tau, \infty)} + f(\delta_{-\tau} * \omega)|_{(-\tau, \infty)} = 0. \tag{5.1}$$

Therefore, viewing γ as an output response on $(0, \infty)$ due to an input and/or initial state occurring in the time interval $(-\infty, 0]$, by (5.1) we have that the input $\delta_{-\tau} * \omega$ (applied during the interval $(0, -\tau]$) drives the output response to zero on $(-\tau, \infty)$.

The objective here is to study controllability in terms of the general framework given in Definition 5.1. All of the following results specialize to state and output function controllability by setting $\lambda = \mu$ or PfI as done above. We begin with the following basic definitions from module theory.

Given an Ω-module M, $x \in M$ is said to be a free element if $\pi x = 0$ for some $\pi \in \Omega$, then $\pi = 0$. If there exists a nonzero $\pi \in \Omega$ such that $\pi x = 0$, x is called a torsion (or nonfree) element. Since Ω is an integral domain (i.e., Ω is a commutative ring with no divisors of zero), the set $T(M)$ of torsion elements of M is a submodule of M.

Let S be a subset of M. The annihilator of S, denoted by Ann (S), is the set of elements $\pi \in \Omega$ such that $\pi x = 0$ for all $x \in S$. For any subset $S \subset M$, Ann (S) is an ideal of the ring Ω. If $S = \{x\}$, we write Ann $(S) =$ Ann (x).

Given an Ω-module homomorphism $\lambda : \Omega^m \to M$, let M_r denote the submodule of M consisting of all reachable elements; that is, $M_r = \lambda(\Omega^m)$. Since Ω^m is a finitely-generated Ω-module, M_r is also finitely generated, in particular,

$$M_r = \sum_{i=1}^{m} \Omega q_i, \quad \text{where } q_i = \lambda(e_i), \quad e_i = (0\ 0 \cdots \underset{\substack{\uparrow \\ i\text{th place}}}{\delta_0} 0 \cdots 0)^{\mathrm{TR}}.$$

It is easily verified that $\text{Ann}(M_r) = \cap_i \text{Ann}(\Omega q_i) = \cap_i \text{Ann}(q_i)$. Using this fact, we can prove the following.

PROPOSITION 5.1. *Suppose that* $M_r \neq \{0\}$. *Then the following are equivalent*:

(i) $\text{Ann}(M_r) \neq \{0\}$,

(ii) $M_r \subset T(M)$,

(iii) *each nontrivial submodule* Ωq_i *contains a nonzero torsion element.*

Proof. Obviously, (i)$\Rightarrow$(ii) and (ii)$\Rightarrow$(iii).

(iii)$\Rightarrow$(i). Suppose that for each $q_i \neq 0$, there exist $0 \neq x_i \in \Omega q_i$ and $0 \neq \pi_i \in \Omega$ such that $\pi_i x_i = 0$. Then since $x_i = \omega_i q_i$ for some $\omega_i \in \Omega$, $\omega_i \neq 0$, we have that $\pi_i x_i = (\pi_i * \omega_i) q_i = 0$. Hence $0 \neq \pi_i * \omega_i \in \text{Ann}(q_i)$ and the product $\prod_i (\pi_i * \omega_i) \neq 0$ annihilates each q_i, $i = 1, 2, \cdots, m$. Thus $\prod_i (\pi_i * \omega_i) \in \cap_i \text{Ann}(q_i) = \text{Ann}(M_r)$.

The following result shows that if $M_r \neq \{0\}$ and any one of the equivalent statements of Proposition 5.1 is not true, then for at least one i such that $q_i \neq 0$, every nonzero state in Ωq_i is uncontrollable.

PROPOSITION 5.2. *If* $M_r \neq \{0\}$ *and for each* i *such that* $q_i \neq 0$, *the submodule* Ωq_i *contains a nonzero controllable element, then* $\text{Ann}(M_r) \neq \{0\}$.

Proof. Suppose that the hypothesis is satisfied. Then for each i such that $q_i \neq 0$, there exist $0 \neq x_i = \omega_i q_i$, $\tau_i < 0$, and $u_i \in \Omega^m$ with $l(u_i) > \tau_i$, such that

$$\delta_{\tau_i}(\omega_i q_i) + \lambda(u_i) = (\delta_{\tau_i} * \omega_i) q_i + \lambda(u_i) = 0.$$

Writing $u_i = \sum_j u_{ij} e_j$, $u_{ij} \in \Omega$, with $l(u_{ij}) > \tau_i$, we have

$$(\delta_{\tau_i} * \omega_i) q_i + \sum_j u_{ij} q_j = 0.$$

Hence

$$(\delta_{\tau_i} * \omega_i + u_{ii}) q_i + \sum_{\substack{j \\ j \neq i}} u_{ij} q_j = 0. \tag{5.2}$$

Now for each i such that $q_i = 0$, we have

$$\delta_0 q_i = 0. \tag{5.3}$$

Let C denote the $m \times m$ matrix consisting of the coefficients of the q_i in equations (5.2–3) such that the diagonal elements of C are $\delta_{\tau_i} * \omega_i + u_{ii}$ or δ_0. Then (Lang [9, p. 335]) the determinant of C, denoted by det C, annihilates each q_i, and thus $\det C \in \text{Ann}(M_r)$. It must be shown that $\det C \neq 0$. By construction, $\det C$ is of the form

$$\det C = \left(\prod_i (\delta_{\tau_i} * \omega_i) \right) + \pi,$$

where $l(\pi) > \sum_i \tau_i$ since $l(u_{ij}) > \tau_i$. Hence the support of π does not intersect the support of the product $\prod_i (\delta_{\tau_i} * \omega_i)$ and since $\prod_i (\delta_{\tau_i} * \omega_i) \neq 0$, $\det C \neq 0$.

COROLLARY 5.1. *If* M_r *contains a free element, then for at least one* i, $q_i \neq 0$ *and every nonzero element of* Ωq_i *is uncontrollable.*

COROLLARY 5.2. *If $m = 1$ and $T(M_r) = \{0\}$, no nonzero element of M is both reachable and controllable.*

PROPOSITION 5.3. *Suppose that for some fixed i, q_i is free and $q_1, \cdots, q_{i-1}, q_{i+1}, \cdots, q_m$ have torsion. Then M_r can be written as an internal direct sum*

$$M_r = M_1 \oplus \Omega q_i, \quad \text{where } M_1 = \sum_{\substack{j=1 \\ j \neq i}}^{m} \Omega q_i$$

and every element $x = x_1 + x_i$ is uncontrollable, where $x_1 \in M_1$, $x_i \in \Omega q_i$, $x_i \neq 0$.

Proof. Clearly, $M_r = M_1 + \Omega q_i$. Suppose that there exists an $x \neq 0$ with $x \in M_1 \cap \Omega q_i$. Then since Ann (M_1) is nontrivial by Proposition 5.1, there exists a $\pi \in \Omega$, $\pi \neq 0$, such that $\pi x = 0$. Now $x = \omega q_i$ for some $\omega \in \Omega$, $\omega \neq 0$, and thus $(\pi * \omega) q_i = 0$, $\pi * \omega \neq 0$, which is a contradiction if q_i is free.

Suppose that $x = x_1 + x_i$ is controllable, where $x_1 \in M_1$ and $x_i = \pi q_i \neq 0$. Then there exist $\tau < 0$, $\omega \in \Omega^m$, $l(\omega) > \tau$, such that $\delta_\tau(x_1 + \pi q_i) + \lambda(\omega) = 0$. Writing $\omega = \sum_j \omega_j e_j$ and since $x_1 = \sum_{j \neq i}^{j} \alpha_j q_j$, $\alpha_j \in \Omega$, we obtain

$$\sum_{\substack{j \\ j \neq i}} (\delta_\tau * \alpha_j + \omega_j) q_j + (\delta_\tau * \pi + \omega_i) q_i = 0.$$

Multiplying both sides of this equation by some $\beta \in \text{Ann}\,(M_1)$, $\beta \neq 0$, gives $\beta * (\delta_\tau * \pi + \omega_i) q_i = 0$, which is a contradiction since q_i is free.

As seen from the following results, the condition Ann $(M_r) \neq \{0\}$ is also related to the controllability of elements that are not necessarily reachable.

PROPOSITION 5.4. *If* Ann (M_r) *is nontrivial, every controllable element of M is contained in $T(M)$, the torsion submodule of M.*

Proof. If $x \in M$ is controllable, there exist $\tau < 0$, $\omega \in \Omega^m$, $l(\omega) > \tau$, such that $\delta_\tau x = -\lambda(\omega)$. Since $\lambda(\omega) \in M_r$, if Ann $(M_r) \neq \{0\}$, there exists $\pi \in \Omega$, $\pi \neq 0$, such that $(\pi * \delta_\tau) x = -\pi\lambda(\omega) = 0$. Hence x is a torsion element.

COROLLARY 5.3. *If* Ann $(M_r) \neq \{0\}$, *every free element of M is not controllable and not reachable.*

PROPOSITION 5.5. *Suppose that for each $i \in \{1, 2, \cdots, m\}$ there exists a nonzero torsion element of M which is controllable with control $\omega_i e_i \neq 0$, $\omega_i \in \Omega$. Then* Ann (M_r) *is nontrivial.*

Proof. Let $x_1, x_2, \cdots, x_m$ be nonzero torsion elements of M such that for each i, there exist $\tau_i < 0$, $\omega_i e_i \neq 0$, $l(\omega_i) > \tau_i$, with $\delta_{\tau_i} x_i + \lambda(\omega_i e_i) = 0$. Then if $\alpha_i x_i = 0$, $\alpha_i \neq 0$, $(\alpha_i * \delta_{\tau_i}) x_i = -\alpha_i \lambda(\omega_i e_i) = -(\alpha_i * \omega_i) q_i = 0$, and thus $\alpha_i * \omega_i \in \text{Ann}\,(q_i)$. Therefore the product $\prod_i (\alpha_i * \omega_i)$ is a nonzero element of Ann (M_r).

It follows from Corollary 5.1 that for every element of M_r to be controllable, it is necessary that Ann (M_r) be nontrivial. Whether or not the reachable states are controllable is an important question, since for any $x_1, x_2 \in M_r$, there exists a control $\omega \in \Omega^m$ which sets up x_2 from x_1 (i.e., $\delta_\tau x_1 + \lambda(\omega) = x_2$ for some $\tau < l(\omega)$) if and only if every element of M_r is controllable. It is interesting to note that when M_r is finite-dimensional as a linear space over $\mathbb{R}$, every $x \in M_r$ is controllable. The easy proof is omitted.

In terms of the module structure, we now develop a necessary and sufficient condition for controllability of M_r. We begin with the following ring-theoretic result.

LEMMA 5.1. *Let A be an ideal of the ring Ω and suppose that there exist $\tau<0$ and $\alpha\in\Omega$, $l(\alpha)>\tau$, such that $\delta_\tau+\alpha\in A$. Given $\omega\in\Omega$, let $s\geqq 0$ be an integer such that $(s+1)(\tau-l(\alpha))<l(\omega)$. Then $\delta_{(s+1)\tau}*\omega+\pi\in A$, where $\pi=(-1)^s\alpha^{s+1}*\omega$, $l(\pi)>(s+1)\tau$, $\alpha^{s+1}=(s+1)$-fold convolution of α.*

Proof. Given $\tau<0$, $\alpha\in\Omega$, $l(\alpha)>\tau$, such that $\delta_\tau+\alpha\in A$, by induction it is easily verified that for any integer $s\geqq 0$,

$$\delta_{(s+1)\tau}+(-1)^s\alpha^{s+1}=(\delta_\tau+\alpha)*\left(\sum_{i=0}^{s}(-1)^i\delta_{(s-i)\tau}*\alpha^i\right).$$

Then since A is an ideal of Ω and $\delta_\tau+\alpha\in A$, $(\delta_{(s+1)\tau}+(-1)^s\alpha^{s+1})*\omega\in A$ for any $\omega\in\Omega$. Now given a fixed $\omega\in\Omega$, we pick an integer $s\geqq 0$ such that $(s+1)(\tau-l(\alpha))<l(\omega)$. Such an integer can always be found since $\tau-l(\alpha)<0$. Then

$$\begin{aligned}
&(s+1)\tau<(s+1)l(\alpha)+l(\omega)\\
\Rightarrow&(s+1)\tau<l(\alpha^{s+1})+l(\omega) && \text{using (3.5)}\\
\Rightarrow&(s+1)\tau<l(\alpha^{s+1}*\omega) && \text{again using (3.5)}\\
\Rightarrow&(s+1)\tau<l(\pi) && \text{by definition of } \pi.
\end{aligned}$$

THEOREM 5.1. *Every $x\in M_r$ is controllable if and only if there exists $\delta_\tau+\alpha\in\operatorname{Ann}(M_r)$ with $l(\alpha)>\tau$.*

Proof. Recall that $M_r=\sum_{i=1}^m\Omega q_i$. If every $x\in M_r$ is controllable, each q_i is controllable, and thus for each i, there exist $\tau_i<0$, $u_i\in\Omega^m$, $l(u_i)>\tau_i$, with $\delta_{\tau_i}q_i+\lambda(u_i)=0$. Using the construction given in the proof of Proposition 5.2, we have that $\operatorname{Ann}(M_r)$ contains an element of the form $\delta_\tau+\pi$, $l(\pi)>\tau\triangleq\sum_i\tau_i$. Conversely, suppose that there exists $\delta_\tau+\alpha\in\operatorname{Ann}(M_r)$, $l(\alpha)>\tau$, and let $x\in M_r$. Then $x=\lambda(\sum_i\omega_ie_i)$, $\omega_i\in\Omega$. Let $s\geqq 0$ be an integer such that $(s+1)(\tau-l(\alpha))<l(\omega_i)$, $i=1,2,\cdots,m$. Then by Lemma 5.1., for each i, $\delta_{(s+1)\tau}*\omega_i+\pi_i\in\operatorname{Ann}(M_r)$, $l(\pi_i)>(s+1)\tau$, $\pi_i=(-1)^s\alpha^{s+1}*\omega_i$. Hence

$$\begin{aligned}
&(\delta_{(s+1)\tau}*\omega_i+\pi_i)q_i=0, && i=1,2,\cdots,m,\\
\Rightarrow&\sum_i(\delta_{(s+1)\tau}*\omega_i+\pi_i)q_i=0\\
\Rightarrow&\delta_{(s+1)\tau}x+\sum_i\pi_iq_i=0.
\end{aligned}$$

Since $l(\pi_i)>(s+1)\tau$, all i, the element $\sum_i\pi_ie_i\in\Omega^m$ is a control for x.

COROLLARY 5.4. *Every $x\in M_r$ is controllable if and only if each generator q_i is controllable.*

Examples for which the condition in Theorem 5.1. is satisfied will be given in § 7.

6. Bounded and minimal time controllability. Given an Ω-module homomorphism $\lambda:\Omega^m\to M$, the submodule $M_r=\lambda(\Omega^m)$ is said to be reachable (resp. controllable) in bounded time N if for each $x\in M_r$, there exists an $\omega\in\Omega^m$ with $l(\omega)>-N$, such that $x=\lambda(\omega)$ (resp. $\delta_{-N}x+\lambda(\omega)=0$). In the first part of this

section, we prove that if every element of M_r is controllable, then M_r is reachable and controllable in bounded time. Then we consider the determination of the smallest time period during which all the elements of M_r can be controlled. In the last part of the section, results are given on the smoothness of the controls constructed here.

Let A be an ideal of the ring Ω and let Ω/A denote the residue class ring of Ω by A. The elements of Ω/A will be denoted by $[\omega]=\omega+A$, $\omega\in\Omega$. Recall that Ω is a subring of $V=\mathscr{D}'_+$, the ring of distributions with support bounded on the left. We also note that for any $\gamma\in V$, $\varphi\in\mathscr{D}$, the multiplication $\gamma\varphi$ by φ is defined by $\langle\gamma\varphi, \chi\rangle=\langle\gamma, \varphi\chi\rangle$, where $\varphi\chi$ is the usual pointwise multiplication of functions.

LEMMA 6.1. *Let A be an ideal of Ω and suppose that there exists a $\beta\in A$ having an inverse $\beta^{-1}\in V$. Let $\tau<l(\beta)$. Then for each $[\omega]\in\Omega/A$, there exists an $\alpha\in[\omega]$ such that $l(\alpha)>\tau$.*

Proof. Assume that there exists $\beta\in A$ with $\beta^{-1}\in V$, and fix $\tau<l(\beta)$. Given $[\omega]\in\Omega/A$, if $l(\omega)>\tau$ there is nothing to prove. Therefore assume that $l(\omega)\leqq\tau$. Now $\beta*(\beta^{-1}*\omega)=\omega$, and thus $l(\omega)\geqq l(\beta)-l(\beta^{-1}*\omega)$. Since $l(\omega)\leqq\tau$ and $\tau<l(\beta)$,

$$l(\beta^{-1}*\omega)\leqq\tau-l(\beta)<0.$$

Choose a_1, a_2, b_1, $b_2\in\mathbb{R}$ such that $-\infty<a_2<a_1<l(\beta^{-1}*\omega)$ and $\tau-l(\beta)<b_1<b_2<0$. By a well-known result of distribution theory (see [6, p. 31]), there exists a $\varphi\in\mathscr{D}$ such that $\varphi(t)=1$ on $[a_1, b_1]$, $\varphi(t)=0$ on $\mathbb{R}-[a_2, b_2]$, and $0\leqq\varphi(t)\leqq 1$, all $t\in\mathbb{R}$. Then supp $[(\beta^{-1}*\omega)\varphi]\subseteq[l(\beta^{-1}*\omega), b_2]$, which implies that $(\beta^{-1}*\omega)\varphi\in\Omega$. Now define $\alpha=-\beta*[(\beta^{-1}*\omega)\varphi-\beta^{-1}*\omega]$. Then $\alpha=-\beta*[(\beta^{-1}*\omega)\varphi]+\omega\in[\omega]$. It is claimed that $l(\alpha)>\tau$. By construction, $(\beta^{-1}*\omega)\varphi=\beta^{-1}*\omega$ on $(-\infty, b_1)$, and thus supp $[-(\beta^{-1}*\omega)\varphi+\beta^{-1}*\omega]\subset[b_1, \infty)$. Then by definition of α, supp $\alpha\subseteq[b_1+l(\beta), 0]$. Therefore $l(\alpha)\geqq b_1+l(\beta)$, but by definition of b_1, $\tau<b_1+l(\beta)$, and hence $l(\alpha)>\tau$.

Using Lemma 6.1, we obtain the following sufficient condition for M_r to be reachable and controllable in bounded time.

THEOREM 6.1. *Given $M_r=\lambda(\Omega^m)$, if* Ann (M_r) *contains an element β having an inverse $\beta^{-1}\in V$, for any $a>0$ and $x_1, x_2\in M_r$, there exists a control $\omega\in\Omega^m$, with $l(\omega)>\tau\triangleq l(\beta)-a$ such that $\delta_\tau x_1+\lambda(\omega)=x_2$.*

Proof. Let β satisfy the hypothesis, fix $a>0$, and set $\tau=l(\beta)-a$. Then given $x_1=\sum_i\omega_i q_i$ and $x_2=\sum_i\pi_i q_i$, by Lemma 6.1 (taking $A=$ Ann (M_r)), for each i there exists an $\alpha_i\in(\delta_\tau*\omega_i-\pi_i)+$ Ann (M_r), with $l(\alpha_i)>\tau$. Hence $\delta_\tau x_1-x_2=\sum_i\alpha_i q_i$, which proves that x_2 can be set up from x_1 by control $-\sum\alpha_i e_i$.

COROLLARY 6.1. *If there exists a $\beta\in$ Ann (M_r) with $\beta^{-1}\in V$, then M_r is reachable and controllable in bounded time $-l(\beta)+a$, where a is an arbitrarily small positive number.*

Referring back to Theorem 5.1, we had that every element of M_r is controllable if and only if there exists $\delta_\tau+\alpha\in$ Ann (M_r) with $l(\alpha)>\tau$. As we shall see, this condition implies that Ann (M_r) contains a β with $\beta^{-1}\in V$, giving the following surprising result.

THEOREM 6.2. *M_r is reachable and controllable in bounded time if and only if every element of M_r is controllable.*

The proof of Theorem 6.2 follows from Lemma 6.2.

LEMMA 6.2. *Any element of the form* $\delta_\tau + \alpha \in \Omega$, $l(\alpha) > \tau$, *has an inverse in* V.

Proof. Given $\delta_\tau + \alpha \in \Omega$, $l(\alpha) > \tau$, consider $\delta_{-\tau}(\delta_\tau + \alpha) = \delta_0 + \delta_{-\tau}\alpha$, which is an element of V. It will be shown that $\delta_0 + \delta_{-\tau}\alpha$ has a (unique) inverse in V. Viewing $(\delta_0 + \delta_{-\tau}\alpha)^{-1}$ as an element in the quotient field of V, we can expand by long division giving

$$(\delta_0 + \delta_{-\tau}\alpha)^{-1} = \sum_{n=0}^{\infty} (-\delta_{-\tau}\alpha)^n. \tag{6.1}$$

Let $\{\gamma_i\}$ denote the sequence of partial sums obtained from the sum (6.1). Now since $l(\delta_{-\tau}\alpha) = a$, some $a > 0$, $l((\delta_{-\tau}\alpha)^n) \geqq na$. Then given $\varphi \in \mathscr{D}$, since φ has compact support there exist an integer i_0 and a constant K such that $\gamma_i(\varphi) = K$, all $i \geqq i_0$. Hence $\{\gamma_i(\varphi)\}$ converges in $\mathbb{R}$, proving that $\{\gamma_i\}$ converges in V. Therefore $(\delta_0 + \delta_{-\tau}\alpha)^{-1}$ is a distribution with support contained in $[0, \infty)$, and since $(\delta_\tau + \alpha)^{-1} = \delta_{-\tau}(\delta_0 + \delta_{-\tau}\alpha)^{-1}$, $\delta_\tau + \alpha$ has an inverse in V.

If M_r is controllable in bounded time, the question then arises as to what is the smallest time interval during which all the elements of M_r can be controlled. This minimal control time, denoted by $N_{\min}$, is defined to be the infinum over all N such that M_r is controllable in time N. We have the following results on the magnitude of $N_{\min}$.

Let $\ker \lambda$ denote the submodule $\{\omega \in \Omega^m : \lambda(\omega) = 0\} \subset \Omega^m$, and define

$$S_1 = \{\omega = (\omega_1, \cdots, \omega_m)^{\mathrm{TR}} \in \ker \lambda : \omega_i^{-1} \in V, \quad i = 1, 2, \cdots, m\},$$

$$S_2 = \{\omega = (\omega_1, \cdots, \omega_m)^{\mathrm{TR}} \in S_1 : \omega_i \in \operatorname{Ann}(q_i), \quad i = 1, 2, \cdots, m\}.$$

In terms of S_1 and S_2, we have the following bounds on $N_{\min}$.

THEOREM 6.3. *If* M_r *is controllable in bounded time, then*

$$\inf_{\omega \in S_1} \{-l(\omega)\} \leqq N_{\min} \leqq \inf_{\omega \in S_2} \{-l(\omega)\}.$$

Proof. If M_r is controllable in time N, for each $i = 1, 2, \cdots, m$, there exists $u_i \in \Omega^m$, $l(u_i) > -N$, such that $\delta_{-N}q_i + \lambda(u_i) = 0$. Thus $\delta_{-N}e_i + u_i \in \ker \lambda$, all i, which implies that $\sum_i (\delta_{-N}e_i + u_i) \in \ker \lambda$. Since $l(u_i) > -N$ all i, $\sum_i (\delta_{-N}e_i + u_i) = \sum_i (\delta_{-N} + \pi_i)e_i$ for some $\pi_i \in \Omega$ with $l(\pi_i) > -N$, all i. By Lemma 6.2, each $\delta_{-N} + \pi_i$ has an inverse in V, and thus $\sum_i (\delta_{-N}e_i + u_i) \in S_1$. Therefore $N_{\min} \geqq \inf_{\omega \in S_1}\{-l(\omega)\}$.

Now S_2 is not empty since $\operatorname{Ann}(M_r) \neq \{0\}$. Let $\omega \in S_2$. Then it follows from Lemma 6.1 that M_r is controllable in bounded time $-l(\omega) + a$, any $a > 0$. Hence $N_{\min} \leqq \inf_{\omega \in S_2}\{-l(\omega)\}$.

When $m = 1$, $\ker \lambda = \operatorname{Ann}(M_r)$ and $S_1 = S_2$, so we have the following.

COROLLARY 6.2. *If* $m = 1$ *and each* $x \in M_r$ *is controllable,* M_r *is controllable in minimal time* $N_{\min} = \inf\{-l(\pi) : \pi \in \operatorname{Ann}(M_r), \pi^{-1} \in V\}$.

In the next section, we shall use this result to compute minimal control times for delay-differential systems.

Given $\beta \in \operatorname{Ann}(M_r)$ with $\beta^{-1} \in V$, by Theorem 6.1. every $x \in M_r$ can be controlled in bounded time-τ for any $\tau < l(\beta)$. In particular, if $x = \sum \omega_i q_i$, by the construction given in the proof of Lemma 6.1, a control $u \in \Omega^m$ for x is

$$u = \sum_i u_i e_i, \quad \text{where } u_i = -\beta * [(\beta^{-1} * \delta_\tau * \omega_i)\varphi_i] + \delta_\tau * \omega_i, \qquad \tau < l(\beta). \tag{6.2}$$

However the control u may be so "rough" that it is not possible to generate an actual signal which is a good approximation to u. For example, this is the case if u contains derivatives of the Dirac distribution. Therefore some indication of the smoothness of the control (6.2) is very desirable. We now consider this by using the concept of the order of a distribution.

Let $\gamma \in \mathscr{D}'$ and let U be an open set contained in $\mathbb{R}$. The order of γ on U, denoted by ord $\gamma|_U$, is the smallest integer r such that $\gamma = h_\gamma^{(r)}$ on U, where $h_\gamma^{(r)}$ is the rth derivative of some continuous function h_γ on U. If no such positive integer r exists, γ is said to be of infinite order on U. If γ is infinitely differentiable on U, we write ord $\gamma|_U = -\infty$. The order of any distribution on a bounded set U is finite or $-\infty$ and so is the order of any distribution on $\mathbb{R}$ having compact support (see [6, p. 95]). It is easily verified that for any u, $v \in \mathscr{D}'$ having order $< +\infty$, ord $(u * v) \leqq (\text{ord } u) + (\text{ord } v)$.

Now given $x = \sum_i \omega_i q_i$, consider the control (6.2). We have the following upper bound on the order of the components of u.

THEOREM 6.4. ord $(u_i) \leqq \text{ord } \beta + \text{ord } (\omega_i) + \text{ord } \beta^{-1}|_{(0, -\tau - l(\omega_i))}$.

Proof. Given $u_i = -\beta * [(\beta^{-1} * \delta_\tau * \omega_i)\varphi_i] + \delta_\tau * \omega_i$, since supp $u_i \subset (\tau, 0)$ and supp $(\delta_\tau * \omega_i) \subseteq [\tau + l(\omega_i), \tau]$, $u_i = -\beta * [(\beta^{-1} * \delta_\tau * \omega_i)\varphi_i]$ on $(\tau, 0)$, $u_i = 0$, otherwise. Thus

$$\text{ord } (u_i) = \text{ord } \beta * [(\beta^{-1} * \delta_\tau * \omega_i)\varphi_i]|_{(\tau, 0)},$$

$$\text{ord } (u_i) \leqq \text{ord } \beta + \text{ord } [(\beta^{-1} * \delta_\tau * \omega_i)\varphi_i]|_{(\tau, 0)},$$

$$\text{ord } (u_i) \leqq \text{ord } \beta + \text{ord } (\beta^{-1} * \delta_\tau * \omega_i)|_{(\tau, 0)} \qquad \text{since } \varphi \in \mathscr{D},$$

$$\text{ord } (u_i) \leqq \text{ord } \beta + \text{ord } (\beta^{-1} * \omega_i)|_{(0, -\tau)},$$

$$\text{ord } (u_i) \leqq \text{ord } \beta + \text{ord } \omega_i + \text{ord } \beta^{-1}|_{(0, -\tau - l(\omega_i))} \qquad \text{since supp } \omega_i \subseteq [l(\omega_i), 0].$$

COROLLARY 6.3. *If β^{-1} is infinitely differentiable on $(0, \infty)$, every element of M_r has a control whose components are infinitely differentiable.*

Proof. In this case, ord $\beta^{-1}|_{(0, \infty)} = -\infty$, so that by the theorem, ord $(u_i) = -\infty$, implying that u_i is infinitely differentiable.

As will be seen in the next section, there exist controls that are infinitely differentiable when M_r is finite-dimensional as a linear space over $\mathbb{R}$.

7. Role of the impulse response matrix in controllability. The results of the preceding two sections reveal that the annihilating ideal Ann (M_r) plays a crucial role in the controllability of $M_r = \lambda(\Omega^m)$. Given a system $\Sigma = (X, \mu, \eta, \psi)$, for the special cases $\lambda = \mu$ and $\lambda = \eta\mu = f^*$, we now investigate the properties of Ann (M_r) by relating it to the impulse response matrix W of the system Σ. Here we obtain particular results on output function and state controllability, expressed in terms of the properties of W.

For the system $\Sigma = (X, \mu, \eta, \psi)$, let $X_r = \mu(\Omega^m)$ and $(\Gamma^k)_r = f^*(\Omega^m)$ denote the finitely-generated submodules of reachable states and reachable outputs, respectively. Letting $\{e_1, \cdots, e_m\}$ denote the standard basis of Ω^m as before, we have that X_r is generated by $g_i \triangleq \mu(e_i)$, $i = 1, 2, \cdots, m$, and $(\Gamma^k)_r$ is generated by $h_i \triangleq f^*(e_i)$, $i = 1, 2, \cdots, m$.

Since f^* is equal to the composition $\eta\mu$, $(\Gamma^k)_r = \eta(X_r)$, and thus $\text{Ann}\,(X_r) \subseteq \text{Ann}\,(\Gamma^k)_r$. However, in general, $\text{Ann}\,(X_r) \neq \text{Ann}\,(\Gamma^k)_r$. A necessary and sufficient condition for equality is given in the following.

PROPOSITION 7.1. $\text{Ann}\,(X_r) = \text{Ann}\,(\Gamma^k)_r$, *if and only if the restriction of* η *to the submodule* Ωg_i *is injective for each* i *such that* $g_i \neq 0$.

The proof of this result is straightforward, and therefore will not be given.

Recalling that the system Σ is completely observable if η is injective, we have the following.

COROLLARY. 7.1. *If* Σ *is completely observable*, $\text{Ann}\,(X_r) = \text{Ann}\,(\Gamma^k)_r$.

Now let $W = (w_{ij})$ denote the impulse response matrix of the system Σ. For each $i = 1, 2, \cdots, k$ and $j = 1, 2, \cdots, m$, define $A_{ij} = \{\pi \in \Omega : w_{ij} * \pi \in \Omega\}$. Each A_{ij} is an ideal of the ring Ω. In terms of the A_{ij}, the following result establishes a direct relationship between the impulse response matrix and the annihilating ideal of $(\Gamma^k)_r$.

PROPOSITION 7.2. $\text{Ann}\,(\Gamma^k)_r = \bigcap_{i,j} A_{ij}$.

Proof. Let $\pi \in \text{Ann}\,(\Gamma^k)_r$. Then $\pi h_j = 0$, $j = 1, 2, \cdots, m$, $h_j = f^*(e_j)$. Hence $(w_{ij} * \pi)|_{(0,\infty)} = 0$, $i = 1, 2, \cdots, k$, $j = 1, 2, \cdots, m$, which implies that $w_{ij} * \pi \in \Omega$, all i, j. Thus $\pi \in \bigcap_{i,j} A_{ij}$. Conversely, let $\pi \in \bigcap_{i,j} A_{ij}$. Then $(w_{ij} * \pi)|_{(0,\infty)} = 0$, all i, j. Thus $f^*(\pi e_j) = 0$, $j = 1, 2, \cdots, m \Rightarrow \pi h_j = 0$, all $j \Rightarrow \pi \in \text{Ann}\,(\Gamma^k)_r$.

From the results of § 6, the reachable states and outputs are controllable if and only if each ideal, $\text{Ann}\,(X_r)$ and $\text{Ann}\,(\Gamma^k)_r$, contains an element that is invertible in $V = \mathscr{D}'_+$. Since $\text{Ann}\,(X_r) \subseteq \text{Ann}\,(\Gamma^k)_r = \bigcap_{i,j} A_{ij}$, controllability is therefore connected to the existence of invertible elements in $\bigcap_{i,j} A_{ij}$, which we now consider. The approach given below is developed in terms of fields and rings of fractions.

Since V (resp. Ω) is an integral domain, the smallest field in which $V(\Omega)$ can be embedded is its quotient field, denoted by $Q(V)$ ($Q(\Omega)$). Let $\mathscr{J} : V \to Q(V) : v \mapsto v/\delta_0$ denote the embedding of V in $Q(V)$. Note that since $\Omega \subset V$, $Q(\Omega)$ is a subfield of $Q(V)$.

PROPOSITION 7.3. *Given a system* Σ *with impulse response matrix* W, $\bigcap_{i,j} A_{ij} \neq \{0\}$ *if and only if* $\mathscr{J}(w_{ij}) \in Q(\Omega)$, *all* i, j.

Proof. The proof is clear.

From Corollary 5.1, we have the following.

COROLLARY 7.2. *If any one of the elements of* W *cannot be embedded in* $Q(\Omega)$, *there exist at least one* $g_i \neq 0$ *and* $h_j \neq 0$ *such that every nonzero state in* Ωg_i *is uncontrollable and every nonzero output in* Ωh_j *is uncontrollable.*

As we now show, a condition for controllability is that the elements of W belong to a ring of fractions of Ω. Let $D = \{\pi \in \Omega : \pi^{-1} \in V\}$, which is clearly a multiplicative subset of the ring Ω. Let $D^{-1}\Omega$ denote the ring of fractions of Ω defined by D. Note that $D^{-1}\Omega$ can be viewed as a subring of V under the embedding $D^{-1}\Omega \to V : \omega/\pi \mapsto \pi^{-1} * \omega$. Then combining Theorems 6.1–6.2 and Propositions 7.1–7.2, we have the following.

THEOREM 7.1. *Given a system* $\Sigma = (X, \mu, \eta, \psi)$ *with impulse response matrix* W, *the following are equivalent*:

(i) $w_{ij} \in D^{-1}\Omega$, *all* i, j;

(ii) *every reachable output is controllable*;

(iii) *the submodule of reachable outputs is reachable and controllable in bounded time.*

Furthermore, for the reachable states to be controllable (or controllable in bounded time) it is necessary that one of these conditions be true.

THEOREM 7.2. *If the restriction of η on each nontrivial submodule Ωg_i is injective or if Σ is completely observable, the following are equivalent:*

(i) $w_{ij} \in D^{-1}\Omega$, *all* i, j;

(ii) *every reachable state is controllable;*

(iii) *the submodule of reachable states is reachable and controllable in bounded time.*

Some important consequences of Theorem 7.2 are given as follows.

COROLLARY 7.3. *If $\sum = (X, \mu, \eta, \psi)$ is completely reachable and observable, then X is completely controllable (or controllable in bounded time) if and only if $w_{ij} \in D^{-1}\Omega$, all i, j.*

COROLLARY 7.4. *A strictly causal i/o operator $f : V^m \to V^k : v \mapsto W * v$ has a canonical realization (X, μ, η, ψ) with X completely controllable if and only if W is over $D^{-1}\Omega$.*

These results show that controllability properties of the systems considered here are nice if the impulse response matrix is over $D^{-1}\Omega$. There exist systems for which this is not the case. For example, consider a single-input single-output system with impulse response $w(t) = e^{-t^2}H(t)$, where $H(t)$ is the Heaviside function. Because $e^{-(t-\tau)^2}$ contains the factor $e^{2\tau t}$, it follows that there does not exist a $\beta \in \Omega$ with $\beta^{-1} \in V$, such that $\beta * w \in \Omega$. The details are rather involved and will not be given.

Examples of classes of systems having impulse response matrix defined over $D^{-1}\Omega$ can be generated in the following manner. Let K be a multiplicative subset of Ω with $K \subseteq D$. Let $\mathcal{K}$ denote the class of all strictly causal systems (Definition 3.1) whose impulse response matrix is over $K^{-1}\Omega \subseteq D^{-1}\Omega$.

Example 1. Let $\mathbb{R}[p] = \{\sum_{i=0}^{n} a_i p^i : a_i \in \mathbb{R},\ n \geqq 0\}$, where p^i denotes the ith derivative of δ_0. Clearly, $\mathbb{R}[p]$ is a subring of Ω. Further, it is well known (see [6]) that every nonzero element of $\mathbb{R}[p]$ has an inverse in V which is infinitely differentiable on $(0, \infty)$. Thus we can take $K = \mathbb{R}[p] - \{0\}$. The resulting class $\mathcal{K}$ of systems includes all finite-dimensional systems. Let Σ be a system in this class. Since for any $\beta \in K$, $l(\beta) = 0$ and β^{-1} is infinitely differentiable on $(0, \infty)$, by Theorem 6.1 and Corollary 6.3 every reachable state and every reachable output of Σ can be controlled in an arbitrarily small time interval by a control whose components are infinitely differentiable.

Example 2. Let $\mathbb{R}[d_1, \cdots, d_r, p]$ denote the smallest subring of Ω containing $d_i = \delta_{a_i}$, $a_i < 0$, $p = \delta_0^{(1)}$, and $b\delta_0$, all $b \in \mathbb{R}$. Any $\beta \in \mathbb{R}[d_1, \cdots, d_r, p]$ can be written as a finite sum

$$\beta = \sum_{j_1, \cdots, j_{r+1}} c_{j_1, \cdots, j_{r+1}} d_1^{j_1} * \cdots * d_r^{j_r} * p^{j_{r+1}},$$

where the j_i are nonnegative integers, $c_{j_1, \cdots, j_{r+1}} \in \mathbb{R}$, and $d_i^j = j$th-fold convolution of d_i.

From the results of Kamen [3], every nonzero element of $\mathbb{R}[d_1, \cdots, d_r, p]$ has an inverse in V, so that we can take $K = \mathbb{R}[d_1, \cdots, d_r, p] - \{0\}$. In this case, the class $\mathcal{K}$ consists of systems having time delays equal to multiples of $-a_i$, $i = 1, 2, \cdots, r$

($d_i = \delta_{a_i}$). In particular, $\mathcal{K}$ contains a subclass of delay-differential systems, i.e., systems whose inputs and outputs are related by delay-differential equations. By Theorems 7.1–7.2 we have the interesting result that for systems with time delays (as defined here), the submodules of reachable states and outputs are reachable and controllable in bounded time.

For the case $K = \mathbb{R}[d_1, \cdots, d_r, p] - \{0\}$, in many instances we can readily compute the minimal control time $N_{\min}$: Let $\Sigma = (X, \mu, \eta, \psi)$ be a single-input single-output completely reachable and observable system belonging to the class $\mathcal{K}$. Then the impulse response w is given by $w = \beta^{-1} * \pi$, some $\beta \in K$, $\pi \in \Omega$, and $\operatorname{Ann}(X) = \{\alpha \in \Omega : w * \alpha \in \Omega\}$. Suppose that $\beta\Omega + \pi\Omega = \Omega$. Then we claim that $\operatorname{Ann}(X) = \beta\Omega$, that is, $\operatorname{Ann}(X)$ is a principal ideal. Let $\alpha \in \operatorname{Ann}(X)$. Then $\beta^{-1} * \pi * \alpha = \sigma$, some $\sigma \in \Omega$, $\Rightarrow \pi * \alpha = \beta * \sigma \Rightarrow \beta | \pi * \alpha$ (i.e., β divides $\pi * \alpha$ in Ω). Now $\beta * \gamma + \pi * \xi = \delta_0$ some $\gamma, \xi \in \Omega$, and since $\beta | \pi * \alpha * \xi$ and $\beta | \beta * \gamma * \alpha$, $\beta | (\beta * \gamma + \pi * \xi)\alpha \Rightarrow \beta | \alpha$. Hence $\operatorname{Ann}(X) \subset \beta\Omega$, while it is clear that $\beta\Omega \subset \operatorname{Ann}(X)$. Now by definition of K, $l(u * v) \leqq l(u)$ for any $u, v \in K$. Thus $l(\alpha) \leqq l(\beta)$ for any $\alpha \in \beta\Omega$, and by Corollary 6.2, $N_{\min} = -l(\beta)$. For example, let $K = \mathbb{R}[\delta_{-1}, p]$ and suppose that the impulse response of the system is

$$w(t) = \sum_{n=0}^{\infty} \frac{(n-t)^n}{n!} e^{-(t-n)} H(t-n).$$

Using the operational calculus given in [3], we have $(\delta_{-1} * p + \delta_{-1} + \delta_0) * w = \delta_{-1}$, so that we can take $\beta = \delta_{-1} * p + \delta_{-1} + \delta_0$ and $\pi = \delta_{-1}$. Since $\beta - (p + \delta_0) * \pi = \delta_0$, $\beta\Omega + \pi\Omega = \Omega$, and thus $N_{\min} = -l(\beta) = 1$. In words, every state of Σ can be controlled to zero within a minimal time period of one second.

8. Discussion of results. In this work, the internal (state) definition of a system is given in terms of a module framework that reflects the convolution module structure of the i/o representation. The module setup can be viewed as an extension of the usual $\mathbb{R}$-linear setting in the sense that the field of scalars ($\mathbb{R}$) of the latter is extended to a ring of convolution operators (Ω) in the former. (Recall that we have the embedding $\mathbb{R} \to \Omega : a \mapsto a\delta_0$.) In other words, with the module structure we can operate on states and input, output functions using convolution operators, rather than just elements of $\mathbb{R}$. The question immediately arises as to why this extension of the scalar multiplication is worth considering.

In the first place, important subspaces of the state space and output function space may be finitely generated as Ω-modules, but infinite-dimensional as $\mathbb{R}$-linear spaces. For example, the submodules of reachable states and output functions are always finitely generated (because the input function module is finitely generated). This finiteness can yield computational procedures involving operations in the convolution ring Ω (see, for example, the method of constructing controls given in the proofs of Theorems 5.1 and 6.1). Computable results can then be obtained since convolution operations in Ω can be performed by a variety of techniques (e.g., transform calculus), even though Ω is infinite-dimensional as a linear space over $\mathbb{R}$. For an illustration of this latter point, see the example given after Example 2, in which convolution operations are used to compute the minimal control time.

Another motivation for considering the convolution module structure is that certain dynamical properties are nicely characterizable in this framework. This is

illustrated by the results in §§ 5 and 6 connecting controllability to the properties of the annihilating ideal of the submodule of reachable elements. The practicability of these results stems primarily from the relationship (established in § 7) between the annihilating ideals of the submodules of reachable states and output functions and the properties of the impulse response matrix W. In particular, as a consequence of this relationship in § 7 a fairly computable criterion, given in terms of W, is obtained for determining when the reachable states and outputs are controllable to zero. The module-theoretic results of §§ 5 and 6 can also be used to obtain information on minimal control time.

In contrast to the algebraic techniques used here, existing results on the controllability of systems with infinite state space are obtained by using functional-analytical methods (see, for example, the papers by Falb [10] and Delfour–Mitter [11]). In addition, computable (algebraic) criteria have been obtained for various types of controllability of hereditary systems which evolve in n-dimensional space (e.g., delay-differential systems). Some of this work is referenced by Banks and Manitius in their survey paper [12].

It is clear that the module structure considered here arises as a consequence of the assumptions of linearity and time invariance. Hence it is not directly extendable to the time-varying case. However, as a result of recent work [13] showing that a modified version of Kalman's $K[z]$-module structure can be applied to time-varying systems, it now appears that module-theoretic techniques can be utilized to study infinite-dimensional time-varying systems. Although this may be true, what is of prime importance here is that the module approach can be used to "reduce" infinite-dimensional problems to computable finite-dimensional problems involving convolution operators.

REFERENCES

[1] R. KALMAN, P. FALB AND M. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[2] R. KALMAN AND M. HAUTUS, *Realization of continuous-time linear dynamical systems*, Proc. Conf. Diff. Equ., NRL Math. Research Center, 1971.

[3] E. KAMEN, *On an algebraic theory of systems defined by convolution operators*, Math. Systems Theory, 9 (1975), pp. 57–74.

[4] L. SCHWARTZ, *Theorie des Distributions*, Hermann, Paris, 1966.

[5] A. BENSOUSSAN AND E. KAMEN, *Continuous-time systems defined over a Banach convolution algebra*, IRIA Report INF 72 023, France, 1972.

[6] A. ZEMANIAN, *Distribution Theory and Transform Analysis*, McGraw-Hill, New York, 1965.

[7] R. SAEKS, *Causality in Hilbert space*, SIAM Rev., 12 (1970), pp. 357–383.

[8] F. TREVES, *Topological Vector Spaces, Distributions, and Kernels*, Academic Press, New York, 1967.

[9] S. LANG, *Algebra*, Addison-Wesley, Reading, Mass., 1967.

[10] P. FALB, *Infinite dimensional control problems I: On the closure of the set of attainable states for linear systems*, J. Math. Anal. Appl., 9 (1964), pp. 12–22.

[11] M. DELFOUR AND S. MITTER, *Controllability and observability for infinite-dimensional systems*, this Journal, 10 (1972), pp. 329–333.

[12] H. BANKS AND A. MANITIUS, *Application of abstract variational theory to hereditary systems – A survey*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 524–533.

[13] E. KAMEN, *A new algebraic approach to linear time-varying systems*, 1974, submitted for publication.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

NESTED DECOMPOSITION OF MULTISTAGE CONVEX PROGRAMS*

RICHARD P. O'NEILL†

Abstract. The multistage or staircase structure appears naturally in many models with time horizons. This paper presents and discusses a method for decomposition when the problem functions are convex. Among the techniques which can be used to solve the subproblems are the Dantzig–Wolfe convex programming algorithm and Bender's decomposition. Furthermore, when the nature of the problem presents certain structural forms, the decomposition allows for the introduction of more efficient techniques.

1. Introduction.

The Multistage Convex Program (MSCP) can be stated as follows:

$$\max \sum_{t=1}^{T} c_t(x_t),$$

$$A_1(x_1) \leqq 0,$$

$$B_{t-1}(x_{t-1}) + A_t(x_t) \leqq 0, \qquad t = 2, \cdots, T,$$

$$x_t \in S_t, \qquad t = 1, \cdots, T.$$

Dantzig and Wolfe [1] addressed the problem in linear form. Because of its appearance when written in a tableau form, it is also said to have staircase structure. Problems with multistage structure occur in optimal control, dynamic multisector economic models and various other engineering and business problems.

The problem in its linear form has also been studied by Dantzig and Wolfe [2], Glassey [9], and Ho and Manne [11]. In this paper, a decomposition structure for MSCP will be given creating a master problem and a sequence of problems which can be considered both subproblems and master problems. It will then be shown that the solution to the decomposed problem can be used to solve MSCP. Finally, some approaches to solving the subproblems will be given.

* Received by the editors July 2, 1974, and in revised form March 22, 1975.

† Department of Computer Science, Louisiana State University, Baton Rouge, Louisiana 70803.

2. The multistage convex program (MSCP). In tableau form, MSCP appears as

$$
\begin{array}{lllllll}
\max c_1(x_1)+c_2(x_2)+ & \cdots & +c_t(x_t)+ & \cdots & +c_T(x_T) & & \\
A_1(x_1) & & & & & \leqq 0, & \\
B_1(x_1)+A_2(x_2) & & & & & \leqq 0, & \\
\quad \ddots & \ddots & & & & \vdots & \\
& B_{t-1}(x_{t-1})+A_t(x_t) & & & & \leqq 0, & \\
& & B_t(x_t)+A_{t+1}(x_{t+1}) & & & \leqq 0, & \\
& & & \ddots & \ddots & \vdots & \\
& & & & B_{T-1}(x_{T-1})+A_T(x_T) & \leqq 0, &
\end{array}
$$

where $x_t \in S_t$ for $t = 1, 2, \cdots, T$, x_t is a vector with dimension n_t, and

$$
\begin{aligned}
&c_t : S_t \to R, && t = 1, \cdots, T,\\
&A_t : S_t \to R^{m_t}, && t = 1, \cdots, T,\\
&B_t : S_t \to R^{m_{t+1}}, && t = 1, \cdots, T-1.
\end{aligned}
$$

Define $S = S_1 \times \cdots \times S_T$, $x = (x_1, \cdots, x_T)$, and $c(x) = \sum_{t=1}^{T} c_t(x_t)$. Let c^* be the optimal value of MSCP and x^* be an optimal solution to MSCP. Also for convenience, define

$$B_0(\cdot) = B_T(\cdot) = 0.$$

It will be assumed that c_t is concave, that A_t and B_t are convex, and that all functions are continuous for $t = 1, \cdots, T$. Further, we will assume S_t is compact and convex.

3. The dual of MSCP. Throughout this paper, it will be assumed there is a point x^0 which satisfies the Slater condition. That is, x^0 is strictly interior to all inequality constraints.

By defining

$$L(x, u) = \sum_{t=1}^{T} \{c_t(x_t) - u_t[B_{t-1}(x_{t-1}) + A_t(x_t)]\}, \qquad u = (u_1, \cdots, u_T), \tag{1}$$

and

$$h(u) = \max_{x \in S} L(x, u), \tag{2}$$

the dual of MSCP can be defined as

$$\text{Dual} \qquad \min h(u), \qquad u \geqq 0. \tag{3}$$

The Slater condition implies the existence of an optimal solution to the dual with the optimal values of MSCP and the dual being equal (see Geoffrion [7, Chap. 2] or Mangasarian [12, Chap. 8]). It is a simple exercise to show that a feasible solution to the dual is an upper bound on the optimal value of MSCP. It should be observed that the dual is separable, in that the summation on t and the maximization can be interchanged.

4. The decomposition of MSCP. The decomposition creates T generalized linear programs coupled in one direction by the multipliers of the preceeding program and in the other direction by column generation (i.e., the columns are sent "up" to higher number programs and the multipliers are sent "down" to lower numbered programs). At each cycle (indexed by k) of the algorithm a sequence of T programs (denoted by SP_t^k for $t = 1, \cdots, T$) are solved. First, SP_T^k is solved for x_T^k, λ_T^k, π_T^k and σ_T^k. Then for $t = T-1, \cdots, 2$, SP_t^k using π_{t+1}^k is solved for x_t^k, λ_t^k, π_t^k and σ_t^k. Finally, SP_1^k using π_2^k is solved for x_1^k and π_1^k. Then P_t^{k+1} and Q_t^{k+1} for $t = 2, \cdots, T$ are defined by the inclusion of p_t^k and q_t^k, and the cycle is repeated.

The algorithm is initiated by setting $k = 1$ and

$$
\begin{aligned}
&p_2^0 = c_1(x_1^0),\\
&p_t^0 = c_{t-1}(x_{t-1}^0) + p_{t-1}^0, \qquad t = 3, \cdots, T\\
&q_t^0 = B_{t-1}(x_{t-1}^0), \qquad t = 2, \cdots, T.
\end{aligned}
$$

At the kth cycle the algorithm appears as

$$
\begin{aligned}
\mathrm{SP}_T^k: z_T^k = \max\ & c_T(x_T) + P_T^k \lambda_T,\\
& A_T(x_T) + Q_T^k \lambda_T \leqq 0, && \pi_T,\\
& e\lambda_T = 1, && \sigma_T,\\
& x_T \in S_T,\\
& \lambda_T \geqq 0.
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{SP}_t^k: z_t^k = \max\ & c_t(x_t) - \pi_{t+1}^k B_t(x_t) + P_t^k \lambda_t,\\
& A_t(x_t) + Q_t^k \lambda_t \leqq 0, && \pi_t,\\
& e\lambda_t = 1, && \sigma_t,\\
& x_t \in S_t,\\
& \lambda_t \geqq 0 \quad \text{for } t = T-1, \cdots, 2.
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{SP}_1^k: z_1^k = \max\ & c_1(x_1) - \pi_2^k B_1(x_1),\\
& A_1(x_1) \leqq 0, && \pi_1,\\
& x_1 \in S_1,
\end{aligned}
$$

where

$$P_t^k = (p_t^0, \cdots, p_t^{k-1}), \qquad t = 2, \cdots, T,$$

$$Q_t^k = (q_t^0, \cdots, q_t^{k-1}), \qquad t = 2, \cdots, T,$$

$$\lambda_t^k = (\lambda_t^{0,k}, \cdots, \lambda_t^{k-1,k})', \qquad t = 2, \cdots, T,$$

$$e = (1, \cdots, 1),$$

$$p_2^k = c_1(x_1^k) \quad \text{(scalar)},$$

$$p_t^k = c_{t-1}(x_{t-1}^k) + P_{t-1}^k \lambda_{t-1}^k \quad \text{(scalar)} \quad t = 3, \cdots, T,$$

$$q_t^k = B_{t-1}(x_{t-1}^k), \qquad t = 2, \cdots, T,$$

x_t^k, λ_t^k is an optimal solution to SP_t^k and

π_t^k and σ_t^k are the corresponding multipliers.

SP_T^k can be considered the master problem, but SP_t^k for $t = T-1, \cdots, 2$ can be considered both master problems (to the programs with a smaller t value) and subproblems (to the programs with a larger t value). SP_1^k is considered just a subproblem. For ease of reference, the term subproblem will refer to any SP_t^k for $t = 1, \cdots, T$. Each subproblem is smaller than MSCP in that it has fewer constraints and fewer nonlinear variables.

The assumption that x^0 satisfies the Slater condition guarantees a feasible solution for SP_t^1, $t = 1, \cdots, T$. The algorithm terminates if after a complete cycle, each subproblem remains unchanged (i.e., the next cycle would produce the same solution).

LEMMA 1. *If for any* t

$$c_t(x_t^k) - \pi_{t+1}^k B_t(x_t^k) + P_t^k \lambda_t^k > \sigma_{t+1}^k, \tag{4}$$

then the column generated by x_t^k *will enter the basis of* SP_{t+1}^k, *and, barring degeneracy, the optimal value of* SP_{t+1}^{k+1} *will increase.*

Proof. Rearranging terms in (4), we get

$$c_t(x_t^k) + P_t^k \lambda_t^k - \pi_{t+1}^k B_t(x_t^k) - \sigma_{t+1}^k > 0.$$

Since $\pi_{t+1}^k, \sigma_{t+1}^k$ are the multipliers for SP_{t+1}^k, (4) indicates the column $(p_t^k, q_t^k, 1)'$ will enter the basis of SP_{t+1}^{k+1} on the next cycle and in the absence of degeneracy, $\text{SP}_{t+1}^{k+1} > \text{SP}_{t+1}^k$ for $t = 1, \cdots, T-1$. □

It will now be shown that a feasible solution is available at each cycle. Define $\bar{x}^k = (\bar{x}_1^k, \cdots, \bar{x}_T^k)$ as

$$\bar{x}_T^k = x_T^k,$$

$$\mu_T^k = \lambda_T^k,$$

$$\bar{x}_t^k = X_t^k \mu_{t+1}^k, \qquad t = T-1, \cdots, 1,$$

$$\mu_{t+1}^k = \Lambda_t^k \cdots \Lambda_{T-1}^k \mu_T^k, \qquad t = T-2, \cdots, 1,$$

where $X_t^k = (x_t^0, \cdots, x_t^{k-1})$ dimensioned $n_t \times k$ and $\Lambda_t^k = (\lambda_t^1, \cdots, \lambda_t^k)$ dimensioned $k \times k$ (Λ_t^k is an upper triangular matrix).

LEMMA 2. *$\bar{x}^k$ is a feasible solution of* MSCP.

Proof.

$$A_T(\bar{x}_T^k) + \sum_{j=0}^{k-1} B_{T-1}(x_{T-1}^j)\lambda_T^{j,k} \leqq 0.$$

By convexity of B_{T-1},

$$A_T(\bar{x}_T^k) + B_{T-1}\left(\sum_{j=0}^{k-1} x_{T-1}^j \lambda_T^{j,k}\right) = A_T(\bar{x}_T^k) + B_{T-1}(\bar{x}_{T-1}^k) \leqq 0.$$

Therefore $\bar{x}_T^k$ and $\bar{x}_{T-1}^k$ satisfy the first set of constraints. Now,

$$A_{T-1}(x_{T-1}^j) + \sum_{l=0}^{j-1} B_{T-2}(x_{T-2}^l)\lambda_{T-1}^{l,j} \leqq 0.$$

By convexity of B_{T-2},

$$A_{T-1}(x_{T-1}^j) + B_{T-2}\left(\sum_{l=0}^{j-1} x_{T-2}^l \lambda_{T-1}^{l,j}\right) \leqq 0.$$

Therefore,

$$\sum_{j=0}^{k-1}\left[A_{T-1}(x_{T-1}^j) + B_{T-2}\left(\sum_{l=0}^{j-1} x_{T-2}^l \lambda_{T-1}^{l,j}\right)\right]\lambda_T^{j,k} \leqq 0.$$

By convexity of A_{T-1} and B_{T-2},

$$A_{T-1}\left(\sum_{j=0}^{k-1} x_{T-1}^j \lambda_T^{j,k}\right) + B_{T-2}\left(\sum_{j=0}^{k-1}\left(\sum_{l=0}^{j-1} x_{T-2}^l \lambda_{T-1}^{l,j}\right)\lambda_T^{j,k}\right)$$

$$= A_{T-1}(\bar{x}_{T-1}^k) + B_{T-2}(\bar{x}_{T-2}^k) \leqq 0.$$

Therefore, $\bar{x}_{T-1}^k$ and $\bar{x}_{T-2}^k$ satisfy the second set of constraints. Continuing recursively, we see that $\bar{x}_{T-2}^k, \cdots, \bar{x}_1^k$ satisfy the remaining constraints. Finally, since $\bar{x}_t^k$ is a convex combination of points in S_t, $\bar{x}_t^k \in S_t$. □

LEMMA 3. *At any cycle k the following inequalities hold*

$$z_T^k \leqq c(\bar{x}^k) \leqq c^* \leqq L(x^k, \pi^k). \tag{5}$$

Proof.

$$z_T^k = c_T(x_T^k) + P_T^k \lambda_T^k, \tag{6}$$

$$P_T^k \lambda_T^k = [c_{T-1}(x_{T-1}^0), \cdots, c_{T-1}(x_{T-1}^{k-1})]\lambda_T^k + (p_{T-1}^0, P_{T-1}^1 \lambda_{T-1}^1, \cdots, P_{T-1}^{k-1}\lambda_{T-1}^{k-1})\lambda_T^k. \tag{7}$$

From the first term on the right-hand side of (7) and since c_{T-1} is concave,

$$[c_{T-1}(x_{T-1}^0), \cdots, c_{T-1}(x_{T-1}^{k-1})]\lambda_T^k \leqq c_{T-1}(X_{T-1}^k \lambda_T^k) = c_{T-1}(\bar{x}_{T-1}^k). \tag{8}$$

Expanding the second term on the right-hand side of (7), we obtain

$$[c_{T-2}(x_{T-2}^0), \cdots, c_{T-2}(x_{T-2}^{k-1})]\Lambda_{T-1}^k \lambda_T^k + (p_{T-2}^0, P_{T-2}^1 \lambda_{T-2}^1, \cdots, P_{T-2}^{k-1}\lambda_{T-2}^{k-1}]\Lambda_{T-1}^k \lambda_T^k. \tag{9}$$

From the first term in (9) and since c_{T-2} is concave (note: $\Lambda_{T-1}^k \lambda_T^k = \mu_{T-1}^k$),

$$[c_{T-2}(x_{T-2}^0), \cdots, c_{T-2}(x_{T-2}^{k-1})]\mu_{T-1}^k \leqq c_{T-2}(X_{T-2}^k \mu_{T-1}^k) = c_{T-2}(\bar{x}_{T-2}^k) \tag{10}$$

Continuing in the same manner, we have

$$[c_t(x_t^0), \cdots, c_t(x_t^{k-1})]\mu_{t+1}^k \leqq c_t(X_t^k \mu_{t+1}^k) = c_t(\bar{x}_t^k) \quad \text{for } t = T-3, \cdots, 1. \tag{11}$$

The second inequality is established by observing that $\bar{x}^k$ is a feasible solution to MSCP, and the third inequality is established by observing that (x^k, π^k) is dual feasible. □

LEMMA 4. *For* $t = 2, \cdots, T$,

$$\sigma_t^k - \pi_t^k A_t(x_t^k) = P_t^k \lambda_t^k. \tag{12}$$

Proof. Consider the following linear program:

$$\begin{aligned} \max\ & P_t^k \lambda_t, \\ & Q_t^k \lambda_t \leqq -A_t(x_t^k), \\ & e\lambda_t = 1, \\ & \lambda_t \geqq 0, \end{aligned}$$

where x_t^k is fixed. λ_t^k is an optimal solution to the above program, with the dual variables, π_t^k and σ_t^k. By duality,

$$P_t^k \lambda_t^k = \sigma_t^k - \pi_t^k A_t(x_t^k). \qquad \square$$

THEOREM 1. *Any cluster point of* $\{\bar{x}^k\}$ *is an optimal solution to* MSCP.

Proof. S is compact since each S_t is compact; therefore $\{x^k\}$ has a cluster point. Since $\bar{x}^k \in S$, $\{\bar{x}^k\}$ must have a cluster point. From the continuity of problems functions and the Slater condition, $\{\pi_t^k\}$ and $\{\sigma_t^k\}$ for $t = 1, \cdots, T$ are contained in a compact set and therefore have a cluster point. Let x^∞, π^∞, σ^∞, $\bar{x}^\infty$ be a cluster point. For any $q \geqq k+1$, the following inequalities must hold:

$$c_t(x_t^k) - \pi_{t+1}^q B_t(x_t^k) + P_t^k \lambda_t^k \leqq \sigma_{t+1}^q \quad \text{for } t = 1, \cdots, T-1, \tag{13}$$

since they are the reduced costs for the columns of SP_{t-1}^q for $t = 1, \cdots, T-1$, respectively. Adding (12) and (13), observing that $\pi_1^k A_1(x_1^k) = 0$, and adding $c_T(x_t^k)$ to both sides of the inequality, one obtains

$$\sum_{t=1}^{T} [c_t(x_t^k) - \pi_t^k (B_{t-1}(x_{t-1}^k) - A_t(x_t^k))] \leqq c_T(x_T^k) + P_T^k \lambda_T^k + \sum_{t=2}^{T} (\sigma_t^q - \sigma_t^k). \tag{14}$$

Passing to the limit on k (subsequentially if necessary), the last term on the right-hand side of (14) vanishes, giving

$$\sum_{t=1}^{T} [c_t(x_t^\infty) - \pi_t^\infty (B_{t-1}(x_{t-1}^\infty) - A_t(x_t^\infty))] \leqq c_T(x_T^\infty) + P_T^\infty \lambda_T^\infty = z_T^\infty. \tag{15}$$

By passing to the limit on k in (5) we obtain

$$z_T^{\infty} \leqq \sum_{t=1}^{T} c_t(\bar{x}_t^{\infty}) \leqq \sum_{t=1}^{T} c_t(x_t^*) \leqq L(x^{\infty}, \pi^{\infty}). \tag{16}$$

Combining (15) and (16), it is observed that the right- and left-hand sides of the inequalities are equal, and therefore, $\bar{x}^{\infty}$ is optimal to MSCP. □

If the algorithm is terminated prematurely, a feasible solution is available, and a bound can be placed on its nearness to the optimal value.

THEOREM 2. *If the algorithm is terminated at cycle k, and if* (5) *does not hold as an equality, then* $\bar{x}^k$ *is feasible and within*

$$\beta = \min_{1 \leqq j \leqq k} L(x^j, \pi^j) - c(\bar{x}^k)$$

of the optimal value.

Proof. From Lemma 3, $\bar{x}^k$ is feasible and $c(\bar{x}^k) \leqq c^*$. Since (x^j, π^j) is dual feasible, $c^* \leqq L(x^j, \pi^j)$ for $j = 1, \cdots, k$. Therefore

$$0 \leqq c^* - c(\bar{x}^k) \leqq \min_{1 \leqq j \leqq k} L(x^j, \pi^j) - c(\bar{x}^k). \qquad \square$$

With an additional assumption, the calculation of $\bar{x}^k$ becomes unnecessary.

THEOREM 3. *If* $L(x, \pi^{\infty})$ *is strictly concave as a function of x in a neighborhood of* x^{∞}, *then* x^{∞} *is an optimal solution to* MSCP.

Proof. Since strictly concave functions have a unique optimal solution, the

$$\max L(x, \pi^{\infty}), \qquad x \in S,$$

has a unique optimal solution, x^{∞}. From Theorem 1,

$$c^* = L(x^{\infty}, \pi^{\infty}).$$

Since x^{∞} maximizes $L(x, \pi^{\infty})$,

$$L(x^*, \pi^{\infty}) \leqq L(x^{\infty}, \pi^{\infty}),$$

where x^* is an optimal solution to MSCP.

Since x^* is feasible,

$$B_{t-1}(x_{t-1}^*) + A_t(x_t^*) \leqq 0 \qquad \text{for } t = 1, \cdots, T.$$

Therefore,

$$L(x^*, \pi^{\infty}) = c^* - \sum_{t=1}^{T} \pi_t^{\infty}[B_{t-1}(x_{t-1}^*) + A_t(x_t^*)] \geqq L(x^{\infty}, \pi^{\infty}).$$

Therefore,

$$L(x^*, \pi^{\infty}) = L(x^{\infty}, \pi^{\infty}),$$

and since x^{∞} is unique, $x^{\infty} = x^*$. □

The assumption of strict concavity does not appear to be a serious restriction, since it is only necessary that $c(x)$ be strictly concave or that one constraint with a positive multiplier be strictly convex at x^{∞}.

5. Methods for solving subproblems. Although any method that produces optimal multipliers can be used to solve the subproblems, two methods that seem to be easily adapted to this situation will be discussed. The subscript t and superscript k in SP_t^k will be omitted in this section since they are not necessary for the discussion (i.e., SP_t^k becomes SP).

$$\text{SP} \qquad \begin{aligned} z = \max\ & f(x) + P\lambda, \\ & A(x) + Q\lambda \leqq 0, \\ & e\lambda = 1, \\ & x \in S, \qquad \lambda \geqq 0, \end{aligned}$$

where

$$f(x) = c(x) - \pi_{t+1} B(x).$$

A method for solving SP that appears to be natural since it also is a column generation method, is the Dantzig–Wolfe convex programming algorithm (see Dantzig [1, Chap. 24]). The subproblem becomes

$$\text{RMSP} \qquad \begin{aligned} \max\ & \sum_{j=0}^{l-1} f(x^j)\alpha_j + P^k\lambda, \\ & \sum_{j=0}^{l-1} A(x^j)\alpha_j + Q^k\lambda \leqq 0, \\ & e\lambda = 1, \\ & \sum_{j=0}^{l-1} \alpha_j = 1, \\ & \alpha_j \geqq 0, \qquad \lambda \geqq 0. \end{aligned}$$

Let α^l, λ^l be an optimal solution to RMSP. The subproblem for RMSP is

$$\text{SUBSP} \qquad \max f(x) - y^l A(x), \qquad x \in S,$$

where y^l are the dual variables from the lth cycle of RMSP and x^l is an optimal solution to SUBSP.

We must now show that this infinite process does not disturb the convergence demonstrated in Theorem 1.

Define

$$\bar{x}^l = \sum_{j=0}^{l-1} x_j \alpha_j^l.$$

THEOREM 4. *$\{\bar{x}^l, \lambda^l\}$ has a cluster point, $(\bar{\bar{x}}, \bar{\lambda})$, that is optimal to* SP, *and $\{y^l\}$ has a cluster point, $\bar{y}$, that is an optimal solution to the dual of* RMSP, *and no duality gap exists.*

Proof. By straightforward application of the Dantzig–Wolfe proof [1, Chap. 24], $(\bar{\bar{x}}, \bar{\lambda})$ is optimal to SP. To demonstrate that there is no duality gap and that $\{y^l\}$ has a cluster point which is optimal to the dual of SP it is sufficient to show that RMSP is a consistent program as defined by Duffin and Karlovitz [4] and has a finite value. By the method defined to generate columns for RMSP, the algorithm either terminates or the new column is pivoted into the basis and a feasible

solution is calculated; hence the program is consistent. The Slater condition is sufficient to bound the optimal value; this completes the proof. □

When this technique is employed to solve each subproblem, the decomposition creates $T-1$ linear programs and T nonlinear optimizations.

Another method for solving SP is Bender's decomposition. The presentation will closely follow that given by Geoffrion [7, p. 50]. SP can be rewritten as

$$\max_{x\in S}\{f(x)+\max_{\lambda\geqq 0}[P\lambda \text{ s.t. } Q\lambda \leqq -A(x),\ e\lambda = 1]\}. \tag{17}$$

In the inner maximization of (17), $-A(x)$ is the right-hand side of a linear program parametrized by x and the dual of this program can be written as

$$\text{BD} \qquad \min y'(-A(x))+y_{m+1}, \tag{18}$$

$$y'Q+y_{m+1}\geqq P, \tag{19}$$

$$y\geqq 0. \tag{20}$$

Letting $\begin{pmatrix} y^1 \\ y^1_{m+1} \end{pmatrix}, \cdots, \begin{pmatrix} y^q \\ y^q_{m+1} \end{pmatrix}$ be the extreme points of (19) and (20), equations (18), (19) and (20) can be rewritten as

$$\min_{j=1,\cdots,q}\{(y^j)'(-A(x))+y^j_{m+1}\}. \tag{21}$$

Therefore, (17) can be rewritten as

$$\begin{aligned} &\max_{x\in S} f(x)+z, \\ &(y^j)(-A(x)+y^j_{m+1})\geqq z, \qquad j=1,\cdots,q. \end{aligned} \tag{22}$$

Note (22) is a convex program since $y^j\geqq 0$. Instead of generating all y^j at once, an iterative scheme can be established in the following way.

Step 1. Solve (22) letting the optimal solution be $(\tilde{x}, \tilde{z})$.

Step 2. If $\tilde{x}$ is optimal to BD, $\tilde{x}$ along with the optimal dual variables to BD solve (17).

If $\tilde{x}$ is not optimal to BD, reoptimize and add the new extreme point as a constraint in (22). Go to Step 1.

Now if

$$f(\tilde{x})+z>\sigma_{t+1}, \tag{23}$$

the column added will enter the basis of the $(t+1)$st subprogram. Since theoretically BD can become a semi-infinite program, it must be shown that there is no duality gap. In a manner analogous to Theorem 4 it can be shown that BD is a regular program, which is all that is necessary.

6. Discussion. In the decomposition of MSCP, the direction in which the decomposition is started is arbitrary. That is, the master problem could start with $t=1$ and the last subproblem would have $t=T$. If this were the case the columns would be generated from the A functions instead of the B functions. However,

the order in which the master and subproblems are solved is important. If the master problem is not solved first with the subproblems following in order, an upper bound in the form of a dual feasible solution will not be available. If the subproblems were solved in numerical order (i.e., $t = 1, \cdots, T$), the solution would be found but no upper bound would be available.

Special structure in formulation of MSCP can be an advantage in the solution of the subproblems. For example if either the A or the B functions are linear, the subproblems can be formulated so that the explicit constraints are linear in each subproblem. Further, if either the A or B functions are linear and the others are quadratic, the subproblems can be formulated so that the subproblems are quadratic.

In each master problem it is usually necessary to generate an infinite number of columns to attain theoretical convergence. From a computational standpoint storage and execution time can be considerable. Fortunately, all but the basic columns of each subproblem may be dropped at each cycle if $z_t^k > z_t^{k-1}$ (see Eaves and Zangwill [5] and Murphy [13]). Fox [6] has suggested a hueristic acceleration device accomplished by introducing columns corresponding to $\bar{x}_t^k$ into the master programs. It requires relatively little computation and does not disturb the convergence properties as long as it is done only a finite number of times at each cycle.

The assumption of compactness of S can be relaxed to include the admission of extreme rays (see Dantzig and Van Slyke [3] and Geoffrion [8]).

7. Acknowledgment. The author would like to thank R. T. Rockafellar for his helpful suggestions and criticisms of an earlier version of the paper.

REFERENCES

[1] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, N.J., 1963.

[2] G. B. DANTZIG AND P. WOLFE, *Decomposition principal for linear programming*, Operations Res., 8 (1960), pp. 101–111.

[3] G. B. DANTZIG AND R. M. VAN SLYKE, *Generalized Linear Programming*, Optimization Methods for Large-Scale Systems, D. A. Wismer, ed., McGraw-Hill, New York, 1972, pp. 75–120.

[4] R. J. DUFFIN AND L. A. KARLOVITZ, *An infinite linear program with a duality gap*, Management Sci., 12 (1965), pp. 122–134.

[5] B. C. EAVES AND W. I. ZANGWILL, *Generalized cutting plane algorithms*, SIAM J. Control, 9 (1971), pp. 529–542.

[6] B. L. FOX, *Stability of the dual cutting-plane algorithm for concave programming*, Rep. RM-6147-PR, The RAND Corp., Santa Monica, Calif., 1970.

[7] A. M. GEOFFRION, *Perspectives on Optimization*, Addison-Wesley, Reading, Mass., 1972.

[8] ———, *Generalized Benders decomposition*, J. Optimization Theory Appl., 10 (1972), pp. 237–260.

[9] C. R. GLASSEY, *Nested decomposition and multi-stage linear programs*, Management Sci., 20 (1973), pp. 282–292.

[10] F. J. GOULD, *Extension of Lagrange multipliers in nonlinear programming*, SIAM J. Appl. Math., 17 (1969), pp. 1280–1297.

[11] J. K. HO AND A. S. MANNE, *Nested decomposition for dynamic models*, Math. Programming, 6 (1974), pp. 121–140.

[12] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[13] F. H. MURPHY, *Column dropping procedures for the generalized programming algorithm*, Management Sci., 19 (1973), pp. 1310–1320.

[14] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

LINEAR QUADRATIC OPTIMAL STOCHASTIC CONTROL WITH RANDOM COEFFICIENTS*

JEAN-MICHEL BISMUT†

Abstract. The purpose of this paper is to apply the methods developed in [1] and [2] to solve the problem of optimal stochastic control for a linear quadratic system.

After proving some preliminary existence results on stochastic differential equations, we show the existence of an optimal control.

The introduction of an adjoint variable enables us to derive extremality conditions: the control is thus obtained in random "feedback" form. By using a method close to the one used by Lions in [4] for the control of partial differential equations, a priori majorations are obtained.

A formal Riccati equation is then written down, and the existence of its solution is proved under rather general assumptions.

For a more detailed treatment of some examples, the reader is referred to [1].

Introduction. Let us consider the linear stochastic differential equation:

$$\begin{aligned} dx &= (Ax + Cu + f)\,dt + (Bx + Du + g)\cdot dw, \\ x(0) &= x, \end{aligned} \tag{1}$$

where w is an m-dimensional Brownian motion and where all the coefficients are supposed to be observable by the controller, who controls function u.

We want to minimize the criteria:

$$I(u) = E\left\{\int_0^T |M_t x_t|^2\,dt + \int_0^T \langle N_t u_t, u_t\rangle\,dt + |M_1 x_T|^2\right\}, \tag{2}$$

where N_t is a family of self-adjoint positive operators, and where again all the coefficients are observable by the controller.

This problem has a classical form. However, we allow in addition the coefficients in (1) and (2) to be random. Moreover, we accept that the noise term $(Bz + Du + g)\cdot dw$ depends *explicitly* on control u.

The purpose of this paper is to derive an existence result for an optimal control and to find the optimal control in a random feedback form.

This represents an important extension of the results given in [3] and [8]. The methods, however, are very different.

We will use functional analysis techniques to solve this problem. These methods are very similar to the methods already used by Lions [4, Chap. 3]. Lions was then solving the problem of the control of a linear parabolic partial differential equation with quadratic criterion. Although our problem is entirely different, the methods used to solve both problems will be identical; basically the functional analysis framework is the same in both cases.

In § 1, we define the notations used in the paper. We refer to various probability theory tools. The reader unfamiliar with martingale theory can suppose that all coefficients are deterministic, and that all martingales are

* Received by the editors December 4, 1972, and in revised form January 11, 1975.

† Marseille, France.

stochastic integrals relative to the Brownian motion, in order to have a simpler view of the functional analysis methods used here.

In § 2, we show that under very general conditions, (1) has a unique solution. Moreover, we prove an existence result for some backward stochastic differential equations.

In § 3, the problem is rigorously defined. The function $I(u)$ is proved to be convex and coercive on a given Hilbert space. An optimal control is then shown to exist.

The function I is proved to be differentiable. Condition

$$I'(u) = 0 \tag{3}$$

is then expressed through a dual variable p, which is the unique solution of a backward stochastic differential equation.

In § 4, the control is found in feedback form through the use of processes P_t and r_t, which are formal solutions of a system of stochastic differential equations. In particular, P_t solves an equation which extends the classical Riccati equation.

In § 5, the previous equations are shown to have unique solutions when the various coefficients appearing in (1) and (2) on one hand, and the Brownian motion on the other hand are assumed to be independent. A priori majorations found in the previous parts allow us to solve some deterministic differential equations with singular terms.

The method appears to be quite powerful and unifies the whole theory of linear quadratic problems in a very general framework. The usefulness of duality methods in optimal stochastic control, given for the first time in [2], is exhibited here quite clearly.

1. Notations.

$(\Omega, \mathcal{F}, P)$ is a complete probability space.

$\{\mathcal{F}_t\}_{t\in R^+}$ is an increasing sequence of complete sub σ-fields of $\mathcal{F}$ which has the following properties:

(a) It is right-continuous. [5, IV-30];

(b) It has no time of discontinuity [5, VII-D39].

This last assumption is not strictly necessary, but we make it to simplify the results.

$\mathcal{T}$ is the σ-field of $\Omega\times[0, +\infty[$ of the well-measurable sets [5, VIII-D14]. $\mathcal{T}^*$ is its completion for the measure $dP \otimes dt$.[1]

V is an n-dimensional vector space ($n \geqq 1$).

w is an m-dimensional Brownian motion on $(\Omega, \mathcal{F}, P)$, nonanticipating relative to $\{\mathcal{F}_t\}_{t\in R^+}$. w may be defined equivalently as a square-integrable a.s. continuous martingale on $(\Omega, \mathcal{F}, P)$ with values in R^m such that, by writing $w = (w_1, \cdots, w_m)$, one has (with the notations of [6])

$$d\langle w_i, w_j\rangle = \delta_{ij}\, dt. \tag{1.1}$$

This definition is correct by the result of P. Levy [6, p. 110]. Moreover, we extend the definitions for $m = 0$ by taking w as the one-dimensional null process.

---

[1] For our purpose $\mathcal{T}$ could simply have been the σ-field of nonanticipating sets.

Since w has continuous paths, (1.1) and the results of [6] show that it is possible to define unambiguously the stochastic integral of a $\mathcal{T}^*$ class of $\mathcal{T}$-measurable processes H such that for any t, one has

$$E\int_0^t |H_s|^2\,ds < +\infty. \tag{1.2}$$

For any stopping time σ, L_2^σ is the space of square-integrable $\mathcal{F}_\sigma$-measurable random variables, with values in V.

T is a strictly positive constant.

L_{21} is the space of the $dP \otimes dt$ classes[2] u of $\mathcal{T}^*$-measurable functions with values in V, such that

$$E\left(\int_0^T |u_t|\,dt\right)^2 < +\infty. \tag{1.3}$$

We define then a norm on L_{21} by

$$\|u\|_{21} = \left\{E\left(\int_0^T |u_t|\,dt\right)^2\right\}^{1/2}. \tag{1.4}$$

$L_{2\infty}$ is the space of the $dP \otimes dt$ classes x of $\mathcal{T}^*$-measurable functions with values in V such that

$$E(\sup_{0\leqq t\leqq T} \operatorname{ess} |x_t|^2) < +\infty. \tag{1.5}$$

We define then a norm on $L_{2\infty}$ by

$$\|x\|_{2\infty} = \{E(\sup_{0\leqq t\leqq T} \operatorname{ess} |x_t|^2)\}^{1/2}. \tag{1.6}$$

L_{22} is the space of the $dP \otimes dt$ classes H of $\mathcal{T}^*$-measurable functions with values in V^m such that

$$E\int_0^T |H_t|^2\,dt < +\infty. \tag{1.7}$$

We define then a norm on L_{22} by

$$\|H\|_{22} = \left(E\int_0^T |H_t|^2\,dt\right)^{1/2}. \tag{1.8}$$

By convention, we assume that the elements of L_{21}, $L_{2\infty}$ and L_{22} are equal to 0 for $t > T$.

Duality brackets are then defined:

(a) between L_2^σ and L_2^σ by the standard scalar product,

---

[2] Two $\mathcal{T}^*$-measurable functions u and u' are said to be *$dP \otimes dt$ equivalent* if they differ on a $dP \otimes dt$ negligible set. This defines an equivalence relation on $\mathcal{T}^*$-measurable functions, and then equivalence classes for this relation.

(b) between L_{21} and $L_{2\infty}$ by

$$E\int_0^T \langle u_t, y_t\rangle\, dt; \tag{1.9}$$

(c) between L_{22} and L_{22} by

$$E\int_0^T \langle H_t, H'_t\rangle\, dt. \tag{1.10}$$

$\underline{L}$ is the space of square-integrable martingales with values in V stopped at T, null at 0. $\underline{L}$ can be identified to a closed subspace of L_2^T, on which we put the induced topology.

W is the subspace of $\underline{L}$ generated by the stochastic integrals relative to w of elements of L_{22}. W is a stable space, in the sense of [6, no. 4, p. 80]. Let $W^\perp$ be the orthogonal of W in $\underline{L}$ in the sense of [6, Thm. 5, p. 81]. In particular, if $\{\mathcal{F}_t\}_{t\in R^+}$ is the family of σ-fields generated by w, a result of Ito [6, p. 135] shows that $W^\perp = \{0\}$.

PROPOSITION 1.1. *Let $(x_0, \dot{x}, H, M)$ and $(p_0, \dot{p}, H', M')$ be two elements of $L_2^0 \times L_{21} \times L_{22} \times W^\perp$. Then, if one defines the right-continuous processes x and p by*

$$\begin{aligned} x_t &= x_0 + \int_0^t \dot{x}_s\, ds + \int_0^t H_s \cdot dw_s + M_t, \\ p_t &= p_0 + \int_0^t \dot{p}_s\, ds + \int_0^t H'_s \cdot dw_s + M'_t, \end{aligned} \tag{1.11}$$

then the process N_t defined by

$$\begin{aligned} N_t = \langle p_t, x_t\rangle - \langle p_0, x_0\rangle - \int_0^t \langle \dot{p}_s, x_s\rangle\, ds - \int_0^t \langle p_s, \dot{x}_s\rangle\, ds \\ - \int_0^t \langle H_s, H'_s\rangle\, ds - \langle M_t, M'_t\rangle \end{aligned} \tag{1.12}$$

is a martingale, null at the origin.

Proof. This simple result is proved in [2]. □

2. Linear stochastic differential equations. We give here some results on linear stochastic differential equations. These results will be very useful in proving the existence of a dual variable in the problem of stochastic control which we solve in the remainder. We already know that in deterministic control, the dual variable is a solution of a deterministic backward equation. In stochastic control, it is then quite natural to think that the dual variable will be a solution of a stochastic differential equation with a stochastic terminal condition. We will prove that, under some simple assumptions, these backward equations have unique solutions.

Let A and $(B_i)_{i=1,\cdots,m}$ be a family of functions defined on $\Omega\times[0, +\infty[$ with values in $V \otimes V$ which are bounded and $\mathcal{T}^*$-measurable.

C_2^T is the space of right-continuous processes x adapted to $\{\mathcal{F}_t\}_{t\in R^+}$ and such that

$$E(\sup_{0\leqq t\leqq T}|x_t|^2)<+\infty. \tag{2.1}$$

A norm is defined on C_2^T by

$$\|x\|=\{E(\sup_{0\leqq t\leqq T}|x_t|^2)\}^{1/2}. \tag{2.2}$$

THEOREM 2.1. *For (Z_0, u, v, M) in $L_2^0\times L_{21}\times L_{22}\times W^{\perp}$, the equation*

$$\begin{aligned} dZ&=(AZ+u)\,dt+(BZ+v)\cdot dw+dM,\\ Z(0)&=Z_0, \end{aligned} \tag{2.3}$$

has one and only one solution with right-continuous paths. Moreover, these paths have no oscillatory discontinuities [5, IV-20].

Z is then in C_2^T, and the linear mapping defined on $L_2^0\times L_{21}\times L_{22}\times W^{\perp}$ with values in C_2^T by

$$(Z_0, u, v, M)\underset{f}{\to} Z$$

is continuous.

Proof. The proof is merely technical, and follows from a fixed point theorem. The proof is given in the Appendix. □

Let φ be the operator which associates to (Z_0, v, M) in $L_2^0\times L_{22}\times W^{\perp}$ Z_T in L_2^T through the equation

$$\begin{aligned} dZ&=AZ\,dt+(v+BZ)\cdot dw+dM,\\ Z(0)&=Z_0. \end{aligned} \tag{2.4}$$

Let ψ be the operator which associates to (p_0, v', M') in $L_2^0\times L_{22}\times W^{\perp}$ p_T in L_2^T through the equation

$$\begin{aligned} dp&=-(A^*p+B^*v')\,dt+v'\cdot dw+dM',\\ p(0)&=p_0. \end{aligned} \tag{2.5}$$

THEOREM 2.2. *φ and ψ are both continuous one-to-one operators, and one has*

$$\psi=(\varphi^*)^{-1}. \tag{2.6}$$

Proof. If v' is in L_{22}, B^*v' is in L_{22} and then in L_{21}. One then applies Proposition 1.1 to the processes Z and p defined in (2.4) and (2.5):

$$\begin{aligned} E\langle p_T, Z_T\rangle&=E\langle p_0, Z_0\rangle+E\int_0^T\langle p_t, A_tZ_t\rangle\,dt\\ &\quad+E\int_0^T\langle -A_t^*p_t-B_t^*v_t', Z_t\rangle\,dt\\ &\quad+E\int_0^T\langle v_t', v_t+B_tZ_t\rangle\,dt+E\langle M_T', M_T\rangle. \end{aligned} \tag{2.7}$$

Then (2.7) can be written

$$E\langle p_T, Z_T\rangle = E\langle p_0, Z_0\rangle + E\int_0^T \langle v'_t, v_t\rangle\, dt + E\langle M'_T, M_T\rangle. \tag{2.8}$$

From (2.8) one deduces necessarily that if $Z_T = 0$, then

$$(Z_0, v, M) = (0, 0, 0). \tag{2.9}$$

φ is then an injection.

Let us prove that for any Z_T in L_2^T, one can find (Z_0, v, M) such that

$$\varphi(Z_0, v, M) = Z_T. \tag{2.10}$$

We define $\tilde{Z}_t$ by

$$d\tilde{Z}_t = A\tilde{Z}\, dt, \qquad \tilde{Z}(T) = Z_T. \tag{2.11}$$

Since A is bounded, it is easily proved that

$$E(\sup_{0\leqq t\leqq T} |\tilde{Z}_t|^2) \leqq kE|Z_T|^2. \tag{2.12}$$

Let Z_t be the process $E^{\mathcal{F}_t}\tilde{Z}_t$. One then has

$$Z_t = E^{\mathcal{F}_t}\tilde{Z}_0 + E^{\mathcal{F}_t}\int_0^t A_s\tilde{Z}_s\, ds. \tag{2.13}$$

A is bounded and $\mathcal{T}^*$-measurable, and $\sup_{0\leqq t\leqq T}|\tilde{Z}_t|$ is integrable. Then, for $t' \geqq t$,

$$E^{\mathcal{F}_t}\int_0^{t'} A_s(\tilde{Z}_s - Z_s)\, ds = E^{\mathcal{F}_t}\int_0^t A_s(\tilde{Z}_s - Z_s)\, ds + E^{\mathcal{F}_t}\int_t^{t'} A_s(\tilde{Z}_s - Z_s)\, ds. \tag{2.14}$$

But for $s \geqq t$,

$$E^{\mathcal{F}_t}E^{\mathcal{F}_s}\tilde{Z}_s = E^{\mathcal{F}_t}\tilde{Z}_s. \tag{2.15}$$

From (2.14) we find that

$$E^{\mathcal{F}_t}\int_0^t A_s(\tilde{Z}_s - Z_s)\, ds$$

is a martingale.

$$Z_t - \int_0^t A_sZ_s\, ds$$

is then a martingale. Let us prove that this martingale is square-integrable, or equivalently that

$$E\left|Z_T - \int_0^T A_sZ_s\, ds\right|^2 < +\infty. \tag{2.16}$$

We know that $Z_T = \tilde{Z}_T$. Moreover,

$$E\left|\int_0^T A_s Z_s\, ds\right|^2 \leqq k' \int_0^T E|\tilde{Z}_s|^2\, ds, \tag{2.17}$$

and (2.17) may be written, using (2.12),

$$E\left|\int_0^T A_s Z_s\, ds\right|^2 < +\infty. \tag{2.18}$$

(2.16) is then proved.

By [6, Thm. 5, p. 81], one can find (Z_0, H, M) in $L_2^0 \times L_{22} \times W^{\perp}$ such that

$$Z_t = Z_0 + \int_0^t A_s Z_s\, ds + \int_0^t H_s \cdot dw_s + M_t. \tag{2.19}$$

This equality can be written

$$\begin{aligned} dZ &= AZ\, dt + H \cdot dw + dM, \\ Z(0) &= Z_0. \end{aligned} \tag{2.20}$$

Theorem 2.1 proves that

$$E(\sup_{o \leqq t \leqq T} |Z_t|^2) < +\infty. \tag{2.21}$$

Since B is bounded, BZ is in L_{22}. If we define v by

$$v = H - B\dot{Z}, \tag{2.22}$$

v is in L_{22}, and moreover, one has

$$\begin{aligned} dZ &= AZ\, dt + (v + BZ) \cdot dw + dM, \\ Z(0) &= Z_0, \end{aligned} \tag{2.23}$$

with $Z(T) = Z_T$.

This is equivalent to

$$\varphi(Z_0, v, M) = Z_T. \tag{2.24}$$

φ is then a continuous one-to-one operator. Since all the considered spaces are Banach spaces, φ has a continuous inverse.

But the relation (2.8) can be written:

$$\langle \psi(p_0, v', M'), Z_T \rangle = \langle (p_0, v', M'), \varphi^{-1}(Z_T) \rangle. \tag{2.25}$$

This proves necessarily that

$$\psi = \varphi^{*-1}. \tag{2.26}$$

□

3. The problem of control. In this section, we define in very general terms the problem of linear quadratic control, i.e., the problem of control of a linear stochastic differential equation with bounded and "observable" coefficients. Then by using the results established in § 2 on backward stochastic differential equations, we are able to prove the existence of an optimal control and to find necessary and sufficient conditions for a given control to be optimal.

H and U are two new finite-dimensional vector spaces.

∘ A, $(B_i)_{i=1,\cdots,m}$ is a family of functions defined on $\Omega\times[0,+\infty[$ with values in $V\otimes V$ which are bounded and $\mathcal{T}^*$-measurable.

∘ C, $(D_i)_{i=1,\cdots,m}$ is a family of functions defined on $\Omega\times[0,+\infty[$ with values in $U\otimes V$ which are bounded and $\mathcal{T}^*$-measurable.

∘ M is a function defined on $\Omega\times[0,+\infty[$ with values in $V\otimes H$ which is bounded and $\mathcal{T}^*$-measurable.

∘ N is a function defined on $\Omega\times[0,+\infty[$ with values in $U\otimes U$ which is bounded and $\mathcal{T}^*$-measurable and such that one can find a $\lambda>0$ for which one has: for any u in U,

$$\langle Nu,u\rangle\geqq\lambda|u|^2\,dP\otimes dt \quad \text{a.s.} \tag{3.1}$$

∘ M_1 is a function defined on Ω with values in $V\otimes H$, bounded and $\mathcal{F}_T$-measurable.

∘ f is an element of L_{21}.

∘ g is an element of L_{22}.

DEFINITION 3.1. L_{22}^U is the set of $dP\otimes dt$ classes of functions u defined on $\Omega\times[0,T]$ with values in U which are $\mathcal{T}^*$-measurable, and are such that

$$E\int_0^T|u_t|^2\,dt<+\infty.$$

A norm is defined on L_{22}^U by

$$\|u\|=\left\{E\int_0^T|u_t|^2\,dt\right\}^{1/2}.$$

L_{22}^U is then a Hilbert space.

DEFINITION 3.2. The *problem of linear quadratic control* (LQC) consists in the minimization of the criteria defined on L_{22}^U by

$$u\underset{I}{\to}E\left\{\int_0^T|M_tx_t|^2\,dt+\int_0^T\langle N_tu_t,\,u_t\rangle\,dt\right\}+E|M_1x_T|^2, \tag{3.2}$$

x being given by

$$\begin{aligned}dx&=(Ax+Cu+f)\,dt+(Bx+Du+g)\cdot dw,\\ x(0)&=x_0,\end{aligned} \tag{3.3}$$

with x_0 in L_2^0.

THEOREM 3.1. *The problem LQC has one unique solution.*

Proof. The result is proved according to classical methods. By Theorem 2.1, the mapping $u\to x$ is affine and continuous from L_{22}^U in C_2^T (here x is a.s. continuous). This proves easily that I is continuous and convex.

Moreover, when $\|u\|\to+\infty$, $I(u)\to+\infty$ by (3.1). L_{22}^U is a Hilbert space. This implies that when α is large enough, $\{u;\,I(u)\leqq\alpha\}$ is convex and weakly compact. I then has an optimum. Since I is strictly convex, this optimum is unique. □

We are now going to write the condition $I'(u)=0$ with the use of a new process p.

THEOREM 3.2. *A necessary and sufficient condition for u to be optimal is: if p is the unique solution of*

$$\begin{aligned} dp &= (M^*Mx - A^*p - B^*H)\,dt + H\cdot dw + dM, \\ p_T &= -M_1^*M_1x_T, \end{aligned} \tag{3.4}$$

with (p_0, H, M) *in* $L_2^0 \times L_{22} \times W^\perp$, *then*

$$Nu = C^*p + D^*H. \tag{3.5}$$

Proof. The proof can be done very rapidly by using the general duality results of [2]. We give here a direct proof.

It is easily shown that I is differentiable. Since I is convex, u is a solution to the problem LQC iff $I'(u) = 0$.

One has

$$\begin{aligned} \langle I'(u), v-u\rangle = 2\Big\{ & E\int_0^T \langle M_t^*M_tx_t^u, x_t^v - x_t^u\rangle\,dt + E\int_0^T \langle N_tu_t, v_t - u_t\rangle\,dt \\ & + E\langle M_1^*M_1x_T^u, x_T^v - x_T^u\rangle\Big\}. \end{aligned} \tag{3.6}$$

Let us prove that the system:

$$\begin{aligned} dp &= (M^*Mx^u - A^*p - B^*H)\,dt + H\cdot dw + dM, \\ p_T &= -M_1^*M_1x_T^u, \end{aligned} \tag{3.7}$$

has a unique solution with (p_0, H, M) in $L_2^0 \times L_{22} \times W^\perp$.

Let q be the unique solution of

$$\begin{aligned} dq &= (M^*Mx^u - A^*q)\,dt, \\ q_0 &= 0. \end{aligned} \tag{3.8}$$

Theorem 2.1 shows that q_T is in L_2^T, because x^u is in C_2^T. It is then equivalent to prove that the system

$$\begin{aligned} dq' &= (-A^*q' - B^*H)\,dt + H\cdot dw + dM, \\ q'_T &= -M_1^*M_1x_T^u - q_T, \end{aligned} \tag{3.9}$$

has a unique solution, with (q'_0, H, M) in $L_2^0 \times L_{22} \times W^\perp$. But Theorem 2.2 says precisely that (3.9) has a unique solution.

By applying Proposition 1.1, one has

$$\begin{aligned} E\langle -M_1^*M_1x_T^u, x_T^v - x_T^u\rangle = & E\int_0^T \langle M^*Mx^u - A^*p - B^*H, x_t^v - x_t^u\rangle\,dt \\ & + E\int_0^T \langle p_t, A_t(x_t^v - x_t^u) + C_t(v_t - u_t)\rangle\,dt \\ & + E\int_0^T \langle H_t, B_t(x_t^v - x_t^u) + D_t(v_t - u_t)\rangle\,dt. \end{aligned} \tag{3.10}$$

(3.10) may be written

$$E\int_0^T \langle M_t^* M_t x_t^u,\, x_t^v - x_t^u\rangle\, dt + E\langle M_1^* M_1 x_T^u,\, x_T^v - x_T^u\rangle$$

$$= -E\int_0^T \langle C_t^* p_t + D_t^* H_t,\, v_t - u_t\rangle\, dt.$$

From (3.6) and (3.10), the relation $I'(u) = 0$ is equivalent to

$$Nu = C^* p + D^* H. \tag{3.11}$$

□

4. The "feedback" problem. The purpose of this part is to find the dual variable in feedback form. The method is very similar to the method used by Lions [4, Chap. 3]. Practically, we solve the problem LQC, but instead of starting at time 0, we start at any time s ($0 \leqq s \leqq T$). We then use a priori majorations derived in the Appendix to write the dual variable p in stochastic feedback form. Some of the proofs proceed exactly as in [4, Chap. 3]. To avoid unnecessary repetitions, we refer to this work when necessary.

PROPOSITION 4.1. *For any s in $[0, T]$ and h in L_2^s, the system*

$$\begin{aligned} d\varphi &= (A\varphi + CN^{-1}C^*\psi + CN^{-1}D^*\chi + f)\, dt \\ &\quad + (B\varphi + DN^{-1}C^*\psi + DN^{-1}D^*\chi + g)\cdot dw, \\ d\psi &= (M^*M\varphi - A^*\psi - B^*\chi)\, dt + \chi \cdot dw + dM, \\ &\qquad \text{with } (\chi, M) \text{ in } L_{22}\times W^{\perp}; \end{aligned} \tag{4.1}$$

$$\begin{aligned} \varphi(s) &= h, \\ \psi(T) &= -M_1^* M_1 \varphi(T), \end{aligned} \tag{4.1'}$$

has a unique solution.

Proof. By using the methods of [4, Chap. 3, Lemma 4.1] and Theorem 3.1. It is easily proved that φ and ψ are respectively the optimal state variable and the dual variable of the problem LQC starting at time s with the "value" h.

PROPOSITION 4.2. *The mapping $h \to \{\varphi, \psi\}$ defined in Proposition 4.1 is continuous and affine from L_2^s into $C_2^T \times C_2^T$.*

Proof. The proof proceeds exactly as the proof of [4, Chap. 3, Lemma 4.2]. One proves that the given mapping is continuous from L_2^s in $C_2^T \times C_2^T$, this last space having its weak topology. All the spaces considered being Banach spaces, the closed graph theorem proves that the affine mapping which is considered is necessarily continuous from L_2^s into $C_2^T \times C_2^T$. □

COROLLARY. *The mapping $h \to \psi(s)$ is continuous and affine from L_2^s into L_2^s.*

Proof. Since $h \to \{\varphi, \psi\}$ is continuous from L_2^s into $C_2^T \times C_2^T$, and $\{\varphi, \psi\} \to \psi(s)$ is continuous from $C_2^T \times C_2^T$ into L_2^s, the result is proved. □

PROPOSITION 4.3. *One can find P_s and r_s which are $\mathcal{F}_s$-measurable such that:*

$$\begin{aligned} &\text{(a)}\ P_s(\omega) \in \mathcal{L}(V, V), \\ &\text{(b)}\ r_s(\omega) \in V, \\ &\text{(c)}\ \psi_s = -(P_s h + r_s). \end{aligned} \tag{4.2}$$

P_s and r_s are then determined in a unique way. Moreover, P_s is essentially bounded and r_s is in L_2^s. $P_s h$ is determined by the solution of (4.1) *with f and g null, and r_s is determined by the solution of* (4.1) *with h null.*

Proof. By the existence and uniqueness of the solution of (4.1), one checks immediately that if A is $\mathcal{F}_s$-measurable, and if h and h' are two elements of L_2^s, then

$$\{\varphi, \psi\}(1_A h + 1_{CA} h') = 1_A\{\varphi, \psi\}(h) + 1_{CA}\{\varphi, \psi\}(h'). \tag{4.3}$$

We consider then two cases.

Case 1. $f = 0$, $g = 0$. If $\mathbf{e}_1, \cdots, \mathbf{e}_n$ is a basis of V, the continuity of the mapping $h \to \psi_s$ proves that

$$A \xrightarrow[\pi_{ij}]{} E(\langle \psi_s(1_A \mathbf{e}_i), \mathbf{e}_j \rangle) \tag{4.4}$$

defines an additive measure on $(\Omega, \mathcal{F}_s, P)$ which is absolutely continous with respect to P. By the theorem of Radon–Nikodym, one can find p_{ij} which is $\mathcal{F}_s$-measurable and integrable such that

$$\pi_{ij}(A) = -\int P_{ij}\, dP. \tag{4.5}$$

Let $P_s(\omega)$ be the operator defined by $(P_{ij}(\omega))$. If h is a step function which is $\mathcal{F}_s$-measurable, the relation (4.5) proves that

$$\psi_s(h) = -P_s h. \tag{4.6}$$

Moreover, the mapping $h \to \psi_s(h)$ being continuous, one can find a $k > 0$ such that

$$E|P_s h|^2 \leqq k^2 E|h|^2. \tag{4.7}$$

The mapping $h \to P_s h$ can then be extended in a unique way to a continuous mapping from L_2^s into itself, because the step functions are dense in L_2^s. One deduces that for any h in L_2^s,

$$\psi_s(h) = -P_s h. \tag{4.8}$$

Moreover, the relation (4.7) proves that P_s is essentially bounded.

Case 2. In the general case, $\psi_s(h) + P_s h$ is a random variable r_s which does not depend on h. Since r_s is equal to $\psi_s(0)$, r_s is square-integrable. □

PROPOSITION 4.4. *P_s is a.s. a self-adjoint positive operator. One can find a $C_2 > 0$ such that for any s, and for any h of V,*

$$|P_s h| \leqq C_2 |h| \quad a.s. \tag{4.9}$$

Proof. Let h and h' be two elements of L_2^s. Let (φ, ψ) (resp. (φ', ψ')) be the solutions of (4.1) for $\varphi_s = h$ (resp. $\varphi'_s = h'$). u_t (resp. u'_t) is the corresponding control. f and g are supposed to be null.

Let $F_s(h, h')$ be the expression defined by

$$F_s(h, h') = E\int_s^T \langle M_t \varphi_t, M_t \varphi'_t \rangle\, dt + E\int_s^T \langle N_t u_t, u'_t \rangle\, dt + E\langle M_1 \varphi_T, M_1 \varphi'_T \rangle. \tag{4.10}$$

$F_s(h, h')$ is symmetric in (h, h'). By the same technique already used in the

previous sections, and by a method comparable to the method used in [4, Chap. 3, Lemma 4.4], one proves that

$$F_s(h, h') = E\langle P_s h, h'\rangle. \tag{4.11}$$

Equation (4.11) shows that P_s is a.s. self-adjoint. Moreover, $F_s(h, h') \geqq 0$. Then P_s is a.s. positive. But the expression of $F_s(h, h)$ is precisely the minimal value of the criteria for the problem LQC starting at s from h.

If x_h is the solution of

$$\begin{aligned} dx_h &= Ax_h\, dt + Bx_h \cdot dw, \\ x_h(s) &= h, \end{aligned} \tag{4.12}$$

one has

$$F_s(h, h) \leqq E\int_s^T |Mx_{h_t}|^2\, dt + E|M_1 x_{h_T}|^2. \tag{4.13}$$

It is proved in the Appendix (Remark A.1) that the mapping

$$h \to_{\tau_s} x_h$$

is continuous from L_2^s into C_2^T and, moreover, one can find a $C_0 > 0$ such that for any s in $[0, T]$,

$$\|\tau_s\| \leqq C_0. \tag{4.14}$$

From (4.11), (4.13) and (4.14), one deduces that one can find $C_1 > 0$ such that for any s in $[0, T]$ and for any h in L_2^s, one has

$$E\langle P_s h, h\rangle \leqq C_1 E|h|^2. \tag{4.15}$$

Since P_s is self-adjoint and positive, one deduces that one can find C_2 such that

$$E|P_s h|^2 \leqq C_2 E|h|^2. \tag{4.16}$$

This implies that $\sup \operatorname{ess} |P_s(\cdot)|$ is bounded by a constant independent of s. □

THEOREM 4.1. *The solution p of the system* (3.4) *is such that for any s, one has*

$$p_s = -(P_s x_s + r_s) \quad a.s. \tag{4.17}$$

Proof. This is obvious from the previous results. □

Remark. The previous theorem says nothing on the trajectories of P.

5. The Riccati equation: A formal approach. A natural idea is to write formally that P_t can be decomposed in the following way:

$$P_t = P_0 + \int_0^t \dot{P}_s\, ds + \int_0^t \mathcal{H}_s \cdot dw_s + \mathcal{M}_t, \tag{5.1}$$

with $\mathcal{H} = (\mathcal{H}_1, \cdots, \mathcal{H}_m)$ being such that

$$E\int_0^T |\mathcal{H}_s|^2\, ds < +\infty$$

and $\mathcal{M}$ being in $W^{\perp}$.

The purpose of this section is to find the formal stochastic differential equation, whose solution is precisely P. P will be proved to satisfy a generalized Riccati equation. The same method is used for r.

PROPOSITION 5.1. *The formal Riccati equation determining P is*

$$\begin{aligned} &dP+\{PA+A^*P+B^*PB+B^*\mathcal{H}+\mathcal{H}B-(B^*PD+PC+\mathcal{H}D)\\ &(N+D^*PD)^{-1}(D^*PB+C^*P+D^*\mathcal{H})+M^*M\}\,dt-\mathcal{H}\cdot dw-d\mathcal{M}=0,\\ &P_T=M_1^*M_1, \end{aligned} \tag{5.2}$$

where $\mathcal{H}=(\mathcal{H}_1,\cdots,\mathcal{H}_m)$ is a family of self-adjoint operators depending on (ω, t) and $\mathcal{T}^*$-measurable, where $\mathcal{M}$ is a martingale of self-adjoint operators belonging to $W^\perp$, and with the conventions

$$B^*PB=\sum_{i=1}^m B_i^*PB_i, \qquad B\mathcal{H}=\sum_{i=1}^m B_i\mathcal{H}_i$$

and the corresponding conventions for all the other terms.

Proof. In (5.1), if we write that P is self-adjoint, necessarily $\dot{P}$, $\mathcal{H}$ and $\mathcal{M}$ are self-adjoint, by the uniqueness of the decomposition (5.1).

If we consider the system (3.3), (3.4), with f and g null, we have

$$\begin{aligned} dx&=(Ax+Cu)\,dt+(Bx+Du)\cdot dw,\\ x(0)&=x_0,\\ dp&=(M^*Mx-B^*H-A^*p)\,dt+H\cdot dw+dM,\\ p_T&=-M_1^*M_1x_T,\\ Nu&=C^*p+D^*H. \end{aligned} \tag{5.3}$$

But by Theorem 4.1, one has, for any s,

$$p_s=-P_sx_s. \tag{5.4}$$

Moreover, if we assume that P can be written in the form (5.1), (5.4) implies that the right-continuous processes p_s and $-P_sx_s$ are equal. If we replace p_s by $-P_sx_s$ in (5.3), one gets

$$\begin{aligned} -\{dPx+P\,dx+\mathcal{H}\cdot(Bx+Du)\,dt\}&\\ =(M^*Mx-B^*H+A^*Px)\,dt+H\cdot dw+dM.& \end{aligned} \tag{5.5}$$

(5.5) can be written

$$\begin{aligned} -\{\dot{P}x+P(Ax+Cu)+\mathcal{H}\cdot(Bx+Du)\}&=M^*Mx-B^*H+A^*Px,\\ -\{\mathcal{H}_ix+P(B_ix+D_iu)\}&=H_i, \qquad i=1,\cdots,m. \end{aligned} \tag{5.6}$$

From (5.3) and (5.6), one gets

$$(N+D^*PD)u=-(D^*PB+C^*P+D^*\mathcal{H})x. \tag{5.7}$$

Since P is positive, $N+D^*PD$ is positive definite. One can write, from (5.7),

$$u=-(N+D^*PD)^{-1}(D^*PB+C^*P+D^*\mathcal{H})x.$$

By replacing u and H by their values, one gets

$$\begin{aligned}(5.8)\qquad &\{\dot P + PA + A^*P + B^*PB + B^*\mathscr{H} + \mathscr{H}B - (B^*PD + PC + \mathscr{H}D)\\ &(N + D^*PD)^{-1}(D^*PB + C^*P + D^*\mathscr{H}) + M^*M\}x = 0.\end{aligned}$$

Since the above holds for any x, (5.2) follows. □

Let us now find the formal equation for r.

PROPOSITION 5.2. *r is the formal solution of*

$$\begin{aligned}(5.9)\qquad dr &= \{(PC + B^*PD + \mathscr{H}D)(N + D^*PD)^{-1}C^* - A^*\}r\,dt\\ &\quad + [\{(PC + B^*PD + \mathscr{H}D)(N + D^*PD)^{-1}D^* - B^*\}\\ &\quad (Pg + h) - Pf - \mathscr{H}g]\,dt + h \cdot dw + dM',\\ r_T &= 0,\end{aligned}$$

with (h, M') in $L_{22} \times W^{\perp}$.

Proof. We write in the same way

$$(5.10)\qquad r_t = r_0 + \int_0^t \dot r_s\,ds + \int_0^t h_s \cdot dw_s + \int_0^t dM',$$

with (h, M') in $L_{22} \times W^{\perp}$.

We now take the complete system (3.3), (3.4), and we know here that

$$(5.11)\qquad p_s = -(P_s x_s + r_s).$$

One then gets

$$\begin{aligned}(5.12)\qquad &-\{\dot P x + P(Ax + Cu + f) + \mathscr{H} \cdot (Bx + Du + g) + \dot r\}\\ &\qquad = M^*Mx - B^*H + A^*Px + A^*r,\\ &-\{\mathscr{H}_i x + P(B_i x + D_i u + g_i) + h_i\} = H_i,\\ &\quad dM' + dMx = -dM.\end{aligned}$$

One then has for u,

$$(5.13)\qquad u = -(N + D^*PD)^{-1}\{C^*r + (C^*P + D^*PB + D^*\mathscr{H})x + D^*Pg + D^*h\}.$$

One then gets

$$\begin{aligned}(5.14)\qquad \dot r &= \{(PC + B^*PD + \mathscr{H}D)(N + D^*PD)^{-1}C^* - A^*\}r\\ &\quad + \{(PC + B^*PD + \mathscr{H}D)(N + D^*PD)^{-1}D^* - B^*\}(Pg + h) - Pf - \mathscr{H}g.\end{aligned}$$ □

COROLLARY. *The formal expression of the optimal control is*

$$(5.15)\qquad u = -(N + D^*PD)^{-1}\{(C^*P + D^*PB + D^*\mathscr{H})x + C^*r + D^*(Pg + h)\}.$$

6. The Riccati equation: Existence of the solution. We have no proof of existence and uniqueness of the solution of equation (5.2) in the general case. We will prove existence and uniqueness in a particular case, which applies especially when the coefficients of the equation, the coefficients of the criteria on the one hand and the Brownian motion on the other hand, are independent.

THEOREM 6.1. *The Riccati equation*

$$\begin{aligned} &dP+\{PA+A^*P+B^*PB-(B^*PD+PC)(N+D^*PD)^{-1}\\ &\qquad (D^*PB+C^*P)+M^*M\}\,dt-d\mathcal{M}=0,\\ &P_T=M_1^*M_1, \end{aligned} \tag{6.1}$$

where $\mathcal{M}$ is a square-integrable martingale of linear operators, has a unique solution in the space of adapted a.s. right-continuous processes $\tilde{P}_T$ with values in $\mathcal{L}(V, V)$ such that one can find $C'>0$ with

$$\begin{aligned} &\sup_{0\leqq t\leqq T}\operatorname{ess\,sup}|\tilde{P}_t|\leqq C',\\ &\sup_{(\omega,t)}\operatorname{ess}\|(N+D^*PD)^{-1}\|\leqq C'. \end{aligned} \tag{6.2}$$

$\mathcal{M}$ is then a martingale of self-adjoint operators, and P is a process of self-adjoint positive operators.

Proof. For P in $\mathcal{L}(V, V)$, let $\varphi_t(P)$ be formally defined by:

$$\begin{aligned} \varphi_t(P)=-\{&PA_t+A_t^*P+B_t^*PB_t-(B_t^*PD_t+PC_t)\\ &(N_t+D_t^*PD_t)^{-1}(D_t^*PB_t+C_t^*P)+M_t^*M_t\} \end{aligned} \tag{6.3}$$

We want to solve the equation

$$\begin{aligned} &dP=\varphi_t(P_t)\,dt+d\mathcal{M},\\ &P_T=M_1^*M_1. \end{aligned} \tag{6.4}$$

Let P' be a self-adjoint positive operator. Then if P is a linear operator such that

$$\|P-P'\|\leqq\frac{\lambda}{2\operatorname{sup\,ess}\|D\|^2}, \tag{6.5}$$

$N+D^*PD$ has $dP\otimes dt$ a.s. an inverse, and moreover:

$$\|(N+D^*PD)^{-1}\|\leqq\frac{2}{\lambda}. \tag{6.6}$$

To prove the first part of this assertion, since $N+D^*P'D$ has an inverse $dP\otimes dt$ a.s., one needs only to prove that, under (6.5),

$$\|D^*(P-P')D\|<\|(N+D^*P'D)^{-1}\|^{-1}. \tag{6.7}$$

But one has necessarily

$$N+D^*P'D\geqq N. \tag{6.8}$$

Then

$$\|(N+D^*P'D)^{-1}\|\leqq\|N^{-1}\|. \tag{6.9}$$

From (6.9) one gets

$$\|(N+D^*P'D)^{-1}\|^{-1} \geqq \|N^{-1}\|^{-1} \geqq \lambda. \tag{6.10}$$

If P satisfies (6.5), then

$$\|D^*(P-P')D\| \leqq \lambda/2, \tag{6.11}$$

and λ being strictly positive,

$$\lambda/2 < \lambda. \tag{6.12}$$

(6.7) follows from (6.11), (6.12) and (6.10). Then necessarily:

$$\begin{aligned} \|(N+D^*PD)^{-1}\| &\leqq \|(N+D^*P'D)^{-1}\|/(1 \\ &\quad -\|D^*(P-P')D\|\|(N+D^*P'D)^{-1}\|) \\ &\leqq 2/\lambda. \end{aligned} \tag{6.13}$$

(6.6) is also proved.

We notice that the different majorations are related only to the fact the P' is self-adjoint and positive.

Let R be defined by

$$R = \frac{\lambda}{2 \operatorname{sup\,ess} \|D\|^2}. \tag{6.14}$$

For $\alpha > 0$, and for $\mathscr{P}$ a function which is $\mathscr{F}_t$-measurable and a.s. bounded with values in $\mathscr{L}(V, V)$, let $\varkappa_{\mathscr{P}}^{\alpha}$ be the set of right-continuous adapted processes defined on $[T-\alpha, T]$, with values in $\mathscr{L}(V, V)$ such that

$$\|P_t - E^{\mathscr{F}_t}\mathscr{P}\| \leqq R. \tag{6.15}$$

We put on $\varkappa_{\mathscr{P}}^{\alpha}$ the distance defined by:

$$d(P, P') = \operatorname{sup\,ess} \sup_{T-\alpha \leqq t \leqq T} \|P'_t - P_t\|. \tag{6.16}$$

Then if $\mathscr{P}$ has self-adjoint positive values, the relations (6.5), (6.14) and (6.6) prove that one can find a positive finite number $M(\mathscr{P})$ such that, if P is in $\varkappa_{\mathscr{P}}^{\alpha}$, then

$$\|\varphi_t(P_t)\| \leqq M(\mathscr{P})\, dP \otimes dt \quad \text{a.s.} \tag{6.17}$$

Moreover, if C_2 is the constant defined in (4.9), the same relation will prove that

$$\sup_{\substack{\mathscr{P} \text{ self-adjoint} \geqq 0 \\ \|\mathscr{P}\| \leqq C_2}} M(\mathscr{P}) < M < +\infty. \tag{6.18}$$

In the same way, the calculation of the derivative in P of φ_t will prove that one can find $k > 0$ such that if P and P' are in $\varkappa_{\mathscr{P}}^{\alpha}$, with $\mathscr{P}$ self-adjoint, positive, and with $\|\mathscr{P}\| \leqq C_2$, then

$$\|\varphi_t(P_t) - \varphi_t(P'_t)\| \leqq k\|P_t - P'_t\|\, dP \otimes dt \quad \text{a.s.} \tag{6.19}$$

Let us notice finally that $\varkappa_{\mathscr{P}}^{\alpha}$ is a metrizable complete space. We take here

$$\mathscr{P} = M_1^* M_1. \tag{6.20}$$

From (4.9), one has necessarily

$$\|M_1^* M_1\| \leqq C_2. \tag{6.21}$$

We take for α the value:

$$\alpha = R/M. \tag{6.22}$$

Let G be the mapping which to $\tilde{P}$ in $\varkappa_{\mathscr{P}}^{\alpha}$ associates $\tilde{Q}$ via

$$\tilde{Q}_t = E^{\mathscr{F}_t}\Big(\mathscr{P} - \int_t^T \varphi_s(\tilde{P}_s)\, ds\Big). \tag{6.23}$$

Then we prove that Q is in $\varkappa_{\mathscr{P}}^{\alpha}$. $\tilde{Q}_t$ is a right-continuous process because one can write

$$\tilde{Q}_t = E^{\mathscr{F}_t}\Big(\mathscr{P} - \int_{T-\alpha}^T \varphi_s(\tilde{P}_s)\, as\Big) + \int_{T-\alpha}^t \varphi_s(\tilde{P}_s)\, ds, \tag{6.24}$$

$$\operatorname{sup\,ess} \sup_{T-\alpha \leqq t \leqq T} \|\tilde{Q}_t - E^{\mathscr{F}_t}\mathscr{P}\| \leqq \frac{R}{M} M(\mathscr{P}) \leqq R. \tag{6.25}$$

For $\tilde{P}$ and $\tilde{P}'$ in $\varkappa_{\mathscr{P}}^{\alpha}$, let us calculate

$$G(\tilde{P}) - G(\tilde{P}'). \tag{6.26}$$

One has

$$(G(\tilde{P}) - G(\tilde{P}'))_t = E^{\mathscr{F}_t}\Big(\int_t^T \varphi_s(\tilde{P}'_s)\, ds - \int_t^T \varphi_s(P_s)\, ds\Big). \tag{6.27}$$

Then, by (6.19), one has

$$\|(G(\tilde{P}) - G(\tilde{P}'))_t\| \leqq E^{\mathscr{F}_t} \int_t^T k\|\tilde{P}_s - \tilde{P}'_s\|\, ds. \tag{6.28}$$

From (6.28), one deduces

$$\|(G^2(\tilde{P}) - G^2(\tilde{P}'))_t\| \leqq k^2 E^{\mathscr{F}_t} \int_t^T ds E^{\mathscr{F}_s} \int_s^T \|\tilde{P}_u - \tilde{P}'_u\|\, du. \tag{6.29}$$

But (6.29) can be written

$$\|(G^2(\tilde{P}) - G^2(\tilde{P}'))_t\| \leqq k^2 E^{\mathscr{F}_t} \int_t^T ds \int_s^T \|\tilde{P}_s - \tilde{P}'_s\|\, ds. \tag{6.30}$$

From (6.30), one deduces

$$d(G^2(\tilde{P}), G^2(\tilde{P}')) \leqq \frac{k^2\alpha^2}{2}\, d(\tilde{P}, \tilde{P}'). \tag{6.31}$$

In the same way, one will have

$$d(G^n(\tilde{P}), G^n(\tilde{P}')) \leqq \frac{k^n \alpha^n}{n!} d(\tilde{P}, \tilde{P}'). \tag{6.32}$$

For n large enough,

$$k_n \alpha^n / n! < 1. \tag{6.33}$$

From [7, II, § 12, Remark 2], one deduces that G has a unique fixed point in the metrizable complete space $\varkappa_{\mathscr{P}}^{\alpha}$, which we call P.

Let then $(\tilde{\Omega}, \tilde{\mathscr{F}}_{t'} \tilde{P})$ be the space of continuous functions defined on $[0, +\infty[$ with values in R^m on which one has put the Brownian measure $\tilde{P}$ relative to a m-dimensional Brownian motion w starting from 0 at time 0.

Let $(\Omega', \mathscr{F}'_{t'} P')$ be the probability space

$$(\Omega \times \tilde{\Omega}, \mathscr{F} \otimes \tilde{\mathscr{F}}_{t-T+\alpha}, P \otimes \tilde{P}), \tag{6.34}$$

with $t \geqq T - \alpha$.

Then on this space, the process $\mathscr{M}$ defined for $t \geqq T - \alpha$ by

$$\mathscr{M}_t = E^{\mathscr{F}_t}\left[M_1^* M_1 - \int_{T-\alpha}^{T} \varphi_s(P_s)\, ds \right] \tag{6.35}$$

is a martingale relative to $\{\mathscr{F}_t \otimes \tilde{\mathscr{F}}_{t-T+\alpha}\}_{t \geqq T-\alpha}$; this follows from the independence of $\mathscr{F}_t$ and $\tilde{\mathscr{F}}_{t-T+\alpha}$. Moreover, $(\mathscr{M}_t - \mathscr{M}_{T-\alpha})$ is a martingale which is orthogonal to w, because $\mathscr{M}_t - \mathscr{M}_{T-\alpha}$ and w are independent variables. P is then a solution on $[T-\alpha, T]$ of

$$\begin{aligned} dP &= \varphi_s(P_s)\, ds + d\mathscr{M}, \\ P_T &= M_1^* M_1, \end{aligned} \tag{6.36}$$

and $\mathscr{M}_t - \mathscr{M}_{T-\alpha}$ is in $W^{\perp}$.

If we come back to the problem of control, we check now that P_t is precisely the operator defined in Proposition 4.3.

To prove this property, we need only to prove that for any s in $[T-\alpha, T]$ and any h in L_2^s, if x is a solution of

$$\begin{aligned} dx &= \{A - C(N + D^*PD)^{-1}(C^*P + D^*PB)\}x\, dt \\ &\quad + \{B - D(N + D^*PD)^{-1}(C^*P + D^*PB)\}x\, dw, \\ x_s &= h, \end{aligned} \tag{6.37}$$

then $(x, -Px)$ is the solution of the system (4.1) with f and g null.

We prove first that (6.37) has a solution. This is obvious, because all the linear operators appearing in (6.37) are bounded (this follows in particular from (6.6)). One then applies Theorem 2.1.

By doing the same calculations as are in the proof of Proposition 5.1, we find easily that $(x, -Px)$ corresponds to (φ, ψ) in (4.1) with

$$\begin{aligned} \chi &= -P\{B - D(N + D^*PD)^{-1}(D^*PB + C^*P)\}x, \\ M_t &= -\int_{T-\alpha}^{t} \langle d\mathscr{M}_s, x_s \rangle, \end{aligned} \tag{6.38}$$

x being C_2^T, χ is necessarily in L_{22}. Moreover, by Proposition 1.1, one has

$$\left\{-\int_{t-\alpha}^{t} \chi_s \cdot dw_s + \int_{T-\alpha}^{t} \langle d\mathcal{M}_s, x_s\rangle\right\}$$

$$= P_t x_t - P_{T-\alpha} x_{T-\alpha} - \int_{T-\alpha}^{t} \varphi_s(P_s) x_s\, ds - \int_{T-\alpha}^{t} P_s \dot{x}_s\, ds, \tag{6.39}$$

Then, as P is bounded, one has

$$E(\sup_{T-\alpha \leqq t \leqq T} |P_t x_t\|^2) < +\infty,$$

$$E\left(\sup_{T-\alpha \leqq t \leqq T} \left|\int_{T-\alpha}^{t} \varphi_s(P_s) x_s\, ds\right|^2\right) \leqq kE(\sup_{T-\alpha \leqq s \leqq T} |x_s|^2) < +\infty, \tag{6.40}$$

$$E\left(\sup_{T-\alpha \leqq t \leqq T} \left|\int_{T-\alpha}^{t} P_s \dot{x}_s\, ds\right|^2\right) \leqq k' E(\sup_{T-\alpha \leqq s \leqq T} |x_s|^2) < +\infty.$$

(6.40) proves that the local martingale M is a square-integrable martingale. $(x, -Px)$ is then the unique solution of the system (4.1), (4, 1'). One then applies Proposition 4.4: a.s., for any s in $[T-\alpha, T]$, P_s is self-adjoint positive, and

$$|P_s| \leqq C_2. \tag{6.41}$$

In particular,

$$|P_{T-\alpha}| \leqq C_2. \tag{6.42}$$

One can then start again the procedure from time $T-\alpha$, and in a finite number of steps reach 0.

Uniqueness is easily proved under the given assumption: if P' is a second solution of (6.1) on $[0, T)$, right-continuous and bounded, one will have

$$P'_t = E^{\mathcal{F}_t}\left(M_1^* M_1 - \int_t^T \varphi_s(P'_s)\, ds\right), \tag{6.43}$$

with $\varphi_s(P_s)$ uniformly bounded by a constant M'. Then if $t \geqq T - R/M'$, one has

$$|P'_t - E^{\mathcal{F}_t} M_1^* M_1| \leqq R. \tag{6.44}$$

We define then α' by

$$\alpha' = \alpha \wedge \frac{R}{M'}.$$

P' is necessarily a fixed point of G on $\varkappa^{\alpha'}_{M_1^* M_1}$. P' is then equal to P on $[T-\alpha', T]$, because G has a unique fixed point. One iterates the procedure a finite number of steps to reach 0. □

Remark 1. The general equation given in (5.2) can not be solved by this method. We were able to solve it, in the case when $\mathcal{H}$ can be taken to be 0 and where the martingale $\mathcal{M}$ is necessarily orthogonal to w. In the general case, we

could try to use the following technique: we define a mapping $(\tilde{P}, \tilde{\mathscr{H}}) \to (\tilde{Q}, \tilde{\mathscr{H}}')$ by

$$\tilde{Q}_t = E^{\mathscr{F}_t}\Big(\tilde{P} - \int_t^T \varphi_s(\tilde{P}_s, \tilde{\mathscr{H}}_s)\, ds\Big),$$

where $\varphi_s(\tilde{P}_s, \tilde{\mathscr{H}}_s)$ is the natural extension of $\varphi_s(P_s)$. $\tilde{Q}_t$ can then be written as

$$\tilde{Q}_t = \tilde{Q}_s + \int_s^t \varphi_s(\tilde{P}_s, \tilde{\mathscr{H}}_s)\, ds + \int_s^t \tilde{\mathscr{H}}'_s \cdot dw + \tilde{\mathscr{M}}_t.$$

Unfortunately, fixed point techniques are difficult to use in this case, because, apparently, $\tilde{\mathscr{H}}$ cannot be taken to vary in a sufficiently regular space.

Remark 2. In the case where all the coefficients are deterministic, the restriction on the boundedness of P and of $\varphi(P)$ is unnecessary to prove the uniqueness of the solution. To prove this point, one needs only to see that—P'_t converging necessarily to $M_1^* M_1$ when $t \to T$, for t close enough to T—one has

$$\|P'_t - M_1^* M_1\| \leqq R.$$

The deterministic differential equation defined by (6.1) with $\mathscr{M} = 0$ then has a unique solution.

THEOREM 6.2. *Under the assumption of Theorem* 6.1, *the equation*

$$\begin{aligned} dr = {} & \{(PC + B^*PD)(N + D^*PD)^{-1}C^* - A^*\} r\, dt \\ & + [\{(PC + B^*PD)(N + D^*PD)^{-1}D^* - B^*\}(Pg + h) - Pf]\, dt \\ & + h \cdot dw + dM', \\ r_T = {} & 0, \end{aligned} \tag{6.45}$$

has a unique solution with (h, M') *in* $L_{22} \times W^{\perp}$.

Proof. One must solve an equation of the type

$$\begin{aligned} dr = {} & -(\mathscr{A}^* r + \mathscr{B}^* h + \varphi_1\, dt) + (h + \varphi_2) \cdot dw + dM', \\ r_T = {} & 0, \end{aligned} \tag{6.46}$$

with $\mathscr{A}$ and $\mathscr{B}$ bounded and (φ_1, φ_2) in $L_{21} \times L_{22}$. Let r_1 be the solution of

$$\begin{aligned} dr_1 = {} & -(\mathscr{A}_1^* r_1 + \varphi_1)\, dt + \varphi_2 \cdot dw, \\ r_1(0) = {} & 0. \end{aligned} \tag{6.47}$$

By Theorem 1.1, $r_1(T)$ is in L_2^T. One needs to find the solution of

$$\begin{aligned} dr_2 = {} & -(\mathscr{A}^* r_2 + \mathscr{B}^* h)\, dt + h \cdot dw + dM', \\ r_2(T) = {} & -r_1(T). \end{aligned} \tag{6.48}$$

One applies Theorem 2.2. □

Example 1. $\mathscr{U} = U_1 \times U_2$. All the operators are supposed to be constant. w is 1-dimensional. We suppose

$$C = (C, 0), \quad D = (0, D), \quad N = \begin{pmatrix} N_1 & 0 \\ 0 & N_2 \end{pmatrix}.$$

The Riccati equation is then

$$\frac{dP}{dt}+PA+A^*P+B^*PB-PCN_1^{-1}C^*P-B^*PD$$

$$(N_2+D^*PD)^{-1}D^*PB+M^*M=0, \tag{6.49}$$

$$P_T=M_1^*M_1.$$

The optimal control is given by

$$\begin{aligned} u_1&=-N_1^{-1}C^*Px,\\ u_2&=-(N_2+D^*PD)^{-1}D^*PBx. \end{aligned} \tag{6.50}$$

Example 2. We consider the equation

$$dx=(Ax+Cu)\,dt+Bx\,dw_1+Du\,dw_2,$$

$$x(0)=x_0,$$

with the criterion

$$E\int_0^T |M_tx_t|^2\,dt+E\int_0^T \langle N_tu_t,u_t\rangle\,dt+E|M_1x_1|^2. \tag{6.51}$$

We suppose that the operators are constant. Then P is a solution of

$$\begin{aligned} &\frac{dP}{dt}+PA+A^*P+B^*PB-PC(N+D^*PD)^{-1}C^*P+M^*M=0,\\ &P_T=M_1^*M_1. \end{aligned} \tag{6.52}$$

u is given by

$$u=-(N+D^*PD)^{-1}C^*Px. \tag{6.53}$$

These formulas are the same as the ones given by Wonham [8].

Example 3. We take the general case with

$$f=0,$$

$$B_i \text{ or } D_i\neq 0\Rightarrow g_i=0.$$

Then the solution of (6.45) is $r=0$. The "random" feedback has no "constant" term.

Example 4. Let (A_1, C_1, M_1, N_1) and (A_2, C_2, M_2, N_2) be two families of constant operators having the properties given in § 3.

Let γ be a positive random variable defined on R^+ having density $\lambda e^{-\lambda t}\,dt$.

Let $\{\mathcal{F}_t\}_{t\in R^+}$ be the family of σ-algebras defined on R by

$$\{\mathcal{F}(\gamma\wedge t)\}_{t\in R^+}.$$

Then by the results of [5, VII, 54b], $\{\mathcal{F}_t\}_{t\in R^+}$ is a right-continuous family of σ-algebras, with no times of discontinuity, and γ is a totally inaccessible stopping time.

Let T be a positive constant, γ_T the stopping time $\gamma \wedge T$. We consider the system

$$\begin{aligned} dx &= 1_{\{t<\gamma_T\}}(A_1 x + B_1 u)\, dt + 1_{\{t\geqq\gamma_T\}}(A_2 x + B_2 u)\, dt, \\ x(0) &= x_0. \end{aligned} \tag{6.54}$$

One wants to minimize

$$\begin{aligned} E\int_0^T \{1_{\{t<\gamma_T\}}(|M_1 x_t|^2 + \langle N_1 u_t, u_t\rangle) \\ + 1_{\{t\geqq\gamma_T\}}(|M_2 x_t|^2\, dt + \langle N_2 u_t, u_t\rangle)\}\, dt. \end{aligned} \tag{6.55}$$

One must solve (6.1).

One has necessarily:

$$\mathcal{M}_t = E^{\mathcal{F}_t}\tilde{\mathcal{M}}_{\gamma_T}. \tag{6.56}$$

But on $t < \gamma_T$, a simple calculation done in [5, VII, 54b] proves that

$$d\mathcal{M}_s = \lambda(\mathcal{M}_s - \tilde{\mathcal{M}}_s)\, ds. \tag{6.57}$$

Moreover, (6.1) proves that

$$P_{\gamma_T} - P_{\gamma_T^-} = \tilde{\mathcal{M}}_{\gamma_T} - \mathcal{M}_{\gamma_T^-}. \tag{6.58}$$

Let P_2 be the solution of

$$\begin{aligned} &\frac{dP_2}{dt} + P_2 A_2 + A_2^* P_2 - P_2 C_2 N_2^{-1} C_2 P_2^* + M_2^* M_2 = 0, \\ &P_{2_T} = 0. \end{aligned} \tag{6.59}$$

P_2 is then self-adjoint and positive. Let P_1 be the solution of

$$\begin{aligned} &\frac{dP_1}{dt} + P_1 A_1 + A_1^* P_1 - P_1 C_1 N_1^{-1} C_1^* P_1 + M_1^* M_1 + \lambda(P_2 - P_1) = 0, \\ &P_{1_T} = 0. \end{aligned} \tag{6.60}$$

(6.60) has a solution, because it can be written

$$\begin{aligned} &\frac{dP_1}{dt} + P_1\left(A_1 - \frac{\lambda}{2} I\right) + \left(A_1^* - \frac{\lambda}{2} I\right) P_1 - P_1 C_1 N_1^{-1} C_1^* P_1 + M_1^* M_1 + \lambda P_2 = 0, \\ &P_{1_T} = 0, \end{aligned} \tag{6.61}$$

and $M_1^* M_1 + \lambda P_2$ is self-adjoint and positive. Then one will check that P is the process

$$P_t = 1_{\{t<\gamma_T\}} P_{1_t} + 1_{\{t\geqq\gamma_T\}} P_{2_t}. \tag{6.62}$$

Appendix. The purpose of this Appendix is to prove some very general results on linear differential stochastic equations. Assumptions and notations are taken from §§ 1 and 2.

Let A and $(B_i)_{i=1,\cdots,m}$ be a family of functions defined on $\Omega\times[0, +\infty[$ with values in $R^n\otimes R^n$, which we suppose to be bounded and $\mathcal{T}^*$-measurable.

THEOREM. *Let $Z_0\in L_2^0$, $u\in L_{21}$, $v=(v_1,\cdots,v_m)\in L_{22}$ and $M\in \underline{L}$. Then the equation*

$$\begin{aligned} &dZ=(AZ+u)\,dt+(BZ+v)\cdot dw+dM,\\ &Z(0)=Z_0, \end{aligned} \tag{A.1}$$

has one and only one solution, whose trajectories are a.s. right-continuous,

For any $T\geqq 0$, one has

$$E(\sup_{0\leqq t\leqq T}|Z_t|^2)<+\infty. \tag{A.2}$$

Proof. We prove existence and uniqueness on any interval $[0, T]$.

Existence. We consider the space of stochastic processes $\{X_t\}_{t\in R^+}$ such that for any t, X_t is $\mathcal{F}_t$-measurable, and such that:

$$\sup_{O\leqq t\leqq T}E|X_t|^2<+\infty.$$

Let B_T be the quotient of this space by the subspace of the processes equivalent to the zero process (i.e., for any t, $X_t=0$ a.s.).

B_T is then a Banach space with

$$\sup_{o\leqq t\leqq T}E|X_t|^2 \tag{A.3}$$

as a norm.

We consider now the space C_2^T defined in § 2 with a norm defined in (2.2). C_2^T is then also a Banach space. There is a continuous injection from C_2^T into B_T:

—if $X_t\in C_2^T$ and if for any t, $X_t=0$ a.s., then X_t is the zero process in C_2^T;

—if $X_t\in C_2^T$, then

$$\sup_{0\leqq t\leqq T}E|X_t|^2\leqq E\sup_{0\leqq t\leqq T}|X_t|^2. \tag{A.4}$$

Let Φ be the function defined on C_2^T with values in C_2^T such that if $Z\in C_2^T$, $\Phi(Z)$ is the process Z':

$$Z'_t=Z_0+\int_0^t(AZ+u)\,dt+\int_0^t(BZ+v)\cdot dw+M_t. \tag{A.5}$$

Z'_t has necessarily right-continuous trajectories. Moreover,

$$|Z'_t|^2\leqq k_1\left[\left|\int_0^t(AZ+u)\,ds\right|^2+\left|\int_0^t(BZ+v)\cdot dw\right|^2+|M_t|^2+|Z_0|^2\right], \tag{A.6}$$

$$\left|\int_0^t(AZ+u)\,ds\right|^2\leqq 2\left(\left|\int_0^t AZ\,ds\right|^2+\left|\int_0^t u\,ds\right|^2\right)$$

$$\leqq 2\left(\int_0^T\|A\|^2\,ds\int_0^T|Z_s|^2\,ds+\left(\int_0^T|u|\,ds\right)^2\right). \tag{A.7}$$

Since A is bounded, necessarily

$$\sup_{0\leqq t\leqq T}\left|\int_0^t (AZ+u)\,ds\right|^2 \leqq \lambda_1\left\{\int_0^T |Z|^2\,ds+\left(\int_0^T |u|\,ds\right)^2\right\}. \tag{A.8}$$

In the same way, by Doob's inequality on martingales (see [5, VI, Remark 2]), we have

$$\begin{aligned} E\left(\sup_{0\leqq t\leqq T}\left|\int_0^t BZ+v\cdot dw\right|^2\right) &\leqq 4E\left|\int_0^T BZ+v\cdot dw_s\right|^2 \\ &= 4E\int_0^T \|BZ+v\|^2\,ds. \end{aligned} \tag{A.9}$$

Moreover,

$$E\int_0^T \|BZ+v\|^2\,ds \leqq \lambda_2 E\left(\int_0^T \|B\|^2|Z|^2+\|v\|^2\right)ds. \tag{A.10}$$

Since B is bounded, we then have

$$E\int_0^T \|BZ+v\|^2\,ds \leqq \lambda_3\left(E\int_0^T |Z|^2\,d_s+E\int_0^T \|v\|^2\,dt\right). \tag{A.11}$$

Finally,

$$E(\sup_{0\leqq t\leqq T}|M_t|^2)\leqq 4E|M_T|^2. \tag{A.12}$$

From (A.8), (A.9), (A.11) and (A.12), we find that

$$E(\sup_{0\leqq t\leqq T}|Z'_t|^2)\leqq \lambda+\lambda' E\int_0^T |Z|^2\,ds\leqq \lambda+\lambda' T\sup_{0\leqq t\leqq T}E|Z_t|^2. \tag{A.13}$$

Φ is then an affine continuous mapping from C_2^T into C_2^T, the first space having the topology induced by B_T.

Since C_2^T is a Banach space, Φ can then be defined in a unique way on the closure of C_2^T in B_T, which we call $\overline{C}_2^T$, with values in C_2^T. For Z in $\overline{C}_2^T$, $\Phi(Z)$ is then well-defined.

Let Z_1 and $Z_1\in C_2^T$; Z and Z' are defined by

$$Z=Z_1-Z_2,\qquad Z'=\Phi(Z_1)-\Phi(Z_2). \tag{A.14}$$

Inequality (A.13), when applied with $Z_0=0$, $u=0$, $v=0$, $M=0$, proves that

$$E(\sup_{0\leqq t\leqq T}|Z'_t|^2)\leqq \lambda' E\int_0^T |Z|^2\,ds. \tag{A.15}$$

If $\Phi^2,\cdots,\Phi^n$ are the powers of Φ, then

$$\sup_{0\leqq t\leqq T}E|\Phi^n(Z_2)_t-\Phi^n(Z_1)_t|^2\leqq\frac{(\lambda' T)^n}{n!}\sup_{0\leqq t\leqq T}E|Z_{2_t}-Z_{1_t}|^2. \tag{A.16}$$

Relation (A.16) is also necessarily true when Z is in $\bar{C}_2^T$. When n is large enough, $(\lambda' T)^n/n! < 1$. Φ then has a unique fixed point, by the result given in [7, Chap. II, § 12]. For any Z in $\bar{C}_2^T$, $\Phi(Z)$ is in C_2^T. The fixed point is then necessarily in C_2^T.

Uniqueness. We prove that if Z'' is a right-continuous process such that

$$\begin{aligned} dZ'' &= AZ''\,dt + BZ''\cdot dw, \\ Z''(0) &= 0, \end{aligned} \tag{A.17}$$

then Z'' is the zero process.

From (A.17), Z'' is a continuous process. Let T_n be the stopping time defined by

$$T_n = \inf\{t: |Z_t''| \geqq n\}. \tag{A.18}$$

Then the process $Z''_{t\wedge T_n}$ is a solution of the equation

$$\begin{aligned} dZ &= 1_{\{t<T_n\}}(AZ\,dt + BZ\cdot dw), \\ Z(0) &= 0, \end{aligned} \tag{A.19}$$

and moreover,

$$E(\sup_{0\leqq t\leqq T} |Z''_{t\wedge T_n}|^2) < +\infty. \tag{A.20}$$

Z'' is then in C_2^T.

The previously proved uniqueness of the fixed point implies then that Z'' is the null process.

COROLLARY A.1. *If Z is the unique solution of* (A.1) *with* $(Z_0, u, v, M) \in L_2^0 \times L_{21} \times L_{22} \times \underline{L}$, *then the mapping* $(Z_0, u, v, M) \to Z$ is continuous from $L_2^0 \times L_{21} \times L_{22} \times \underline{L}$ into C_2^T.

Proof. From (A.8), (A.9), (A.10) and (A.12), we have

$$E(\sup_{0\leqq t\leqq T} |Z_t|^2) \leqq k(\|Z\|^2_{L_2^0} + \|u\|^2_{21} + \|v\|^2_{22} + \|M\|^2_{\underline{L}_2^T}) + \lambda' E\int_0^T |Z|^2\,ds. \tag{A.21}$$

Let b be defined by

$$b = k(\|Z\|^2_{L_2^0} + \|u\|^2_{21} + \|v\|^2_{22} + \|M\|^2_{\underline{L}_2^T}). \tag{A.22}$$

Then

$$E(\sup_{0\leqq t\leqq T} |Z_t|^2) \leqq b + \lambda' E\int_0^T |Z|^2\,ds. \tag{A.23}$$

This implies

$$E|Z_T|^2 \leqq b + \lambda' E\int_0^T |Z|^2\,ds, \tag{A.24}$$

and similarly

$$E|Z_t|^2 \leqq b + \lambda' E\int_0^t |Z|^2\,ds. \tag{A.25}$$

This implies, by Gronwall's lemma,

$$E|Z_t|^2 \leqq b\, e^{\lambda' t}. \tag{A.26}$$

From (A.23) and (A.26), we find

$$E\Big(\sup_{0\leqq t\leqq T} |Z_t|^2\Big) \leqq b\, e^{\lambda' T}. \tag{A.27}$$

The corollary follows from inequality (A.27). □

Remark A.1. For $s \leqq T$, and $h \in L_2^s$, let $\varkappa_h$ be the solution of:

$$\begin{cases} d\varkappa_h = A\varkappa_h\, dt + B\varkappa_h \cdot dw \\ \varkappa_h(s) = h. \end{cases}$$

Then by (A.8)–(A.12) and (A.27), it is easily seen that the norm of the mapping τ_s: $h \to \varkappa_h$ defined on L_2^s with values in C_2^T ($\varkappa_h$ is supposed to be null for $t < s$) can be bounded by a constant independent of s.

COROLLARY A.2. *Under the assumptions of Corollary* 1, *the mapping* $(Z_0, u, v, M) \to Z_T$ *is linear and continuous from* $L_2^0 \times L_{21} \times L_{22} \times L$ *into* L_2^T.

Proof. $Z \to Z_T$ is continuous from C_2^T into L_2^T. The result follows from Corollary 1.

REFERENCES

[1] J. M. BISMUT, *Analyse convexe et probabilités*, Doctoral dissertation, Faculté des Sciences de Paris, 1973.

[2] ———, *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. Appl., 44 (1973), pp. 384–404.

[3] U. G. HAUSSMAN, *Optimal stationary control with state and control dependent noise*, this Journal, 9 (1971), pp. 184–198.

[4] J. L. LIONS, *Contrôle Optimal des Systèmes Gouvernés par des Équations aux Derivées Partielles*, Dunod, Paris, 1968, English transl., Springer-Verlag, New York, 1971.

[5] P. A. MEYER, *Probabilités et Potentiel*, Hermann, Paris. English transl., Blaisdell, Waltham. Mass., 1966.

[6] ———, *Intégrales stochastiques, Séminaire de probabilités no. 1*, Lecture Notes in Mathematics, no. 39, Springer-Verlag, Berlin, 1967, pp. 72–142.

[7] L. SCHWARTZ, *Cours de l'École Polytechnique*, Hermann, Paris.

[8] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 312–326.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

A GENERAL UNIQUENESS THEOREM FOR SOLUTIONS OF STOCHASTIC DIFFERENTIAL EQUATIONS*

THOMAS C. GARD†

Abstract. We give a Lyapunov-type comparison theorem to obtain pathwise uniqueness for solutions of Ito stochastic differential equations in one dimension. This theorem contains basic criteria which generalize Ito's result, in which f and g satisfy Lipschitz conditions in the second variable. In the case of t-dependent moduli of continuity, we obtain as a corollary some new uniqueness results.

1. Introduction. The main result of this paper is an attempt to unify what is already known and facilitate the extension of the theory of the pathwise uniqueness property for the Ito equation

$$X_t = X_0 + \int_0^t f(s, X_s)\, ds + \int_0^t g(s, X_s)\, d\beta_s, \tag{1}$$

where β_t is a Brownian motion process, and the integrals in (1) are mean-square and Ito integrals, respectively.

Watanabe and Yamada [10] have shown that the pathwise uniqueness property implies both uniqueness in the law sense (solutions have the same distributions) and that solutions are measurable functions of the initial condition and the Brownian motion process. It is this fact that motivates the study of pathwise uniqueness. An example, attributed to Tanaka, is given in [10] to show that pathwise uniqueness and law uniqueness are not equivalent.

Ito's result [7] shows that (1) has the pathwise uniqueness property if f and g satisfy two-sided Lipschitz conditions in the second variable. However, Skorohod [9] has demonstrated the existence of a solution of (1) given that f and g are continuous. Thus it is appropriate to consider the question of uniqueness, apart from existence.

The theorem given here contains as special cases the results of Conway [2] and Watanabe and Yamada [10] both of which generalize Ito's criterion [7]. Conway assumes that f and g satisfy one-sided and two-sided Lipschitz conditions in the second variable, respectively, while Watanabe and Yamada assume f and g satisfy two-sided moduli of continuity conditions in the second variable weaker than the Lipschitz condition. It has not been shown, as yet, that the theorem represents a complete unification of the theory of this property for (1). For example, the criterion established by Skorohod [9], which removes moduli of continuity conditions of f at the expense of requiring positivity of g, thus far has eluded inclusion. However, an important special case of Skorohod's result and some new results assuming t-dependent modulus of continuity conditions are given as corollaries of the theorem.

* Received by the editors December 18, 1974, and in revised form March 24, 1975.

† Department of Mathematics, University of Georgia, Athens, Georgia 30602. This work is part of the author's dissertion under the direction of J. W. Heidel at the University of Tennessee and was supported in part by a National Science Foundation traineeship.

Although Ito's result as well as some of the abovementioned results are known for the case in which the random functions in (1) are vector-valued and recently, for the Ito's result at least, more generally operator-valued [8], for simplicity, only the scalar case will be discussed here.

2. Preliminaries. Definitions 1 and 2. Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $\{\mathcal{F}_t : t \in [0, T]\}$ a nondecreasing sequence of sub-σ-algebras of $\mathcal{F}$; i.e., if $s < t$, then $\mathcal{F}_s \subseteq \mathcal{F}_t$. Assume β_t is a Brownian motion process adapted to $\mathcal{F}_t$; i.e., β_t is a continuous $\mathcal{F}_t$ martingale satisfying

(i) $\beta_0 \equiv 0$,

(ii) for each s, and $t > s$, $E\{(\beta_t - \beta_s)^2 / \mathcal{F}_s\} = \sigma^2(t-s)$ for some constant σ.

Assume $f(t, x)$ and $g(t, x)$ are real-valued functions defined in $[0, T] \times R^1$ and Borel measurable in (t, x). By a *solution* of the Ito stochastic equation (1) is meant any a.s. sample continuous process Z_t adapted to $\mathcal{F}_t$ satisfying a.s.

$$Z_t = Z_0 + \int_0^t f(s, Z_s)\, ds + \int_0^t g(s, Z_s)\, d\beta_s, \qquad 0 \leqq t \leqq T,$$

where the second integral is understood as the Ito stochastic integral.

Equation (1) has the *pathwise uniqueness property* if any pair of solutions X_t and Y_t agreeing initially have a.s. identical sample paths, i.e.,

$$X_0 = Y_0 \quad \text{a.s.} \Rightarrow P\{X_t = Y_t,\ 0 \leqq t \leqq T\} = 1.$$

The principal tool used in the proof of the main theorem is the following special case of a result due to Ito [6, p. 187] which allows integration of smooth functions of solution processes.

ITO'S LEMMA. *Let $M > 0$ and F be a real-valued function on $[0, T] \times [-M, M] \times [-M, M]$ which is C^2 in x and y and C^1 in t. Assume $f_i(t, \omega)$ and $g_i(t, \omega)$ are Borel measurable real-valued functions defined on $[0, T] \times \Omega$, with $f_i(t, \cdot)$ and $g_i(t, \cdot)$ being also $\mathcal{F}_t$-measurable, $i = 1, 2$. If*

$$\begin{aligned} X_t(\omega) &= X_t(0) + \int_0^t f_1(s, \omega)\, ds + \int_0^t g_1(s, \omega)\, d\beta_s, \\ Y_t(\omega) &= Y_t(0) + \int_0^t f_2(s, \omega)\, ds + \int_0^t g_2(s, \omega)\, d\beta_s \end{aligned} \tag{2}$$

satisfy $|X_t| \leqq M$ and $|Y_t| \leqq M$, $0 \leqq t \leqq T$, a.s., then

$$\begin{aligned} F(t, X_t, Y_t) &= F(0, X_0, Y_0) \\ &+ \int_0^t \left[g_1 \frac{\partial F}{\partial x}(s, X_s, Y_s) + g_2 \frac{\partial F}{\partial y}(s, X_s, Y_s) \right] d\beta_s \\ &+ \int_0^t \left[f_1 \frac{\partial F}{\partial x}(s, X_s, Y_s) + f_2 \frac{\partial F}{\partial y}(s, X_s, Y_s) + \frac{\partial F}{\partial t}(s, X_s, Y_s) \right. \\ &\left. + \frac{1}{2} g_1^2 \frac{\partial^2 F}{\partial x^2}(s, X_s, Y_s) + \frac{1}{2} g_2^2 \frac{\partial^2 F}{\partial y^2}(s, X_s, Y_s) + g_1 g_2 \frac{\partial^2 F}{\partial x \partial y}(s, X_s, Y_s) \right] ds. \end{aligned}$$

3. A general uniqueness theorem.

LEMMA 1. *Suppose $F(t)$ is a continuous nonnegative function on $[0, T]$ with $F(0) = 0$. Assume there exists a scalar function $w(t, u) = \phi(t)\psi(u)$ satisfying*

(i) *ϕ is continuous and nonnegative on $(0, T]$,*

(ii) *ψ is continuous nondecreasing on $[0, \infty)$ and $\psi(0) = 0$,*

(iii) *$w(s, F(s)) \to 0^+$, as $s \to 0^+$,*

(iv) *the only solution $u(t)$ of*

$$u' = w(t, u) \tag{3}$$

on any interval such that $u(t)/t \to 0$ as $t \to 0^+$ is $u(t) = 0$.

If

$$F(t) \leqq \int_0^t w(s, F(s))\, ds, \qquad 0 < t \leqq T, \tag{4}$$

then $F(t) \equiv 0$ on $[0, T]$.

Proof. Define $W(t) = \int_0^t w(s, F(s))\, ds$, $0 \leqq t \leqq T$. $W(t) \geqq 0$ on $[0, T]$. Since w is continuous in both variables, and F is continuous, $w(t, F(t))$ is continuous in t; thus W is differentiable and

$$W'(t) = w(t, F(t)) = \phi(t)\psi(F(t)) \leqq \phi(t)\psi(W(t)) = w(t, W(t)),$$

the inequality following by (4), the fact that ψ is nondecreasing and the fact that ϕ is nonnegative. Also

$$W(t) \leqq t[\sup_{0 \leqq s \leqq t} w(s, F(s))].$$

So by (iii), $W(t)/t \to 0$ as $t \to 0^+$.

Now assume $W(t_0) > 0$ for some $t_0 \in (0, T]$. Let $\mathbf{u}(t)$ be the minimal solution of (3) such that $u(t_0) = W(t_0)$ existing on some interval to the left of t_0 (such a solution will exist by Hartman [4, p. 25]). If $\mathbf{u}(t_1) = 0$ for some $t_1 \in (0, t_0)$, then $\mathbf{u}(t)$ can be continued to the entire interval $(0, t_0]$ as a solution of (3) by setting $\mathbf{u}(t) = 0$ for $0 < t < t_1$, since by (ii), $u(t) \equiv 0$ is a solution of (3). This would contradict (iv) as $\mathbf{u}(t)$ would be a nontrivial ($\mathbf{u}(t_0) > 0$) solution of (3) satisfying $u(t)/t \to 0/t = 0$ as $t \to 0^+$. Thus $u(t) > 0$ as far as $\mathbf{u}(t)$ exists to the left of t_0. Now since $\mathbf{u}(t)$ is the minimal solution of (3), $\mathbf{u}(t_0) = W(t_0)$, and $W'(t) \leqq w(t, W(t))$ on $(0, t_0]$, it can be concluded that $0 < \mathbf{u}(t) \leqq W(t)$ as far as $\mathbf{u}(t)$ exists to the left of t_0 (see, for example, Hartman [4, p. 27], and make a time substitution). This means $\mathbf{u}(t)$ can be continued to the entire interval $(0, t_0]$ as a solution of (3) and $0 < \mathbf{u}(t) \leqq W(t)$. Furthermore, since $0 < \mathbf{u}(t)/t < W(t)/t$ and $\mathbf{u}(t)$ is a nontrivial solution of (3), this contradicts (iv).

Thus no such t_0 can exist. The conclusion is that $W(t)$, and hence $F(t)$ must vanish identically on $[0, T]$.

Remark. This lemma generalizes Hille's theorem 1.5.3 [5, p. 16].

Remark. The following lemma is an easy consequence of the bounded convergence theorem.

LEMMA 2. *Let X_t and Y_t be a.s. sample continuous processes on $[0, T]$. Assume there is a constant $M > 0$ such that $|X_t| \leqq M$ and $|Y_t| \leqq M$, $0 \leqq t \leqq T$, a.s. Then if $V(t, x, y)$ is a real-valued continuous function on $[0, T] \times [-M, M] \times [-M, M]$, $E\{V(t, X_t, Y_t)\}$ is a continuous function on $[0, T]$.*

Remark. Note that in the previous two lemmas all expectations exist and are finite because the expressions being averaged are bounded in each case.

DEFINITION 3. A scalar function $w(t, u) = \phi(t)\psi(u)$ is *admissible* if w satisfies conditions (i), (ii) and (iv) of Lemma 1, and ψ is concave.

THEOREM. *Suppose there exist scalar functions* $V(t, x, y)$, $w(t, u)$, *and a sequence of scalar functions* $\{V_n(t, x, y)\}$ *such that*

(i) V *is continuous and nonnegative on* $[0, T]\times R^2$;

(ii) $V(t, x, y) = 0$ *if and only if* $x = y$;

(iii) $w(t, u)$ *is admissible, and satisfies condition* (iii) *of Lemma* 1 *with* $F(t) = E\{V(t, X_t, Y_t)\}$ *for any pair of solutions of* (1) *with* $X_0 = Y_0$ *a.s., and* $|X_t| \leqq M$, $|Y_t| \leqq M$, $0 \leqq t \leqq T$, *a.s., where* M *is some positive constant*;

(iv) *for each* n, V_n *is nonnegative,* C^2 *in* x *and* y, *and* C^1 *in* t *on* $[0, T]\times R^2$, *and* $V_n(t, x, y) = 0$ *if* $x = y$;

(v) *for each* $t \in [0, T]$, $V_n(t, x, y) \to V(t, x, y)$ *in* R^2.

If there is a sequence $\{f_n\}$ *of functions such that* $f_n \to 0$ *in* $L^1[0, T]$ *and, for sufficiently large* n,

$$\frac{\partial V_n}{\partial t} \leqq w(t, V) + f_n,$$

$$DV_n \leqq w(t, V) + f_n$$

hold on $(0, T]\times R^2$, *where* D *is the differential operator*

$$f(t, x)\frac{\partial}{\partial x} + f(t, y)\frac{\partial}{\partial y} + \frac{\partial}{\partial t} + \frac{1}{2}g^2(t, x)\frac{\partial^2}{\partial x^2} + \frac{1}{2}g^2(t, y)\frac{\partial^2}{\partial y^2} + g(t, x)g(t, y)\frac{\partial^2}{\partial x \partial y},$$

then (1) *has the pathwise uniqueness property.*

Proof. Let X_t and Y_t be a pair of solutions of (1). Fix $M > 0$. Let τ_x and τ_y be the first exit stopping times of the processes X_t and Y_t, respectively, relative to M. Then $\tau = \tau_x \wedge \tau_y$, the minimum of τ_x and τ_y, is a stopping time, and the corresponding stopped processes $\hat{X}_t = X_{t\wedge\tau}$ and $\hat{Y}_t = Y_{t\wedge\tau}$ satisfy

$$Z_t = Z_0 + \int_0^t \hat{f}(s, Z_s)\, ds + \int_0^t \hat{g}(s, Z_s)\, d\beta_s, \tag{5}$$

where $\hat{f}(t, Z_t) = I_{\{t<\tau\}} f(t, Z_t)$ and $\hat{g}(t, Z_t) = I_{\{t<\tau\}} g(t, Z_t)$.

Now Ito's lemma can be applied to $V_n(t, \hat{X}_t, \hat{Y}_t)$, taking, in (2), f_1, f_2, g_1 and g_2 to be $\hat{f}(t, X_t(\omega))$, $\hat{f}(t, Y_t(\omega))$, $\hat{g}(t, X_t(\omega))$ and $\hat{g}(t, Y_t(\omega))$, respectively, to obtain for $0 \leqq t \leqq T$,

$$\begin{aligned} V_n(t, \hat{X}_t, \hat{Y}_t) = {} & \int_0^t \left[\hat{g}(s, \hat{X}_s)\frac{\partial V_n}{\partial x}(s, \hat{X}_s, \hat{Y}_s) + \hat{g}(s, \hat{Y}_s)\frac{\partial V_n}{\partial y}(s, \hat{X}_s, \hat{Y}_s)\right] d\beta_s \\ & + \int_0^t \left[\hat{f}(s, \hat{X}_s)\frac{\partial V_n}{\partial x}(s, \hat{X}_s, \hat{Y}_s) + \hat{f}(s, \hat{Y}_s)\frac{\partial V_n}{\partial y}(s, \hat{X}_s, \hat{Y}_s)\right. \\ & \left. + \frac{\partial V_n}{\partial t}(s, \hat{X}_s, \hat{Y}_s)\right. \\ & \left. + \frac{1}{2}g^2(s, \hat{X}_s)\frac{\partial^2 V_n}{\partial x^2}(s, \hat{X}_s, \hat{Y}_s) + \frac{1}{2}\hat{g}^2(s, \hat{Y}_s)\frac{\partial^2 V_n}{\partial y^2}(s, \hat{X}_s, \hat{Y}_s)\right. \end{aligned} \tag{6}$$

$$+\hat{g}(s,\hat{X}_s)\hat{g}(s,\hat{Y}_s)\frac{\partial^2 V_n}{\partial x\partial y}(s,\hat{X}_s,\hat{Y}_s)\Big]\,ds$$

$$=I_1+I_2,$$

say (noting that $\hat{X}_0=\hat{Y}_0$ a.s. $\Rightarrow V_n(0,\hat{X}_0,\hat{Y}_0)=0$ a.s.). Since V_n has bounded partial derivatives on the compact set $[0,T]\times[-M,M]\times[-M,M]$, and f and g are bounded, $E\{V_n\}$, $E\{I_1\}$ and $E\{I_2\}$ are finite, and $E\{I_1\}=0$.

Now an estimate is obtained for $E\{I_2\}$ involving the function w. Fix $\omega\in\Omega$. Suppose $s<\tau(\omega)$. Then, denoting by $\hat{D}V_n(s,\hat{X}_s,\hat{Y}_s)$ the integrand in I_2,

$$\begin{aligned}\hat{D}V_n(s,\hat{X}_s,\hat{Y}_s)&=f(s,\hat{X}_s)\frac{\partial V_n}{\partial x}(s,\hat{X}_s,\hat{Y}_s)+f(s,\hat{Y}_s)\frac{\partial V_n}{\partial y}(s,\hat{X}_s,\hat{Y}_s)\\&\quad+\frac{\partial V_n}{\partial t}(s,\hat{X}_s,\hat{Y}_s)+\frac{1}{2}g^2(s,\hat{X}_s)\frac{\partial^2 V_n}{\partial x^2}(s,\hat{X}_s,\hat{Y}_s)\\&\quad+\frac{1}{2}g^2(s,\hat{Y}_s)\frac{\partial^2 V_n}{\partial y^2}(s,\hat{X}_s,\hat{Y}_s)\\&\quad+g(s,\hat{X}_s)g(s,\hat{Y}_s)\frac{\partial^2 V_n}{\partial x\partial y}(s,\hat{X}_s,\hat{Y}_s)\\&\leqq w(s,V(s,\hat{X}_s,\hat{Y}_s))+f_n(s)\quad\text{a.s.}\end{aligned}$$

On the other hand, if $s\geqq\tau(\omega)$, $\hat{f}$ and $\hat{g}$ vanish a.s., so

$$\hat{D}V_n(s,\hat{X}_s,\hat{Y}_s)=\frac{\partial V_n}{\partial t}(s,\hat{X}_s,\hat{Y}_s)\leqq w(s,V(s,\hat{X}_s,\hat{Y}_s))+f_n(s)\quad\text{a.s.}$$

(the inequalities follow by assumptions on V_n, V, w and f_n).

Thus, taking expectation in (6),

$$\begin{aligned}E\{V_n(t,\hat{X}_t,\hat{Y}_t)\}=E\{I_2\}&\leqq E\Big\{\int_0^t w(s,V(s,\hat{X}_s,\hat{Y}_s))\,ds+\int_0^t f_n(s)\,ds\Big\}\\&=E\Big\{\int_0^t w(s,V(s,\hat{X}_s,\hat{Y}_s))\,ds\Big\}+\int_0^t f_n(s)\,ds.\end{aligned}\tag{7}$$

But,

$$\int_0^t w(s,V(s,\hat{X}_s,\hat{Y}_s))\,ds=\int_0^t \phi(s)\psi(V(s,\hat{X}_s,\hat{Y}_s))\,ds.$$

$E\{V(s,\hat{X}_s,\hat{Y}_s)\}<\infty$, since V is bounded on $[0,T]\times[-M,M]\times[-M,M]$. Thus by Jensen's inequality, using concavity of ψ,

$$E\{\psi(V(s,\hat{X}_s,\hat{Y}_s))\}\leqq\psi(E\{V(s,\hat{X}_s,\hat{Y}_s)\}).\tag{8}$$

So $\phi(s)E\{\psi(V(s,\hat{X}_s,\hat{Y}_s))\}$ is an integrable function of s, noting that $\phi(s)\psi(E\{V(s,\hat{X}_s,\hat{Y}_s)\})\to 0$ as $s\to 0^+$ by (iii). Also $\int_0^t\phi(s)\psi(V(s,\hat{X}_s,\hat{Y}_s))\,ds$ is an integrable function of ω as $\phi(s)\psi(E\{V(s,\hat{X}_s,\hat{Y}_s)\})\to 0$ as $s\to 0^+$ and (4) imply that $\phi(s)\psi(V(s,\hat{X}_s,\hat{Y}_s))\to 0$ a.s. as $s\to 0^+$. Hence Fubini's theorem and (8) can be

applied to (7) to yield

$$E\{V_n(t, \hat{X}_t, \hat{Y}_t)\} \leqq \int_0^t \phi(s) E\{\psi(V(s, \hat{X}_s, \hat{Y}_s)\}\, ds + \int_0^t f_n(s)\, ds$$

(9)

$$\leqq \int_0^t \phi(s)\psi(E\{V(s, \hat{X}_s, \hat{Y}_s)\})\, ds + \int_0^t f_n(s)\, ds.$$

Now by (v), $V_n(t, \hat{X}_t(\omega), \hat{Y}_t(\omega)) \to V(t, \hat{X}_t(\omega), \hat{Y}_t(\omega))$, so taking $\underline{\lim}_{n\to\infty}$ in (9) and applying Fatou's lemma,

$$E\{V(t, \hat{X}_t, \hat{Y}_t)\} \leqq \int_0^t \phi(s)\psi(E\{V(s, \hat{X}_s, \hat{Y}_s)\})\, ds. \tag{10}$$

Set $F(t) = E\{V(t, \hat{X}_t, \hat{Y}_t)\}$. By Lemma 2, $F(t)$ is continuous on $[0, T]$. Since V is nonnegative, F is nonnegative. $F(0) = 0$, as $\hat{X}_0 = \hat{Y}_0$ a.s. and V satisfies (ii). Assumption (iii) gives that (iii) of Lemma 1 holds. Noting that (10) is the final hypothesis needed in Lemma 1, this result can be applied to yield

$$E\{V(t, \hat{X}_t, \hat{Y}_t)\} = F(t) \equiv 0 \quad \text{on } [0, T].$$

By nonnegativity of V, this means $V(t, \hat{X}_t, \hat{Y}_t) = 0$ a.s. Thus by (ii), $\hat{X}_t = \hat{Y}_t$ a.s., for each $t \in [0, T]$. Since the processes $\hat{X}_t$, $\hat{Y}_t$ are a.s. sample continuous,

$$P\{\omega : \hat{X}_t(\omega) = \hat{Y}_t(\omega),\ 0 \leqq t \leqq T\} = 1$$

(see Yeh [11, p. 2]). Finally, since M was arbitrary, and the processes X_t, Y_t are a.s. sample continuous, then

$$P\{\omega : X_t(\omega) = Y_t(\omega),\ 0 \leqq t \leqq T\} = 1,$$

completing the proof.

Remark. If nonnegativity of V_n is replaced by $V_n(t, x, y) \leqq V(t, x, y)$, the result is obtained by application of the Lebesgue dominated convergence theorem instead of Fatou's lemma.

4. Examples.

COROLLARY 1 (Conway [2]). *Assume there are positive constants K and L such that*

(i) $f(t, x) - f(t, y) \leqq K(x - y)$, $-\infty < y < x < \infty$,

(ii) $|g(t, x) - g(t, y)| \leqq L|x - y|$, $x, y \in R$.

Then (1) *has the pathwise uniqueness property.*

Proof. The Theorem is applied with

$$V(t, x, y) = \tfrac{1}{2}(x - y)^2 \exp[-2K - L^2]t,$$

$$V_n(t, x, y) = V(t, x, y), \quad \text{all } n,$$

$$w(t, u) \equiv 0 \quad \text{and} \quad f_n(t) \equiv 0, \quad \text{all } n.$$

Clearly, conditions (i)–(v) of the Theorem are satisfied. It remains to verify the differential inequalities

$$\frac{\partial V}{\partial t} \leqq 0, \tag{11}$$

$$f(t,x)\frac{\partial V}{\partial x}+f(t,y)\frac{\partial V}{\partial y}+\frac{\partial V}{\partial t}+\frac{1}{2}g^2(t,x)\frac{\partial^2 V}{\partial x^2}$$
$$+\frac{1}{2}g^2(t,y)\frac{\partial^2 V}{\partial y^2}+g(t,x)g(t,y)\frac{\partial^2 V}{\partial x\partial y}\leqq 0 \tag{12}$$

in order to apply the theorem. To do this, the following partial derivatives are computed:

$$\frac{\partial V}{\partial t}=\left(-K-\frac{L^2}{2}\right)(x-y)^2\exp[-2K-L^2]t,$$

$$\frac{\partial V}{\partial x}=(x-y)\exp[-2K-L^2]t=-\frac{\partial V}{\partial y},$$

$$\frac{\partial^2 V}{\partial x^2}=\frac{\partial^2 V}{\partial y^2}=\exp[-2K-L^2]t=-\frac{\partial^2 V}{\partial x\partial y}.$$

These computations show that (11) is satisfied, and that the left-hand side of (12) can be written

$$[(f(t,x)-f(t,y))(x-y)+(-K-L^2/2)(x-y)^2$$
$$+\tfrac{1}{2}(g(t,x)-g(t,y))^2]\exp[-2K-L^2]t. \tag{13}$$

Now (i) is equivalent to

(i′) $\qquad (f(t,x)-f(t,y))(x-y)\leqq K(x-y)^2, \qquad x, y\in R,$

and squaring both sides of (ii) yields

(ii′) $\qquad (g(t,x)-g(t,y))^2\leqq L^2(x-y)^2, \qquad x, y\in R.$

Conditions (i′) and (ii′) imply that the first factor of (13) is nonpositive, and so (12) is valid.

COROLLARY 2 (Watanabe and Yamada [10]). *Suppose there exist positive and nondecreasing functions χ and ρ defined on $(0,\infty)$ with χ concave and*

$$\int_{0^+}\rho^{-2}(u)\,du=+\infty=\int_{0^+}\chi^{-1}(u)\,du$$

such that, for $x, y\in R$,

(i) $|f(t,x)-f(t,y)|\leqq\chi(|x-y|)$,

(ii) $|g(t,x)-g(t,y)|\leqq\rho(|x-y|)$.

Then (1) has the pathwise uniqueness property.

Proof. Watanabe and Yamada [10] demonstrate the existence of a sequence $\{h_n\}$ of nonnegative C^2-functions on R satisfying

$$h_n(0)=h_n'(0)=h_n''(0)=0, \tag{14}$$

$$h_n(u)\to|u|, \tag{15}$$

$$|h_n'(u)|\leqq 1 \quad\text{and}\quad |h_n''(u)|\leqq\frac{2}{n}\rho^{-2}(|u|), \qquad u\neq 0. \tag{16}$$

The Theorem is applied with

$$V(t, x, y) = |x - y|, \qquad V_n(t, x, y) = h_n(x - y),$$
$$w(t, u) = \chi(u) \quad \text{and} \quad f_n(t) = 1/n.$$

Clearly, conditions (i), (ii), (iv) and (v) of the Theorem are satisfied. Since χ is concave on $(0, \infty)$, $-\chi$ is convex, and hence continuous (see, for example, [1, p. 28]) and so χ is continuous. Since $\int_{0^+} 1/\chi(u)\, du = +\infty$, $\chi(u) \to 0$, as $u \to 0$, so that χ can be continuously extended to $[0, \infty)$ by defining $\chi(0) = 0$. Also, since $E\{|X_t - Y_t|\} \to 0$ by Lemma 2 as $t \to 0^+$, for any pair of bounded solutions of (1) with $X_0 = Y_0$ a.s., $\chi(E\{|X_t - Y_t|\}) \to 0$ as $t \to 0^+$. To complete the verification of condition (iii) of the Theorem for the function χ, condition (iv) of Lemma 1 must be shown to hold. To this end, let $u(t)$ be any solution of the differential equation $u' = \chi(u)$ such that $u(t) \to 0$ as $t \to 0^+$. (Certainly, if it is shown that any such solution must be trivial, then (iv) of Lemma 1 will hold, as $u(t)/t \to 0$ as $t \to 0^+$ implies that $u(t) \to 0$, as $t \to 0^+$.) Assume $u(t) \not\equiv 0$. Let $t_1 > 0$ such that $u(t_1) > 0$. (Since χ is defined only on $[0, \infty)$, $u(t) \geqq 0$.) Let

$$t_0 = \begin{cases} \sup\{t : 0 < t \leqq t_1 \quad \text{and} \quad u(t) = 0\}, \\ 0, \quad \text{if } u(t) \neq 0, \text{ for all } t \in (0, t_1]. \end{cases}$$

By continuity and the assumption that $u(t_1) > 0$, $t_0 < t_1$. Now if $t_0 < t < t_1$, then

$$\int_{u(t)}^{u(t_1)} \frac{1}{\chi(u)}\, du = t_1 - t \tag{17}$$

since $u(t)$ is a solution of $u' = \chi(u)$. Letting $t \to t_0$, the right-hand side of (17) approaches $t_1 - t_0$, while the left-hand side of (17) approaches $+\infty$ by assumption on χ. This contradiction means that $u(t) \equiv 0$, and so (iii) of the Theorem is finally verified.

Once again, it remains to verify the differential inequalities

$$\frac{\partial V_n}{\partial t} \leqq \chi(V) + \frac{1}{n}, \tag{18}$$

$$DV_n \leqq \chi(V) + \frac{1}{n} \tag{19}$$

in order to apply the theorem.

Since $\partial V_n / \partial t \equiv 0$, for each n, (18) is satisfied, noting that χ is nonnegative. Since F_n is a function of $x - y$, then

$$\frac{\partial V_n}{\partial x} = -\frac{\partial V_n}{\partial y} \quad \text{and} \quad \frac{\partial^2 V_n}{\partial x^2} = \frac{\partial^2 V_n}{\partial y^2} = -\frac{\partial^2 V_n}{\partial x \partial y}.$$

Using these simplifications, and the estimates in (16),

$$\begin{aligned} DV_n &= (f(t, x) - f(t, y))\frac{\partial V_n}{\partial x} + \frac{1}{2}(g(t, x) - g(t, y))^2 \frac{\partial^2 V_n}{\partial x^2} \\ &\leqq |f(t, x) - f(t, y)| + \frac{1}{2}(g(t, x) - g(t, y))^2 \cdot \frac{2}{n}\rho^{-2}(|x - y|). \end{aligned} \tag{20}$$

By assumptions (i) and (ii) applied to (20),

$$DV_n \leqq \chi(|x-y|)+\frac{1}{n}=\chi(V)+\frac{1}{n}$$

verifying (19), and so the Theorem can be applied to give the desired result.

Remark. Recall that in the proof of the Theorem, stopping times were introduced essentially in order to limit the analysis to bounded solutions. The differential inequalities assumed were only needed on compact sets containing the ranges of these solutions. It suffices, therefore, for the conditions (i) and (ii) in Corollary 1 and Corollary 2, which imply the validity of the differential inequality conditions of the Theorem to hold locally.

COROLLARY 3. *Suppose* $f(t, x)=f(x)$ *is continuous and bounded and* $g(t, x)=g(x)>c>0$ *is bounded and satisfies a local Holder continuity condition of order* $\alpha \geqq \frac{1}{2}$; *i.e., given constant* $M>0$, *there exist constants* $L>0$ *and* $\alpha \geqq \frac{1}{2}$ *such that*

$$|g(x_1)-g(x_2)| \leqq L|x_1-x_2|^{\alpha}, \qquad |x_1|, |x_2| \leqq M.$$

Then (1) *has the pathwise uniqueness property.*

Proof. Let

$$h(x)=\int_0^x \exp\left[-2\int_0^y \frac{f(z)}{g^2(z)}\,dz\right] dy.$$

Since f and g are continuous, h is a C^2-function on R. Because of the positivity of the exponential function, h is 1–1. Furthermore, h^{-1} is a C^1-function:

$$h'(x)=\exp\left[-2\int_0^x \frac{f(z)}{g^2(z)}\,dz\right]>0,$$

so h^{-1} is differentiable and $(h^{-1})'(h(x))=1/h'(x)$; from this formula and the continuity of the functions h^{-1} and h', continuity of $(h^{-1})'$ can be concluded.

Now let X_t be a solution of (1). Since h and its derivatives are bounded, Ito's lemma can be applied to $h(X_t)$. (The form of Ito's lemma given here requires boundedness of the processes, but this was just to insure that the relevant functions of the processes would be bounded.)

$$\begin{aligned} h(X_t) &= h(X_0)+\int_0^t [f(X_s)h'(X_s)+\tfrac{1}{2}g^2(X_s)h''(X_s)]\,ds \\ &\quad +\int_0^t h'(X_s)g(X_s)\,d\beta_s. \end{aligned} \tag{21}$$

By computing the derivatives h' and h'', it is easy to see that

$$f(x)h'(x)+\tfrac{1}{2}g^2(x)h''(x)\equiv 0. \tag{22}$$

Setting $Y_t=h(X_t)$, and making use of (22), equation (21) can be written

$$Y_t = Y_0+\int_0^t G(Y_s)\,d\beta_s, \tag{23}$$

where $G(y)=h'(h^{-1}(y))\cdot g(h^{-1}(y))$.

The properties of h imply that Y_t is an a.s. sample continuous stochastic process on $[0, T]$ adapted to the same sequence of sub-σ-algebras as X_t, and that the pathwise uniqueness for (1) is equivalent to that for (23).

Since h' and h^{-1} are C^1-functions, they satisfy local Lipschitz conditions; i.e., for each $M>0$, and $N>0$, there are $L_M>0$ and $L_N>0$ such that

$$|h'(x_1)-h'(x_2)|\leqq L_M|x_1-x_2|, \qquad |x_1|, |x_2|\leqq M, \tag{24}$$

$$|h^{-1}(y_1)-h^{-1}(y_2)|\leqq L_N|y_1-y_2|, \qquad |y_1|, |y_2|\leqq N. \tag{25}$$

It is now asserted that G is locally Hölder continuous, because g is locally Hölder continuous. The proof of this assertion follows:

$$\begin{aligned}|G(y_1)-G(y_2)| &= |h'(h^{-1}(y_1))\cdot g(h^{-1}(y_1))-h'(h^{-1}(y_2))\cdot g(h^{-1}(y_2))|\\ &\leqq |h'(h^{-1}(y_1))|\,|g(h^{-1}(y_1))-g(h^{-1}(y_2))|\\ &\quad +|g(h^{-1}(y_2))|\,|h'(h^{-1}(y_1))-h'(h^{-1}(y_2))|.\end{aligned}$$

Now, assume $y_1, y_2\in[-N, N]$. Then $h^{-1}[-N, N]\subseteq[-M, M]$ for some $M>0$. By continuity and the Hölder continuous property of g, there are constants M_1, M_2, and L such that

$$\begin{aligned}&|h'(x)|\leqq M_1, \quad |g(x)|\leqq M_2, \quad \text{and for some } \alpha\geqq\tfrac{1}{2},\\ &|g(x_1)-g(x_2)|\leqq L|x_1-x_2|^{\alpha} \quad \text{for } x, x_1, x_2\in[-M, M].\end{aligned}$$

Using these estimates, and (24) and (25),

$$\begin{aligned}|G(y_1)-G(y_2)| &\leqq M_1L|h^{-1}(y_1)-h^{-1}(y_2)|^{\alpha}+M_2|h'(h^{-1}(y_1))-h'(h^{-1}(y_2))|\\ &\leqq M_1LL_N^{\alpha}|y_1-y_2|^{\alpha}+M_2L_ML_N|y_1-y_2|\\ &\leqq [M_1LL_N^{\alpha}+2M_2L_ML_NN^{1-\alpha}]|y_1-y_2|^{\alpha},\end{aligned}$$

verifying the assertion.

Pathwise uniqueness for (23) follows, since the function $\rho(u)=Ku^{\alpha}$, for any positive constant K, and $\alpha\geqq\frac{1}{2}$ satisfies the assumptions in Corollary 2.

Remark. The transformation h made use of in the proof of the preceding corollary is given in Gihman and Skorohod [3, p. 34].

Remark. The following result is well known.

LEMMA 3. *Let X_t and Y_t be solutions of* (1), *with $X_0=Y_0$ a.s., and assume there is a constant $M>0$ such that $|X_t|\leqq M$ and $|Y_t|\leqq M$, $0\leqq t\leqq T$, a.s.*

(a) *If f and g are bounded and Borel measurable, then*

$$E\{|X_t-Y_t|\}=O(t^{1/2}) \quad \text{as } t\to 0.$$

(b) *If f and g are continuous, then*

$$E\{|X_t-Y_t|\}=o(t^{1/2}) \quad \text{as } t\to 0.$$

COROLLARY 4. *Suppose there are constants $A>0$, $\alpha\geqq\frac{1}{2}$, a nonnegative function $\lambda(t)$, continuous and square integrable on $(0, T]$, and a function $\rho(u)$ as in Corollary 2 such that*

(i) $|f(t, x)-f(t, y)|\leqq(A/t^{\alpha})|x-y|$,

(ii) $|g(t, x)-g(t, y)|\leqq\lambda(t)\rho(|x-y|)$

for all $x, y\in R$, $t\in(0, T]$.

If f and g are also continuous in both variables, then (1) *has the pathwise uniqueness property.*

Proof. The Theorem is applied with $V(t, x, y) = |x - y|$, $V_n(t, x, y) = h_n(x - y)$, where $\{h_n\}$ is as in Corollary 2, $w(t, u) = A/t^\alpha \cdot u$, and $f_n(t) = (1/n)\lambda^2(t)$. Conditions (i), (ii), (iv) and (v) of the Theorem are satisfied. It is clear that $w(t, u)$ satisfies (i) and (ii) of Lemma 1 and that $\psi(u) = u$ is concave. Thus to verify that w is admissible, it remains to demonstrate (iv) of Lemma 1. But this follows from the fact that nontrivial solutions of the differential equation $u' = w(t, u)$ have the form

$$u(t) = K \exp\left[\frac{1}{1-\alpha} t^{1-\alpha}\right]$$

for K constant. Thus no nontrivial solution $u(t)$ can satisfy $u(t) \to 0$ as $t \to 0^+$. So $w(t, u)$ is admissible. Also, since by Lemma 3(b), for any pair X_t, Y_t of bounded solutions of (1) with $X_0 = Y_0$ a.s., $E\{|X_t - Y_t|\} = o(t^{1/2})$ as $t \to 0$, $w(t, E\{|X_t - Y_t|\}) = A/t^\alpha \cdot E\{|X_t - Y_t|\} \to 0$ as $t \to 0^+$, since $\alpha \geqq \frac{1}{2}$. So (iii) of the Theorem is completely verified.

It remains to verify the differential inequalities

$$\frac{\partial V_n}{\partial t} \leqq A/t^\alpha \cdot V + \frac{1}{n}\lambda^2, \tag{26}$$

$$DV_n \leqq A/t^\alpha \cdot V + \frac{1}{n}\lambda^2. \tag{27}$$

Since $\partial V_n/\partial t \equiv 0$, for each n, (26) is satisfied. As in the proof of Corollary 2, it can be shown that

$$DV_n \leqq |f(t, x) - f(t, y)| + \tfrac{1}{2}(g(t, x) - g(t, y))^2 \frac{2}{n}\rho^{-2}(|x - y|). \tag{28}$$

Applying assumptions (i) and (ii) to (28), the following estimate is obtained:

$$DV_n \leqq \frac{A}{t^\alpha}|x - y| + \frac{1}{n}\lambda^2(t);$$

i.e., (27) is verified.

Remark. Continuity of f and g was used only in the application of Lemma 3 to verify that

$$w(t, E\{|X_t - Y_t|\}) \to \frac{A}{t^\alpha} \cdot E\{|X_t - Y_t|\} \to 0 \quad \text{as } t \to 0^+.$$

If f and g are assumed to be bounded and Borel measurable, Lemma 3(a) can be applied to give the required result, provided $\alpha < \frac{1}{2}$.

COROLLARY 5. *Assume there are constants $A > 0$ and $B > 0$ with $A + B^2/2 \leqq \frac{1}{2}$ such that for $t \in (0, T]$,*

(i) $f(t, x) - f(t, y) \leqq (A/t)(x - y)$, $-\infty < y < x < \infty$,

(ii) $|g(t, x) - g(t, y)| \leqq (B/t^{1/2})|x - y|$, $x, y \in R$.

If f and g are also continuous in both variables, then (1) *has the pathwise uniqueness property.*

Proof. The Theorem is applied with $V(t, x, y) = \frac{1}{2}(x-y)^2$, $V_n(t, x, y) \equiv V(t, x, y)$, all n, $w(t, u) = u/t$, and $f_n \equiv 0$, all n.

Once again, it is clear that conditions (i), (ii), (iv) and (v) of the Theorem are satisfied, that w satisfies (i) and (ii) of Lemma 1, and that $\psi(u) = u$ is concave.

Now, nontrivial solutions $u(t)$ of the differential equation $u' = w(t, u)$ are of the form

$$u(t) = Kt$$

for K constant. Thus w is admissible, since no nontrivial solution $u(t)$ of this form can satisfy $u(t)/t \to 0$ as $t \to 0^+$.

Furthermore, it is shown in proving Lemma 3 that $E\{(X_t - Y_t)^2\} = o(t)$ as $t \to 0$, for any pair of bounded solutions X_t and Y_t of (2) with $X_0 = Y_0$ a.s. Thus

$$w(t, E\{(X_t - Y_t)^2\}) = 1/t \cdot E\{(X_t - Y_t)^2\} \to 0 \quad \text{as } t \to 0^+,$$

and so (iii) is completely verified.

Finally, it remains to verify the differential inequalities

$$\frac{\partial V}{\partial t} \leqq \frac{1}{t} \cdot V, \tag{29}$$

$$DV \leqq \frac{1}{t} \cdot V. \tag{30}$$

Since $\partial V/\partial t \equiv 0$, (29) is satisfied. Because V is a function of $x - y$,

$$\begin{aligned} DV &= (f(t, x) - f(t, y))\frac{\partial V}{\partial x} + \frac{1}{2}(g(t, x) - g(t, y))^2 \frac{\partial^2 V}{\partial x^2} \\ &= (f(t, x) - f(t, y))(x - y) + \tfrac{1}{2}(g(t, x) - g(t, y))^2 \end{aligned} \tag{31}$$

as $\partial V/\partial x = x - y$, and $\partial^2 V/\partial x^2 = 1$. Now applying the assumptions (i) and (ii) to (21), noting that (i) is equivalent to

$$(f(t, x) - f(t, y))(x - y) \leqq (A/t)(x - y)^2, \qquad x, y \in R, \tag{i'}$$

the following estimate is obtained:

$$DV \leqq \frac{A}{t}(x-y)^2 + \frac{1}{2}\frac{B^2}{t}(x-y)^2 = \frac{1}{t}\left(A + \frac{B^2}{2}\right)(x-y)^2.$$

Thus since $A + B^2/2 \leqq \frac{1}{2}$, (30) is verified.

REFERENCES

[1] R. B. ASH, *Real Analysis and Probability*, Academic Press, New York, 1972.

[2] E. CONWAY, *Stochastic equations with discontinuous drift*, Trans. Amer. Math. Soc., 157 (1971), pp. 235–245.

[3] I. I. GIHMAN AND A. V. SKOROHOD, *Stochastic Differential Equations*, Springer-Verlag, New York, 1972.

[4] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.

[5] E. HILLE, *Lectures on Ordinary Differential Equations*, Addison-Wesley, Reading, Mass., 1969.

[6] K. ITO, *Lectures on Stochastic Processes*, Tata Institute, Bombay, India, 1960.
[7] ———, *On stochastic differential equations*, Mem. Amer. Math. Soc., 4 (1951), pp. 1–51.
[8] H. H. KUO, *On operator-valued stochastic integrals*, Bull. Amer. Math. Soc., 79 (1973), pp. 207–210.
[9] A. V. SKOROHOD, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Mass., 1965.
[10] S. WATANABE AND T. YAMADA, *On the uniqueness of solutions of stochastic differential equations*, J. Math. Kyoto Univ., 11-1 (1971), pp. 155–167, 11-3 (1971), pp. 553–563.
[11] J. YEH, *Stochastic Processes and the Wiener Integral*, Dekker, New York, 1973.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

AN EXISTENCE THEOREM FOR A GENERAL BOLZA PROBLEM*

A. D. IOFFE†

Abstract. An existence theorem is proved for a general Bolza problem covering various types of constrained optimal control problems. This theorem seems to be the most general; it covers some well-known results of Cesari, Olech, Rockafellar and certain others. The proof of the theorem is based upon a new growth condition which is a combination of those of Cinquini and the author on the one hand and those of Olech and Rockafellar on the other hand. Under this condition, the integrand of the Bolza problem is allowed to decrease arbitrarily fast in the state variable, but the decrease in the state variable and the increase in the control variable must be consistent in some sense.

1. Introduction. We consider here the problem of minimizing of the functional

$$I(T, x(\cdot)) = \int_0^T L(t, x(t), \dot{x}(t))\, dt + l(T, x(0), x(T)) \tag{1.1}$$

over the set $\mathfrak{S}$ of all pairs $(T, x(\cdot))$, where $T > 0$ and $x(\cdot)$ is an absolutely continuous mapping from $[0, T]$ into R^n. Here L and l are *extended-real-valued* functions (that is, they may assume infinite values as well as real values). We put aside the justification of the problem, referring the reader to [5], [7]. We note only that this problem covers, in particular, optimal control problems with various constraints.

In what follows, we suppose that L and l satisfy the following assumptions:

(i) the integrand $L(t, x, y)$ is convex in y for all x and almost all $t \in [0, \infty)$;

(ii) the integrand L is lower semicontinuous in (x, y) for almost all $t \in [0, \infty)$;

(iii) the integrand L is $\mathscr{L} \otimes \mathscr{B}$-measurable, that is, measurable with respect to the σ-algebra generated in $[0, \infty) \times R^n \times R^n$ by products of Lebesgue measurable subsets of $[0, \infty)$ and Borel subsets of $R^n \times R^n$;

(iv) the terminal function $l(t, x, z)$ is lower semicontinuous.

These assumptions are not restrictive at all. They are present in practically every existence theorem, usually in even more rigid forms. In fact, the assumptions (i), (ii) and (iv) ensure in essence the lower semicontinuity of the functional I (cf. [5] and [8]), while the measurability assumption (iii) ensures in particular the Lebesgue measurability of certain functions, for example, the function

$$t \to L(t, x(t), y(t)) \tag{1.2}$$

for any measurable $x(\cdot)$ and $z(\cdot)$.

To establish the existence of a solution to (1.1), it is sufficient to prove that I is lower semicontinuous and has nonempty and compact level sets in some appropriate topology in $\mathfrak{S}$. As we have already mentioned, lower semicontinuity of I is essentially connected with the convexity assumption (i) and the semicontinuity assumptions (ii) and (iv). The main differences between existence theorems lie in the criteria which guarantee compactness, though such criteria are always connected with so-called growth conditions. Here we suggest a new compactness criterion which seems to be the most general and covers many of

* Received by the editors February 10, 1975.

† Profsojunznaja ul., 97.k.1, kv.203, Moscow B-279, USSR.

the known results. In this way, a very general existence theorem is established. We indicate now two known results which are special cases of our theorem. In quoting these results, we confine ourselves for simplicity to the fixed-time problem, where

$$l(t, x, z) = \begin{cases} l(x, z) & \text{if } t = T_0, \\ +\infty & \text{otherwise.} \end{cases}$$

Let

$$H(t, x, p) = \sup_y (\langle p, y \rangle - L(t, x, y)) \tag{1.3}$$

(here $\langle \cdot, \cdot \rangle$ denotes the scalar product in R^n).

Then under the assumptions (i)–(iv) and under the assumption that $I(T_0, x(\cdot)) < \infty$ for at least one $(T_0, x(\cdot)) \in \mathfrak{S}$, the existence of a solution of (1.1) is guaranteed by either of the two following conditions (we shall refer to the corresponding results as Theorem I and Theorem II, respectively):

I. $$H(t, x, p) \leqq \mu(t, |p|) + |x|(\sigma(t) + \rho(t)|p|),$$

$$l(x, z) \geqq l_0(x) + l_1(z), \tag{1.4}$$

where $\sigma(t)$, $\rho(t)$ and $\mu(t, v)$ are finite and summable as functions of t (for any $v \geqq 0$) $\sigma(t) \geqq 0$, $\rho(t) \geqq 0$, l_0 and l_1 are everywhere $> -\infty$ and

$$\liminf_{|x| \to \infty} \frac{l_0(x)}{|x|} = \infty, \qquad \liminf_{|z| \to \infty} \frac{l_1(z)}{|z|} > -\infty. \tag{1.5}$$

II. $$L(t, x, y) \geqq \varphi(|y|) - \psi(|x|) + a(t), \tag{1.6}$$

$$l(x, z) = \begin{cases} 0 & \text{if } x = x_0, z = x_1, \\ \infty & \text{otherwise,} \end{cases} \tag{1.7}$$

where φ and ψ are finite, nonnegative and nondecreasing functions on $[0, \infty)$, φ is convex, $a(t)$ is a summable function on $[0, T_0]$ and

$$\lim_{r \to \infty} r^{-1} \varphi(r) = \infty, \tag{1.8}$$

$$\lim_{r \to \infty} \left[\varphi\left(\frac{2r}{T_0}\right) - \psi(r + \max(|x_0|, |x_1|)) \right] = \infty. \tag{1.9}$$

The first result is proved by Rockafellar [8] and generalizes well-known theorems of Olech [6], who in turn was inspired by Cesari's works [1]. The second result is contained in the author's paper [4], though a very similar result for a classical variational problem was proved much earlier by Cinquini [2], as was pointed out in [3].

It is easy to see that the two quoted results are different. Many problems can be found to which one of the results can suitably be applied but not the other. Indeed, in the situation considered in Theorem I, the integrand L cannot decrease in x (for t and y fixed) faster than a linear function. Theorem II cannot be applied to the Bolza problem, and in addition it imposes rather rigid requirements on the

behavior of L as a function of t, though L is allowed to decrease in x at an arbitrary rate.

In § 4, we show how these results follow from our main theorem, proved in § 3. The general growth condition used in our theorem may be regarded, to some degree, as a combination of the growth conditions used by Cinquini and the author on the one hand and by Olech and Rockafellar on the other hand.

2. Definition and auxiliary results. Let $(T, x(\cdot)) \in \mathfrak{S}$. We shall usually extend $x(t)$ outside of $[0, T]$ as follows:

$$x(t) = x(T), \quad \text{if } t > T. \tag{2.1}$$

Thus $x(\cdot)$ is absolutely continuous on the whole half-line $R_+ = [0, \infty)$. We denote by $H^n_{1,1}(T)$ the Banach space of all absolutely continuous mappings from $[0, T]$ into R^n with the norm

$$\|x(\cdot)\|_{1,1} = |x(0)| + \int_0^T |\dot{x}(t)|\, dt, \tag{2.2}$$

where $|x|$ denotes the Euclidean norm of x. The Banach space of all continuous mappings $x(\cdot) : [0, T] \to R^n$ with the usual uniform norm

$$\|x(\cdot)\|_C = \max_{0 \leqq t \leqq T} |x(t)|$$

is denoted by $C^n(T)$.

Now we shall define convergence in $\mathfrak{S}$. We say that the sequence $\{(T_k, x_k(\cdot))\} \subset \mathfrak{S}$ converges to $(T, x(\cdot)) \in \mathfrak{S}$ if $T_k \to T$ and $x_k(\cdot) \to x(\cdot)$ weakly in $H^n_{1,1}(T)$ ($x_k(\cdot)$ being extended according to (2.1)).

Let T^* be the upper bound of those $t \geqq 0$ which satisfy

$$\inf_{x,z} l(t, x, z) < \infty. \tag{2.3}$$

It is reasonable to postulate that $T^* > 0$.

We say, following Rockafellar [8], that the integrand L satisfies the *basic growth condition* if, for any bounded $S \subset R^n$ and any $p \in R^n$, there exists a measurable function $\psi_p(t)$ defined on $[0, \infty)$ summable on every finite interval and satisfying

$$H(t, x, p) \leqq \psi_p(t) \tag{2.4}$$

for all $x \in S$ and almost all $t \in [0, T^*]$.

Let

$$I_1(T, x(\cdot)) = \int_0^T L(t, x(t), \dot{x}(t))\, dt.$$

PROPOSITION 1. *Let L satisfy the assumptions* (i)–(ii) *and the basic growth condition. Then for every real α and positive $\tau \leqq T^*$ and r the set*

$$\mathfrak{S}(\tau, r, \alpha) = \{(T, x(\cdot)) \in \mathfrak{S} \mid T \leqq \tau, \|x(\cdot)\|_C \leqq r, I_1(T, x(\cdot)) \leqq \alpha\}$$

is sequentially compact with respect to the convergence in $\mathfrak{S}$.

Proof. Let

$$\tilde{H}(t,x,p) = \max(0, H(t,x,p)),$$
$$\tilde{L}(t,x,y) = \sup_p (\langle p, y\rangle - \tilde{H}(t,x,p)).$$

Then $\tilde{L}$ obviously satisfies properties (i)–(iii) and also satisfies the basic growth condition, with $\tilde{\psi}_p(t) = \max(0, \psi_p(t))$. Moreover, for all (t,x,y), one has

$$\tilde{L}(t,x,y) \leqq L(t,x,y) \tag{2.5}$$

and

$$\tilde{L}(t,x,0) \leqq 0. \tag{2.6}$$

For any pair $(T, x(\cdot)) \in \mathfrak{S}(\tau, r, \alpha)$, we have from (2.5), (2.6), that

$$\int_0^\tau \tilde{L}(t,x(t),\dot{x}(t))\,dt = \int_0^T \tilde{L}(t,x(t),\dot{x}(t))\,dt + \int_T^\tau \tilde{L}(t,x(T),0)\,dt$$
$$\leqq \int_0^T L(t,x(t),\dot{x}(t))\,dt \leqq \alpha.$$

Making use of the semicontinuity theorem proved in [8], we get that the set

$$\left\{ x(\cdot) \in H^n_{1,1}(\tau) \,\middle|\, \int_0^\tau \tilde{L}(t,x(t),\dot{x}(t))\,dt \leqq \alpha,\ \|x(\cdot)\|_C \leqq r \right\}$$

is compact in the weak topology of $H^n_{1,1}(\tau)$. In $H^n_{1,1}$, weak compactness implies sequential weak compactness. (Indeed, any weak compact set in $H^n_{1,1}(\tau)$ is also norm compact in $C^n(\tau)$; and if $x_k(\cdot) \to x(\cdot)$ in the norm topology of $C^n(\tau)$ and $\|x_k(\cdot)\|_{1,1} \leqq N < \infty$ for all k, then $\dot{x}_k(\cdot)$ converges weakly to $x(\cdot)$ in L_1, and hence $x_k(\cdot) \to x(\cdot)$ weakly in $H^n_{1,1}(\tau)$.)

If now $(T_k, x_k(\cdot)) \in \mathfrak{S}(\tau, r, \alpha)$ $(k = 1, 2, \cdots)$, then there exists a subsequence $(T_{k_s}, x_{k_s}(\cdot))$ such that $T_{k_s} \to T_0$ and $x_{k_s}(\cdot) \to x_0(\cdot)$ weakly in $H^n_{1,1}(\cdot)$. Obviously $\|x_0(\cdot)\|_C \leqq r$, $T_0 \leqq \tau$, and $x_0(t)$ is constant outside of $[0, T_0]$.

Choose $\psi_p(t)$ according to the basic growth condition and let $S = \{x \in R^n \mid |x| \leqq r\}$. Then, in particular,

$$L(t,x,y) \geqq -\psi_0(t) \quad \text{if } x \in S, \quad t \in [0, T_0], \quad y \in R^n. \tag{2.7}$$

If $T_0 = 0$, then there is nothing to prove, inasmuch as

$$\liminf I_1(T_k, x_k(\cdot)) \geqq \lim \int_0^{T_k} (-\psi_0(t))\,dt = 0.$$

If $T_0 > 0$, then for any $\varepsilon > 0$, $x_{k_s}(\cdot) \to x_0(\cdot)$ weakly in $H^n_{1,1}(T_0 - \varepsilon)$, and because of lower semicontinuity of the integral functional,

$$x(\cdot) \to \int_0^{T_0 - \varepsilon} L(t,x(t),\dot{x}(t))\,dt$$

with respect to the weak topology of $H^n_{1,1}(T_0 - \varepsilon)$ (see, for instance, the same

semicontinuity theorem in [8]) we have

$$\begin{aligned}\alpha &\geqq \liminf \int_0^{T_{k_s}} L(t, x_{k_s}(t), \dot{x}_{k_s}(t))\, dt \\ &\geqq \liminf \int_0^{T_0-\varepsilon} L(t, x_{k_s}(t), \dot{x}_{k_s}(t))\, dt - \lim \int_{T_0-\varepsilon}^{T_{k_s}} \psi_0(t)\, dt \\ &\geqq \int_0^{T_0-\varepsilon} L(t, x_0(t), \dot{x}_0(t))\, dt - \int_{T_0-\varepsilon}^{T_0} \psi_0(t)\, dt.\end{aligned}$$

Since $\psi_0(\cdot)$ is summable, it follows that

$$\alpha \geqq \lim_{\varepsilon \to 0} \int_0^{T_0-\varepsilon} L(t, x_0(t), \dot{x}_0(t))\, dt = I_1(T_0, x_0(\cdot)),$$

and hence $(T_0, x_0(\cdot)) \in \mathfrak{S}(\tau, r, \alpha)$. This completes the proof.

Let $g(t, r, w)$ be an extended-real-valued function on $R_+ \times R_+ \times R_+$ ($R_+ = [0, \infty)$). We set

$$G(T, r, w) = \int_0^T g(t, r, w)\, dt \quad \text{(if this exists)},$$

$$G^*(T, r, v) = \sup_{w \geqq 0} (wv - G(T, r, v)),$$

$$F(t, r, v) = \min \{G^*(T, r, v - \min(|x|, |z|)) + l(T, x, z)| \\ x, z \in R^n, \max(|x|, |z|) \leqq v\}.$$

We say that the functional $I(T, x(\cdot))$ satisfies the *general growth condition* if there exists an extended-real-valued function $g(t, r, w)$ on $R_+ \times R_+ \times R_+$ and a nonnegative function $\rho(t)$ on R_+ such that (α) g is $\mathscr{L} \otimes \mathscr{B}$ measurable; for every fixed $r \geqq 0$, $w \geqq 0$, the function $t \to g(t, r, w)$, as well as $\rho(t)$, is summable on every finite interval;

(β) $H(t, x, p) \leqq g(t, |x|, |p|) + \rho(t)|x|\,|p|$ if $0 \leqq t \leqq T^*$, x, p arbitrary;

(γ) $g(t, r, w)$ is nondecreasing in r;

(δ) $F(T, r, \lambda(T)r) \to \infty$ as $T \to \infty$, $r \to \infty$,

where

$$\lambda(T) = \exp\left(-\int_0^T \rho(t)\, dt\right).$$

PROPOSITION 2. *Let $L(t, x, y)$ be $\mathscr{L} \otimes \mathscr{B}$-measurable. Then the basic growth condition holds if and only if there exist $g(t, r, w)$ and $\rho(t)$ satisfying assumptions (α)–(γ) of the general growth condition.*

Proof. Let (α)–(γ) be satisfied. Then given a bounded set $S \subset R^n$, we set

$$\psi_p(t) = g(t, r, |p|) + \rho(t)|p|,$$

where $r = \sup \{|x| \mid x \in S\}$.

Conversely, let the basic growth condition hold. Then (see [8, Prop. 5]) the function

$$h(t, r, w) = \sup \{H(t, x, p) \mid |x| \leqq r, |p| \leqq w\}$$

is $\mathscr{L} \otimes \mathscr{B}$-measurable, nondecreasing in (r, w), convex in w, lower semicontinuous in (r, w), and satisfies

$$\int_0^T h(t, r, w)\, dt < \infty \quad \text{for all } r \geqq 0, \quad w \geqq 0, \quad T < \infty, \quad T \leqq T^*.$$

Let $g(t, r, w) = \max(0, h(t, r, w))$. Then, as may easily be seen, (α)–(γ) holds.

3. Main theorem. *Let I satisfy properties* (i)–(iv) *stated in* § 1. *Assume further that I satisfies the general growth condition. Then the level sets of I are weakly sequentially compact in $\mathfrak{S}$. In particular, if $I(T, x(\cdot)) < \infty$ for at least one pair $(T, x(\cdot)) \in \mathfrak{S}$, then the problem* (1.1) *has a solution.*

Proof. Inasmuch as $F(T, r, \lambda(T)r) \to \infty$ as $T \to \infty$ and $r \to \infty$ for any given $\alpha \in R$, there exist T_α and r_α such that $T \leqq T_\alpha$, $r \leqq r_\alpha$, whenever $F(T, r, \lambda(T)v) \leqq \alpha$. We shall show that the inequalities

$$T \leqq T_\alpha, \quad \|x(\cdot)\|_C \leqq r_\alpha, \tag{3.1}$$

hold for every pair $(T, x(\cdot)) \in \mathfrak{S}$ such that $I(T, x(\cdot)) \leqq \alpha$. The inequalities (3.1), along with Proposition 1, imply weak sequential compactness of all the level sets

$$\operatorname{lev}_\alpha I = \{(T, x(\cdot)) \in \mathfrak{S} \mid I(T, x(\cdot)) \leqq \alpha\}.$$

Let α_0 be the infimum of I on $\mathfrak{S}$. According to the hypothesis, $\alpha_0 < \infty$. Hence for every $\alpha > \alpha_0$, the level set $\operatorname{lev}_\alpha I$ is nonempty and

$$\operatorname{lev}_{\alpha_0} I = \bigcap_{K=1}^{\infty} \operatorname{lev}_{\alpha_0 + (1/k)} I \neq \varnothing,$$

as the intersection of a decreasing countable family of nonempty sequentially compact sets. Each element of $\operatorname{lev}_{\alpha_0} I$ is a solution of the problem, by definition. Therefore, it remains only to establish (3.1).

Let $(T, x(\cdot)) \in \mathfrak{S}$ and $I(T, x(\cdot)) \leqq \alpha$. Let $r = \|x(\cdot)\|_C$. We have

$$\begin{aligned}
\int_0^T L(t, x(t), \dot{x}(t))\, dt &= \int_0^T \sup_p (\langle p, \dot{x}(t)\rangle - H(t, x(t), p))\, dt \\
&\geqq \int_0^T \sup_p [\langle p, \dot{x}(t)\rangle - g(t, |x(t)|, |p|) - \rho(t)|x(t)|\,|p|]\, dt \\
&\quad \text{(in view of } (\gamma)) \\
&\geqq \sup_p \int_0^T [|p|(|\dot{x}(t)| - \rho(t)|x(t)|) - g(t, r, |p|)]\, dt.
\end{aligned}$$

Let

$$\omega(t) = \max(0, |\dot{x}(t)| - \rho(t)|x(t)|).$$

Then

$$\int_0^T L(t, x(t), \dot{x}(t))\, dt \geqq \sup_{w \geqq 0} \int_0^T [w\omega(t) - g(t, r, w)]\, dt = G^*\left(T, r, \int_0^T \omega(t)\, dt\right). \tag{3.2}$$

It follows from the proof of the second existence theorem in [8] that

$$\int_0^T \omega(t)\, dt \geqq \lambda(T)r - |x(0)|.$$

Similarly,

$$\int_0^T \omega(t)\, dt \geqq \lambda(T)r - |x(T)|,$$

that is,

$$\int_0^T \omega(t)\, dt \geqq \lambda(T)r - \min(|x(0)|, |x(T)|). \tag{3.3}$$

The function $v \to G^*(T, r, v)$ is nondecreasing in v. Indeed, if $v' \geqq v \geqq 0$, then

$$\begin{aligned} G^*(T, r, v') &= \sup_{w \geqq 0} (wv' - G(T, r, w)) \\ &\geqq \sup_{w \geqq 0} (wv - G(T, r, w)) = G^*(T, r, v). \end{aligned} \tag{3.4}$$

Making use of (3.2)–(3.4), we get

$$\begin{aligned} \alpha &\geqq \int_0^T L(t, x(t), \dot{x}(t))\, dt + l(T, x(0), x(T)) \\ &\geqq G^*(T, r, \lambda(T)r - \min(|x(0)|, |x(1)|)) + l(T, x(0), x(T)) \end{aligned}$$

which implies (3.1).

4. Applications. We show here how the results stated in the Introduction can be deduced from our main theorem.

Proof of Theorem I. Let

$$g(t, r, w) = \mu(t, w) + r\sigma(t).$$

Then (α)–(γ) are obviously satisfied. Let

$$M(T, w) = \int_0^T \mu(t, w)\, dt, \qquad M^*(t, v) = \sup_{w \geqq 0} (wv - M(T, w)).$$

Since M is everywhere finite,

$$\lim_{v \to \infty} v^{-1} M^*(T, v) = \infty. \tag{4.1}$$

We have

$$G(T, r, w) = M(T, w) + rk(T),$$

where

$$k(T) = \int_0^T \sigma(t)\,dt,$$

$$G^*(T, r, v) = M^*(T, v) - rk(T).$$

It follows from (1.2) that F can be finite only if $T = T_0$. We have

$$F(r, v) = F(T_0, r, v) = \min \{M^*(T_0, v - \min(|x|, |z|)) + l_0(x) + l_1(z)| \\ x, z \in R^n, \max(|x|, |z|) \leqq v\} - rk(T).$$

Suppose that the minimum is attained in (x_v, z_v), and let $\lambda = \lambda(T_0)$. We have $\lambda > 0$ and

$$2F(r, \lambda r) \geqq \left[M^*(T_0, \lambda r - |x_r|) - (\lambda r - |x_r|)\frac{k(T_0)}{\lambda}\right] \\ + \left[2l_0(x_r) - |x_r|\frac{k(T_0)}{\lambda}\right] \\ + \left[M^*(T_0, \lambda r - |z_r|) - (\lambda r - |z_r|)\frac{k(T_0)}{\lambda}\right] \\ + \left[2l_1(z_r) - |z_r|\frac{k(T_0)}{\lambda}\right].$$

If $r \to \infty$, then either $\lambda r - |x_r| \to \infty$ or $|x_r| \to \infty$. In the first case, the first member of the sum tends to ∞ (because of (4.1)) while the other remains bounded below (because of (1.5) and (4.1)). Analogously, in the second case, the second member of the sum tends to ∞, while the other remains bounded below. In either case, $F(r, \lambda r) \to \infty$, and hence the last requirement (δ) of the general growth condition is satisfied. It remains only to apply the main theorem.

Proof of Theorem II. This proof is not so direct, because of the coefficient 2 in (1.9). Let us consider two auxiliary problems:

$$(P_z') \quad \text{minimize } I'(x(\cdot)) = \int_0^{T_0/2} L(t, x(t), \dot{x}(t))\,dt$$

subject to

$$x(0) = x_0, \qquad x(T_0/2) = z;$$

and

$$(P_z'') \quad \text{minimize } I''(x(\cdot)) = \int_{T_0/2}^{T_0} L(t, x(t), \dot{x}(t))\,dt$$

subject to

$$x(T_0/2) = z, \quad x(T_0) = x_1.$$

We shall actually deal only with the first of these. The second is treated in the same manner.

Let

$$g(t, r, w) = \varphi^*(w) + \psi(r) - a(t),$$

where φ^* is the Fenchel conjugate to φ. Then $g(t, r, w)$ is summable for all (r, w) by virtue of (1.8), and

$$G(r, w) = (T_0/2)(\varphi^*(w) + \psi(r)) - a',$$

where

$$a' = \int_0^{T_0/2} a(t)\, dt,$$

$$G^*(r, v) = (T_0/2)\left(\varphi\left(\frac{2v}{T_0}\right) - \psi(r)\right) + a'.$$

It follows from (1.6) and (1.3) that

$$H(t, x, p) \leqq g(t, |x|, |p|);$$

hence (α)–(γ) are satisfied with $\rho(t) \equiv 0$ and $\lambda(T_0/2) = 1$. Since $\min(|x_0|, |z|) \leqq \max(|x_0|, |x_1|)$, it follows from (1.9) that the expression

$$F(r, r) = \varphi\left(\frac{2}{T_0}(r - \min(|x_0|, |z|))\right) - \psi(r) + a'$$

tends to ∞ as $r \to \infty$. Hence for any z, (P_z') has a solution by the theorem. The same arguments show that (P_z'') has a solution as well. Let $x_z'(\cdot)$ and $x_z''(\cdot)$ be solutions of (P_z') and (P_z''), respectively. Since the level sets of I' and I'' are compact by the theorem, there exists z_0 such that

$$I'(x_{z_0}'(\cdot)) + I''(x_{z_0}''(\cdot)) = \min(I'(x_z'(\cdot)) + I''(x_z''(\cdot))). \tag{4.2}$$

Let

$$x_0(t) = \begin{cases} x_{z_0}'(t) & \text{if } 0 \leqq t \leqq T_0/2, \\ x_{z_0}''(t) & \text{if } T_0/2 \leqq t \leqq T_0. \end{cases}$$

Then (4.2) shows that $x_0(t)$ is a solution to the original problem.

REFERENCES

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

[2] S. CINQUINI, *Sopra l'esistenza della solusione nei problemi di calcolo delle variazioni di ordine n*, Ann. Scuola Norm. Sup. Pisa, 5 (1936), pp. 169–190.

[3] ———, *A proposito della esistenza dell'estremo assoluto in campi illimitati*, Rend. Ist. Lombardo di Sci Lettere, 107 (1973), pp. 460–472.

[4] A. D. IOFFE, *An existence theorem for problems of the calculus of variations*, Dokl. Akad. Nauk USSR, 205 (1972), pp. 277–280 = Soviet Math. Dokl., 13 (1972), pp. 919–923.

[5] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of External Problems*, Nauka, Moscow, 1974. (In Russian.)

[6] C. OLECH, *Existence theorems for optimal problems with vector-valued cost functions*, Trans. Amer. Math. Soc., 136 (1969), pp. 157–180.

[7] R. T. ROCKAFELLAR, *Optimal arcs and the minimum value function in problems of Lagrange*, Ibid., 180 (1973), pp. 53–84.

[8] ———, *Existence theorems for general control problems of Bolza and Lagrange*, Advances in Math., to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

THE CLASSIFICATION OF LINEAR STATE VARIABLE CONTROL LAWS*

BRADLEY W. DICKINSON†

Abstract. Two fundamental classes of control laws for linear time-invariant systems were introduced by Kalman [4]. Purely feed-forward control laws do not alter the open loop eigenvalues and purely feedback control laws do not alter the cyclic structure of the open loop system matrix. Here the decomposition of an arbitrary control law into the sum of three laws, two from one of the classes and one from the other, is obtained. The uniqueness of the decomposition is studied. The notion of a covariant control law is introduced to give a decomposition of control laws related to the invariant description of reachable linear systems given by Popov [9]. Two applications of covariant control laws are illustrated, including their use in obtaining maximally unobservable canonical forms for linear multivariable systems under an equivalence relation induced by control laws and state-space basis transformations.

1. Introduction. Let (A, B) be a linear time-invariant system representation defined over an arbitrary field $\mathbf{k}$, described by a set of difference equations

$$x_{t+1} = Ax_t + Bu_t, \tag{1.1}$$

where $x \in \mathbf{k}^n$ is the state vector, $u \in \mathbf{k}^m$ is the control vector, and $t \in \mathbf{Z}$, the integers. Thus A and B are $n \times n$ and $n \times m$ $\mathbf{k}$-matrices respectively.

Frequently, the input in (1.1) is chosen as a linear function of the state

$$u_t = Kx_t, \tag{1.2}$$

where K is an $m \times n$ $\mathbf{k}$-matrix; we call such a matrix K a *linear state variable control law*, or more simply a *control law*, for (A, B). To facilitate the study of (A, B) subjected to a control law K, we will assume that (A, B) is a *reachable* system; that is, we assume the matrix $[B, AB, \cdots, A^{n-1}B]$ has rank n. For notational simplicity, the matrix B is assumed to have rank m.

This paper explores the structure of linear state variable control laws. Rosenbrock's control structure theorem is reviewed as the principal tool for establishing existence of certain control laws. Some important classes of control laws, each preserving various structural properties of (A, B), are defined. Some of these classes of control laws were originally studied by Kalman [4], although we have refined some definitions in order to obtain stronger results. A new class of control laws, called covariant control laws, is introduced to study the interaction of control laws and state-space basis transformations; this idea is rooted in the work of Popov [9] on invariant descriptions of reachable systems. Our structural results in § 3 describe the decomposition of an arbitrary control law into constituent parts from various classes of control laws.

* Received by the editors October 8, 1974, and in revised form April 29, 1975.

† Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08540. This research was supported in part by the National Science Foundation under Grant ENG75-10533 and in part at Stanford University by the Air Force Office of Scientific Research AF Systems Command, under Contract AF-44-620-74-C-0068 and by the Joint Services Electronics Program, under Contract N-00014-67-A-0112-0044, through Professor Thomas Kailath.

In § 4 we give two examples to illustrate these results. Two applications of the class of covariant control laws are discussed in § 5. We show that arbitrary pole assignment can always be accomplished with a covariant control law. Finally, we consider an equivalence relation on the set of reachable linear systems induced by control laws and state-space basis transformations. Covariant control laws are used to transform a particular set of canonical forms to a second set of maximally unobservable canonical forms.

2. Control laws and invariant factors. We first review Rosenbrock's control structure theorem [10, pp. 190–192]. The theorem describes how the *open loop system matrix*, A, can be modified by a control law K. Let $\phi_1(z), \cdots, \phi_q(z)$ be the *invariant factors* of the *closed loop system matrix* $A+BK$. That is, they are the nonunit invariant polynomials of the polynomial matrix $zI-A-BK$, ordered so that

$$\phi_{i+1}(z) \mid \phi_i(z), \qquad 1 \leqq i \leqq q-1, \tag{2.1}$$

(reading $\phi_{i+1}(z)$ divides $\phi_i(z)$). See [5] for further discussion of the significance of the invariant factors. In addition, there is an ordered list of positive integers $\nu_1 \geqq \nu_2 \geqq \cdots \geqq \nu_m$, whose sum is n, that can be uniquely associated with $(A+BK, B)$, and this list of *controllability indices* is independent of the choice of K[1], [17]. Rosenbrock's theorem shows that the controllability indices provide bounds on the degrees of the invariant factors of the closed loop system matrix.

THEOREM 2.1. *Let (A, B) be a reachable system with controllability indices $\nu_1 \geqq \nu_2 \geqq \cdots \geqq \nu_m$. Let $\{\phi_1(z), \cdots, \phi_q(z)\}$ be any set of monic polynomials in $\mathbf{k}[z]$ satisfying the divisibility properties in* (2.1), *and let $q \leqq m$. Then there is a control law K such that the given polynomials are the invariant factors of $A+BK$ if and only if*

$$\sum_{i=1}^{r} \deg \phi_i(z) \geqq \sum_{i=1}^{r} \nu_i, \qquad 1 \leqq r \leqq q. \tag{2.2}$$

This slight modification of Rosenbrock's statement [10, p. 192] follows from the ordering of the controllability indices and the fact

$$\sum_{i=1}^{q} \deg \phi_i(z) = n = \sum_{i=1}^{m} \nu_i. \tag{2.3}$$

Some further discussion of this result can be found in [2] and [4].

Kalman made the first observations on the classification of control laws that will be examined here. Certain control laws preserve key properties of the system matrices of the open and closed loop systems.

DEFINITIONS 2.1 (Kalman [4]). A control law J is called *purely feedforward* (relative to (A, B)) if the open and closed loop system matrices, in this case A and $A+BJ$, respectively, have the same characteristic polynomial. A control law L is called *purely feedback* (relative to (A, B)) if the ordered list of degrees of the invariant factors of the open and closed loop system matrices are the same. A control law M is called *neutral* if it is both purely feedforward and purely feedback.

In short, purely feedforward control laws preserve eigenvalues. The control theoretic significance of purely feedback control laws is less obvious. Certainly the

degrees of the minimal polynomials of the open and closed loop system matrices are equal when a purely feedback control law is used. More generally, the dimensions of the cyclic components of the underlying state module [5] remain fixed.

A natural question that arises at this point concerns the decomposition of an arbitrary control law into component parts. Kalman's simple additive decomposition [4] is not generally valid. For a simple counterexample, illustrated in § 4, the system (A, B) is taken over the field $\mathbf{Q}$ of rational numbers, with $n = 6$, $m = 2$. The controllability indices are chosen to be $\nu_1 = \nu_2 = 3$, and the invariant factors are chosen to be $\phi_1(z) = \phi_2(z) = z^3 + 2$. By Theorem 2.1, there is a control law K giving a closed loop system matrix with invariant factors $\hat{\phi}_1(z) = (z^2+1)^2$ and $\hat{\phi}_2(z) = z^2 + 1$. Because the roots of $\phi_2(z)$ and $\hat{\phi}_2(z)$ are algebraically independent over $\mathbf{Q}$, there can be no simple decomposition $K = J + L$, where J is purely feedforward and L is purely feedback.

The Appendix discusses a counterexample for the case of any field that is not algebraically closed; this includes the real number field and all finite fields which are the cases of most practical significance.

3. Decompositions of control laws. There are two natural decompositions of a control law that each involve *three* summands. It is interesting to note that Kalman actually used one of these when assuming that certain eigenvalues were normalized to zero [4], but then neglected to account for it in his decomposition. An additional definition facilitates a sharp description of the uniqueness of the decomposition.

DEFINITION 3.1. A control law N is called *strongly neutral* (relative to (A, B)) if the invariant factors of A and $A + BN$ are identical.

THEOREM 3.1. *Let (A, B) be a reachable system and let K be any control law. Then*

(a) *$K = L_1 + J - L_2$, where L_1 is purely feedback relative to (A, B), $A + BL_1$ is nilpotent, J is purely feedforward relative to $(A + BL_1, B)$ and L_2 is purely feedback relative to $(A + BK, B)$;*

(b) *$K = J_1 + L - J_2$, where J_1 is purely feedforward relative to (A, B), $A + BJ_1$ is cyclic, L is purely feedback relative to $(A + BJ, B)$, and J_2 is purely feedforward relative to $(A + BK, B)$. All the component laws are unique up to an appropriate strongly neutral law.*

Proof. For (a), let L_1 be a control law shifting all closed loop eigenvalues to zero and preserving the degrees of the invariant factors. Then let J be a control law that preserves the characteristic polynomial of $A + BL_1$ while giving invariant factors whose degrees are equal to those of $A + BK$. Then the control law $L_2 = L_1 + J - K$ is purely feedback relative to $(A + BK, B)$ as required. For (b), let J_1 be a control law that preserves the characteristic polynomial of A and makes $A + BJ_1$ *cyclic*; that is, its characteristic polynomial is its only invariant factor. Let L be a control law that preserves cyclicity and makes the characteristic polynomial of $A + BJ_1 + BL$ equal to that of $A + BK$. Then $J_2 = J_1 + L - K$ is purely feedforward relative to $(A + BK, B)$ as required. Clearly each law is only unique up to an appropriate strongly neutral law. The existence of all the control laws in these decompositions follows from Theorem 2.1. □

Let N be a strongly neutral control law relative to (A, B). Because A and $A+BN$ have the same invariant factors, there is a nonsingular matrix T so that

$$T^{-1}AT = A + BN. \tag{3.1}$$

It may also be the case that

$$T^{-1}B = B; \tag{3.2}$$

(see the example in § 4). When this is true, the closed loop and open loop systems differ only by a change in state-space basis. This remarkable situation indicates that a closer examination of control laws and their interaction with basis changes in the underlying state-space of a reachable system (A, B) is warranted.

One approach to this topic follows Popov's results [9] on a complete set of independent invariants for the set of reachable systems (A, B) acted on by the group of state-space basis transformations.

DEFINITION 3.2. Two systems, (A_1, B_1) and (A_2, B_2), are said to be *state equivalent*, or *similar*, if there is a nonsingular matrix T such that

$$T^{-1}A_1T = A_2, \qquad T^{-1}B_1 = B_2. \tag{3.3}$$

By reachability, the matrix T in (3.3) is unique; see [18].

Popov [9] describes a complete set of independent invariants for this equivalence relation. His result can be used to construct various sets of canonical forms for systems under similarity, each set containing one representative system from each similarity equivalence class [13].

One particular set of canonical forms, here called *s-canonical forms*, is obtained by using a procedure of Popov [9, Thm. 2]. The s-canonical form of a system (A, B) is defined as $(T^{-1}AT, T^{-1}B)$, where the matrix T (Popov's matrix M) is uniquely determined by (A, B). The columns of T are particular linear combinations of the first set of linearly independent columns obtained by examining, in lexicographic order, the columns of the matrix $[B, AB, \cdots, A^{n-1}B]$. By reachability, we see that a system (A, B) is in s-canonical form if and only if the corresponding matrix T is the identity matrix.

We will briefly examine the structure of the s-canonical forms. If (A, B) is in s-canonical form, then there is a permutation $\pi(\cdot)$ of the first m positive integers defining a reordered set of controllability indices $n_1, n_2, \cdots, n_m$, where

$$n_i = \nu_{\pi(i)} \tag{3.4}$$

and, furthermore,

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & & \vdots \\ A_{m1} & \cdots & A_{mm} \end{bmatrix}, \qquad B = \begin{bmatrix} B_1 \\ \vdots \\ B_m \end{bmatrix}, \tag{3.5}$$

where A_{ij} is $n_i \times n_j$ and B_i is $n_i \times m$. Using "x" to denote a possibly nonzero entry,

the blocks have the forms

$$A_{ii} = \begin{bmatrix} 01 & & 0 \\ & \ddots & \\ & & 1 \\ xx & \cdots & x \end{bmatrix},$$

$$A_{ij} = \begin{bmatrix} & & 0 & & \\ x, & x & \cdots & x & 0 \cdots 0 \end{bmatrix}, \qquad i \neq j. \tag{3.6}$$

For $i \neq j$, the last row of A_{ij} can have possibly nonzero entries only in the first min (n_i, n_j)-columns. The matrix B_i has all zero entries except that in row n_i it has a 1 in column i and an "x" in every column $j > i$ for which $n_j < n_i$.

Some observations, following generally from Popov [9, Thms. 2 and 3], should also be made. If $(\hat{A}_1, \hat{B}_1)$ is the s-canonical form of (A, B) and if $(\hat{A}_2, \hat{B}_2)$ is the s-canonical form of $(A + BK, B)$, then $\hat{B}_1 = \hat{B}_2$. Thus control laws preserve the canonical form of the input matrix. Similarly, the ordering of the indices $\{n_i\}$ is preserved. If (A, B) is in s-canonical form, a control law K can be chosen to place any desired elements of $\mathbf{k}$ in the entries denoted by "x" in (3.6). However, some control laws will not leave the resulting closed loop system in s-canonical form because they alter some "sacred zeros" in one or more of the blocks A_{ij} for $i \neq j$. We give a name to the special class of control laws that preserve s-canonical form.

DEFINITION 3.3. Let (A, B) be a reachable system and let $(T^{-1}AT, T^{-1}B)$ be its s-canonical form. A control law K_c is called *covariant* (relative to (A, B)) if $(T^{-1}(A + BK)T, T^{-1}B)$ is in s-canonical form.

The name covariant is chosen to indicate that changes occur in the control law when the basis is changed in the state-space of the system. For example, if (A_1, B_1) and (A_2, B_2) are given by (3.3) and K_1 is covariant relative to (A_1, B_1), then

$$K_2 = K_1 T \tag{3.7}$$

is covariant relative to (A_2, B_2). Notice that when all the controllability indices are equal, every control law is covariant because there are no "sacred zeros" to alter.

The corresponding decomposition of an arbitrary control law can now be given.

THEOREM 3.2. *Let (A, B) be a reachable system and let K be any control law. Then $K = K_c - N$, where K_c is covariant relative to (A, B) and N is strongly neutral relative to $(A + BK, B)$. The decomposition is unique up to a law that is both covariant and strongly neutral.*

Proof. Let $(T^{-1}AT, T^{-1}B)$ be in s-canonical form and let $(\hat{A}, \hat{B})$ be the s-canonical form of $(A + BK, B)$. Then $T^{-1}B = \hat{B}$ and $T^{-1}AT + T^{-1}B\hat{K} = \hat{A}$ for some control law $\hat{K}$ that is covariant relative to $(T^{-1}AT, T^{-1}B)$. Thus $K_c = \hat{K}T^{-1}$ is covariant relative to (A, B). Let $N = K_c - K$. Since the systems $(A + BK_c, B)$ and $(A + BK, B)$ have the same s-canonical form, $A + BK_c$ and $A + BK$ are similar and so have the same set of invariant factors. Thus N is strongly neutral relative to $(A + BK, B)$. Since a control law may be both covariant and strongly neutral, as an example in § 4 shows, the uniqueness result follows directly. □

4. Examples. We first consider the system defined over Q, the rational numbers, used in the counterexample of § 2:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -2 & 0 & 0 \end{bmatrix}; \qquad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The control law K,

$$K = \begin{bmatrix} 2 & -1 & 0 & 1 & 0 & 1 \\ -1 & 0 & -1 & 2 & -1 & 0 \end{bmatrix},$$

gives a closed loop system matrix with invariant factors $(z^2+1)^2$ and (z^2+1). Its decompositions are

(a)
$$K = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}$$
$$- \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & -1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} = L_1 + J - L_2,$$

(b)
$$\mathrm{K} = \begin{bmatrix} 2 & 0 & 0 & 1 & 0 & 0 \\ -4 & 0 & 0 & -2 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & -3 & 2 & -3 & 0 \end{bmatrix}$$
$$- \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & -2 & -2 & -2 & 0 \end{bmatrix} = \mathrm{J}_1 + \mathrm{L} - \mathrm{J}_2.$$

As an example of nonuniqueness, notice that in (a), J may be replaced by any nonzero multiple of J with L_2 then being modified accordingly.

As a second example, consider a system defined over the real number field:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -2 & -3 & 0 \\ 0 & 0 & -3 \end{bmatrix}; \qquad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Notice that (A, B) is in s-canonical form.

The control law K,

$$K = \begin{bmatrix} 0 & 0 & 0 \\ 3 & k & 0 \end{bmatrix}, \qquad k \neq 0,$$

is not covariant, but it is strongly neutral. Thus it has the trivial decomposition $K = 0 + K$.

However, we also have

$$K = \begin{bmatrix} 0 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & -k & 0 \end{bmatrix} = K_c - N,$$

where K_c is both strongly neutral and covariant. Now the s-canonical form of $(A+BK, B)$ is $(\hat{A}, \hat{B})$,

$$\hat{A} = \begin{bmatrix} 0 & 1 & 0 \\ -2 & -3 & 0 \\ 3(1-k) & 0 & -3 \end{bmatrix}; \qquad \hat{B} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix};$$

so that when $k=1$, $(A+BK, B)$ and (A, B) are similar; that is, there is a nonsingular matrix T so that

$$T^{-1}(A+BK)T = A,$$

$$T^{-1}B = B.$$

5. Applications. The classes of control laws introduced by Kalman [4] are significant because of the control theoretic interpretations that apply (see the remarks after Definitions 2.1) to these particular sets of control laws. The significance of covariant control laws will be illustrated with two examples from control theory.

For a practical application, we will consider a generalization of the familiar pole assignment problem [16].

DEFINITION 5.1. Two control laws, K_1 and K_2, for (A, B) are *indifferent* if $A+BK_1$ and $A+BK_2$ are similar matrices.

This is a definition of convenience because we see that K_1 and K_2 are indifferent if and only if K_2-K_1 is strongly neutral relative to $(A+BK_1, B)$. However, we can now give a corollary to Theorem 3.2.

COROLLARY 5.1. *Let (A, B) be a reachable system and let K be an arbitrary control law. Then there is a covariant control law K_c (relative to (A, B)) such that K and K_c are indifferent.*

Proof. The covariant law K_c constructed in the proof of Theorem 3.2 suffices. □

Thus, in the case that any member of a particular set of indifferent control laws is a satisfactory solution to a particular control problem, there will always be at least one covariant solution. A common example is the use of a control law to determine the invariant factors of the closed loop system matrix (usually the closed loop system matrix is chosen to be cyclic); recall Theorem 2.1. This application indicates that covariant control laws form a large enough class to be useful in meaningful control problems.

We now turn to a theoretical application. Here we will need to demonstrate existence of a suitable covariant control law because we want to preserve s-canonical form while satisfying a certain objective.

Let (A, B, C) be a linear time-invariant system representation defined by the equations

$$x_{t+1} = Ax_t + Bu_t, \tag{5.1}$$

$$y_t = Cx_t; \tag{5.2}$$

$y \in \mathbf{k}^p$ is an output vector and C is a $p \times n$ $\mathbf{k}$-matrix, with (5.1) identical to (1.1). Thus we have added an output equation to our previous system representation.

We say (A, B, C) is reachable if and only if (A, B) in (5.1) is reachable as defined in § 1, and this will continue to be a standing assumption.

DEFINITION 5.2. Two systems (A_1, B_1, C_1) and (A_2, B_2, C_2) are *control equivalent*, or *c-equivalent*, if there is a nonsingular matrix T and a control law K such that

$$T^{-1}(A_1 + B_1K)T = A_2, \tag{5.3}$$

$$T^{-1}B_1 = B_2, \tag{5.4}$$

$$C_1T = C_2. \tag{5.5}$$

A slight modification of the proof of Theorem 3.2 shows that $(A_1, B_1, 0)$, here 0 is the $p \times m$ matrix of zeros, and $(A_2, B_1, 0)$ are c-equivalent if and only if $(A_1 + B_1K, B_1)$ is similar to (A_2, B_2) for some covariant control law K relative to (A_1, B_1). Popov [9] noted that a set of canonical forms for reachable systems (A, B) (here imbedded as systems $(A, B, 0)$) under c-equivalence can be obtained by choosing the s-canonical forms, (3.4)–(3.6), and setting every element denoted by "x" in (3.6) equal to zero.

Clearly every c-equivalence class of systems (A, B, C) contains some systems with (A, B) in s-canonical form. It is thus natural to seek a set composed of one representative system from each c-equivalence class; that is, a set of *c-canonical forms*, requiring in addition that each c-canonical form, say (A^*, B^*, C^*), has (A^*, B^*) in s-canonical form. One construction for a set Σ of c-canonical forms has been given by Wang and Davison [12]. If $(A^*, B^*, C^*) \in \Sigma$, then (A^*, B^*) is in s-canonical form with every "x" in (3.6) set equal to zero. We will construct a second set, Σ^+, of c-canonical forms that exhibit additional structural properties; the basic step in this construction is the application of an appropriate covariant control law to each element of Σ.

A system (A, B, C) is called *observable* if the matrix

$$\mathcal{O}(A, C) = [C', A'C', \cdots, (A')^{n-1}C']' \tag{5.6}$$

(prime denotes transpose) has rank n. It is well known that the ranks of $\mathcal{O}(A + BK, C)$ and $\mathcal{O}(A, C)$ may differ; in other words, linear state variable control laws can affect the observability of a system. We say that a system (A, B, C) is *maximally unobservable*, see [8], if

$$\operatorname{rank} \mathcal{O}(A, C) \leqq \operatorname{rank} \mathcal{O}(A + BK, C) \tag{5.7}$$

for all control laws K. As pointed out by Silverman and Payne [11], if (A, B, C) is maximally unobservable, then the nullspace of $\mathcal{O}(A, C)$ is the largest (A, B)-invariant subspace contained in the nullspace of C. (Recall that a subspace is (A, B)-invariant if it is $(A + BK)$-invariant for some control law K; see Morse and Wonham [7] and references therein.)

There are many applications of these concepts to problems of system inversion, disturbance isolation, and decoupling; Silverman and Payne [11] and Morse and Wonham [7] are representative references. We will show that every c-equivalence class contains a maximally unobservable system (A, B, C) with (A, B) in s-canonical form.

THEOREM 5.1. *Let* (A, B, C) *be a system with* (A, B) *in s-canonical form. Then there is a c-equivalent system* $(\hat{A}, \hat{B}, \hat{C})$ *that is maximally unobservable and satisfying*

$$\hat{A} = A + BK_c, \qquad \hat{B} = B, \tag{5.8}$$

where K_c *is a covariant control law relative to* (A, B).

Proof. We choose a K, using the "structure algorithm" of Silverman and Payne [11], so that $(A + BK, B, C)$ is maximally unobservable. K may be made unique by choosing a particular version of the structure algorithm [11]. If K is covariant relative to (A, B), set $(\hat{A}, \hat{B}, \hat{C}) = (A + BK, B, C)$. If K is not covariant relative to (A, B), let $(\hat{A}, \hat{B})$ be the s-canonical form of $(A + BK, B)$, and define the matrix T as

$$T^{-1}(A + BK)T = \hat{A},$$
$$T^{-1}B = \hat{B}.$$

Recall that T is unique by reachability. Note also that $\hat{B} = B$ because (A, B) is in s-canonical form. Thus $A + BK_c = \hat{A}$, where K_c is a covariant control law relative to (A, B). Let $\hat{C} = CT$. Then $(\hat{A}, \hat{B}, \hat{C})$ is c-equivalent to (A, B, C) and $\mathcal{O}(\hat{A}, \hat{B}, \hat{C}) = \mathcal{O}(A + BK, B, C)T$ so that $(\hat{A}, \hat{B}, \hat{C})$ is maximally unobservable. □

We can use this theorem to construct the set Σ^+ of maximally unobservable c-canonical forms by starting with the set Σ of c-canonical forms described by Wang and Davison [12]. A third set, say Σ^0, can be obtained by using a covariant control law on each element of Σ^+ to zero the "x" elements in (3.6). It would be interesting to know if the covariant control laws used in the construction of Σ^+ from Σ could be chosen so that $\Sigma = \Sigma^0$. This question is still under investigation.

6. Concluding remarks. Recently Wolovich [15] has described a simplified construction of a complete invariant for c-equivalence (and implicitly a set of c-canonical forms) by using frequency domain methods. Dickinson [3] has given a similar construction including, in addition, a frequency domain approach to maximally unobservable c-canonical forms. This is based on an extension of other work of Wolovich [14] relating pole-zero cancellation in transfer functions to loss of observability when a control law is used. These connections are beyond the scope of this paper, however.

Morse [6] also investigated structural invariants of c-equivalence, but failed to obtain a complete invariant. His approach was a geometric one; it appears that the maximally unobservable c-canonical forms give a solution in the geometric spirit of his work. Further investigation of the frequency domain approach in [3] offers promise of more explicit descriptions of the output matrix and of a natural choice of the control law K of Theorem 5.1.

We must also point out that solutions to many important control problems are not always covariant control laws. In particular, choice of K using quadratic regulator theory may not give a covariant control law. This brings up an interesting point: quadratic regulator theory is often used to determine closed loop system matrix pole locations, yet pole placement problems always have a covariant solution! The reconciliation of these two design procedures is still an interesting problem.

Appendix. An arbitrary control law K for (A, B) can be written as $K = J + L$, where J is purely feedforward relative to (A, B) (resp. $(A + BL, B)$) and L is purely feedback relative to $(A + BJ, B)$ (resp. (A, B)), whenever (A, B) is defined over an *algebraically closed* field. This follows from the fact that any polynomial over such a field factors into a product of first degree polynomials over the field. Closure is also a necessary condition for this decomposition as demonstrated by the following example.

Let $\mathbf{k}$ be a field that is not algebraically closed and let $p(z) \in \mathbf{k}[z]$ be a monic irreducible polynomial of degree $d \geqq 2$. Let (A, B) be a reachable system with $n = 11d + 2$ and $m = 5$ and with controllability indices $\nu_1 = 3d + 1$, $\nu_2 = \nu_3 = 3d$, $\nu_4 = d + 1$, $\nu_5 = d$. Let the invariant factors of the matrix A be $\phi_1(z) = \phi_2(z) = zp(z)^3$, $\phi_3(z) = p(z)^3$, $\phi_4(z) = \phi_5(z) = p(z)$. (Theorem 1 can be used to show that such an A exists.)

By Theorem 1, there is a control law K so that the invariant factors of $A + BK$ are $\psi_1(z) = z^{2d+1}p(z)^2$, $\psi_2(z) = z^{2d-1}p(z)$. $\psi_3(z) = z^{d+1}p(z)$, $\psi_4(z) = z^d p(z)$, $\psi_5(z) = z$. This control law cannot be written in the form $K = J + L$, where J and L are purely feedforward and purely feedback laws, respectively. In the practical situation of $\mathbf{k} = R$, the real number field, the choice of $p(z) = z^2 + 1$ will suffice.

Acknowledgment. The author is grateful to Professor T. Kailath of Stanford University for encouraging this research.

REFERENCES

[1] P. BRUNOVSKY, *A classification of linear controllable systems*, Kybernetika (Prague), 3 (1970), pp. 173–187.

[2] B. W. DICKINSON, *On the fundamental theorem of linear state variable feedback*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 577–579.

[3] ———, *Properties and applications of matrix fraction description of linear systems*, Ph.D. thesis, Stanford University, Palo Alto, Calif., 1974.

[4] R. E. KALMAN, *Kronecker invariants and feedback*, Ordinary Differential Equations, L. Weiss, ed., Academic Press, New York, 1972.

[5] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[6] A. S. MORSE, *Structural invariants of linear multivariable systems*, this Journal, 11 (1973), pp. 446–465.

[7] A. S. MORSE AND W. M. WONHAM, *Status of non-interacting control*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 568–581.

[8] H. J. PAYNE AND L. M. SILVERMAN, *On the discrete-time algebraic Riccati equation*, Ibid., AC-18 (1973), pp. 226–234.

[9] V. M. POPOV, *Invariant description of linear, time-invariant controllable systems*, this Journal, 10 (1972), pp. 252–264.

[10] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, John Wiley, New York, 1970.

[11] L. M. SILVERMAN AND H. J. PAYNE, *Input-output structure of linear systems with application to the decoupling problem*, this Journal, 9 (1971), pp. 199–233.

[12] S. H. WANG AND E. J. DAVISON, *Canonical forms of linear multivariable systems*, Control System Rep. 7203, Dept. of Elec. Engrg., University of Toronto, Canada, 1972.

[13] H. WEINERT AND J. ANTON, *Canonical forms for multivariable system identification*, IEEE Decision and Control Conf., New Orleans La., 1972.

[14] W. A. WOLOVICH, *On the cancellation of multivariable system zeros by state feedback*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 276–277.

[15] ———, *Equivalence and invariants in linear multivariable systems*, Proc. 1974 JACC, Austin, Texas, pp. 177–185.

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

[17] W. M. WONHAM AND A. S. MORSE, *Feedback invariants of linear multivariable systems*, Automatica, 8 (1972), pp. 93–100.

[18] H. P. ZEIGER, *Ho's algorithm, commutative diagrams, and the uniqueness of minimal linear systems*, Information and Control, 11 (1967), pp. 71–79.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

LINEAR HILBERT NETWORKS CONTAINING FINITELY MANY NONLINEAR ELEMENTS*

VACLAV DOLEZAL†

Abstract. In this paper we establish conditions for the existence and uniqueness of a regime in a linear (finite or infinite) Hilbert network which contains several nonlinear elements. These conditions are given in terms of the driving point set impedance of the linear network, and the operator describing the nonlinear elements. They are easy to test in specific cases. Two examples illustrating the application are given.

1. Introduction. The objective of the paper is to give relatively simple conditions which guarantee the existence and uniqueness of a regime in a linear Hilbert network containing finitely many nonlinear (possible multivalued) elements. To this end, we first prove a theorem giving conditions under which the driving point set impedance of a (nonlinear, in general) Hilbert network is an operator. Then we establish the main theorem on existence and uniqueness of a regime in a network under consideration. It turns out that the necessary and sufficient conditions in question are given in terms of the mapping $R+Z^+$, where R is the driving point set impedance of the linear part of the network, and Z^+ is the operator describing the nonlinear elements.

As examples we consider a finite R, L, C network with constant elements, which contains either a nonlinear resistor or a nonlinear inductance, and a DC-current network containing several nonlinear resistors.

2. Results. In the sequel, we will use several results obtained in [2] and [3], which are slightly modified for the sake of our present purposes. To facilitate reading the paper, let us first list various concepts and theorems we shall need.

Let X, Y be nonempty sets and let $\sigma(Y)$ be the collection of all nonempty subsets of Y; a mapping $A : X \to \sigma(Y)$ will be called a set mapping from X to Y.

If $\mathcal{D} \subset X$, $\mathcal{D} \neq \emptyset$, we denote $(A\mathcal{D})^0 = \bigcup_{x \in \mathcal{D}} Ax$. Moreover, if A is a set mapping such that Ax is a singleton for each $x \in X$, then A will be called an operator.

Let $A : X \to \sigma(Y)$ be a set mapping, and let $\mathcal{D} \subset X$, $\mathcal{D} \neq \emptyset$; then the set mapping $A^- : (A\mathcal{D})^0 \to \sigma(\mathcal{D})$ defined by

$$A^- y = \{x : x \in \mathcal{D},\ y \in Ax\}$$

will be called the quasi-inverse of A on $\mathcal{D}$.

It is clear that if both A and A^- are operators, then A^- coincides with the ordinary inverse A^{-1}.

Note that in [3] the quasi-inverse A^- was defined only for a simple set mapping A, and then A^- was an operator. The present approach constitutes the essence of the modification mentioned above.

* Received by the editors October 8, 1974, and in revised form April 17, 1975.

† Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York 11794. This research was supported by the National Science Foundation under Grant PO 33568-X00.

It is easy to see that the quasi-inverse has the following properties:

(i) If $x \in \mathcal{D}$, then $z \in Ax \Leftrightarrow x \in A^- z$.

(ii) $(A^-(A\mathcal{D})^0)^0 = \mathcal{D}$.

(iii) For the set mapping $(A^-)^- : \mathcal{D} \to \sigma(Y)$ we have $(A^-)^- = A$.

Next, let $A : X \to \sigma(Y)$ be a set mapping, and let $\mathcal{D} \subset X$, $\mathcal{D} \neq \emptyset$. A will be called simple on $\mathcal{D}$ if $x_1, x_2 \in \mathcal{D}$, $x_1 \neq x_2 \Rightarrow (Ax_1) \cap (Ax_2) = \emptyset$.

Then we have the proposition: A is simple on $\mathcal{D} \Leftrightarrow$ the quasi-inverse $A^- : (A\mathcal{D})^0 \to \sigma(\mathcal{D})$ is an operator.

Finally, given $A : X \to \sigma(Y)$ and an operator $B : Y \to Z$, we define the set mapping $BA : X \to \sigma(Z)$ by $(BA)x = B(Ax) \subset Z$ for each $x \in X$; the definition of AC is analogous. Also, if $A_1, A_2 : X \to \sigma(Y)$ are set mappings and Y is a linear space, we define $A_1 + A_2$ in the obvious way.

Let G be a locally finite oriented graph [1] which has the set of branches $\{b_1, b_2, b_3, \cdots\}$ with cardinal $c_2 \leqq \aleph_0$, the set of vertices $\{v_1, v_2, v_3, \cdots\}$ with cardinal $c_1 \leqq \aleph_0$, and let d be the incidence matrix of G (having type $c_2 \times c_1$). Let $a = [a_{ik}] = K \cdot d^T$, where $K = \operatorname{diag}(k_1, k_2, k_3, \cdots)$ of type $c_1 \times c_1$ is chosen so that the number $k_j \neq 0$ for all j's and $\sum_{i,k} |a_{ik}|^2 < \infty$.

Furthermore, let H be a fixed separable Hilbert space, and let $\hat{a} : H^{c_2} \to H^{c_1}$ be defined by $\hat{a}x = a \cdot x$. (For the definition of H^c, $c \leqq \aleph_0$, see [1]). Then $\hat{a}$ is a linear bounded operator on H^{c_2}, and its nullspace $N_{\hat{a}}$ is closed in H^{c_2} and does not depend on the choice of the matrix K.

Next, let X be a $c_2 \times c_0$ matrix whose columns constitute an orthonormal basis in the solution space of the equation $a \cdot \xi = 0$, $\xi \in R^{c_2}$ (here, R^{c_2} is the Euclidean space for $c_2 < \aleph_0$, and l_2 for $c_2 = \aleph_0$), and let $\hat{X} : H^{c_0} \to H^{c_2}$ be defined by $\hat{X}z = X \cdot z$.

Note that if G is finite, i.e., $c_1, c_2 < \aleph_0$, then X can be easily constructed from a complete set of linearly independent loops in G.

As shown in [1], $\hat{X}$ has the following properties:

(a) $\hat{X}$ is a norm-preserving isomorphism between H^{c_0} and $N_{\hat{a}} \subset H^{c_2}$.

(b) $\hat{a}\hat{X} = 0$ on H^{c_0}.

(c) If $\hat{X}^*$ is the adjoint of $\hat{X}$, then $\hat{X}^*\hat{X} = I$ on H^{c_0}, $\hat{X}^*$ maps H^{c_2} onto H^{c_0}, and $\hat{X}^* v = \bar{X}^T \cdot v$ for all $v \in H^{c_2}$.

(d) $N_{\hat{X}^*} = N_{\hat{a}}^{\perp}$, where $N_{\hat{X}^*} = \{x : x \in H^{c_2}, \hat{X}^* x = 0\}$.

(e) If P is the orthogonal projection from H^{c_2} onto $N_{\hat{a}}$, then $P = \hat{X}\hat{X}^*$.

Now, let $\mathcal{D} \subset H^{c_2}$, $\mathcal{D} \neq \emptyset$, and let $Z : \mathcal{D} \to \sigma(H^{c_2})$ be a set mapping; then the ordered pair $\mathfrak{N} = (\hat{Z}, G)$ will be called a Hilbert network.

Clearly, G in the pair $(\hat{Z}, G)$ describes the structure of a network, and $\hat{Z}$ the behavior of its elements.

DEFINITION. Given $\mathfrak{N} = (\hat{Z}, G)$ and $e \in H^{c_2}$, an element $i \in H^{c_2}$ will be called a *solution of* $\mathfrak{N}$ *corresponding to* e if

K_1: there exists $v \in \hat{Z}i$ such that $v - e \in N_{\hat{a}}^{\perp}$,

K_2: $i \in N_{\hat{a}} \cap \mathcal{D}$.

Obviously, if e is interpreted as a vector of EMF's in branches of our network, then i and v has the meaning of a vector of currents and voltage drops in individual branches, respectively. The justification of this definition of a solution is discussed in detail in [1] and [3].

If $N_{\hat{a}} \cap \mathscr{D} \neq \emptyset$, let $\mathfrak{K} \subset H^{c_0}$ be defined by $\hat{X}\mathfrak{K} = N_{\hat{a}} \cap \mathscr{D}$, and let $Q(\mathscr{D}) = (\hat{X}H^{c_0})^{\perp} + (\hat{Z}\hat{X}\mathfrak{K})^0$; if $N_{\hat{a}} \cap \mathscr{D} = \emptyset$, we put $\mathfrak{K} = \emptyset$, $Q(\mathscr{D}) = \emptyset$. Note that $\mathfrak{K}$ is defined uniquely, since $\hat{X}$ is a bijection from H^{c_0} onto $N_{\hat{a}}$.

Finally, let the set mapping $W : \mathfrak{K} \to \sigma((W\mathfrak{K})^0)$ be defined by $W = \hat{X}^*\hat{Z}\hat{X}$.

THEOREM A. *Let $\mathfrak{N} = (\hat{Z}, G)$ be a Hilbert network, and let $e \in H^{c_2}$; then $\mathfrak{N}$ possesses a solution corresponding to e iff $e \in Q(\mathscr{D})$. In this case, the set $\mathscr{T}$ of all solutions corresponding to e is given by*

$$\mathscr{T} = \hat{X}W^{-}\hat{X}^*e, \tag{1}$$

where W^{-} denotes the quasi-inverse of W on $\mathfrak{K}$.

Obviously, $Q(\mathscr{D})$ has the physical meaning of a set of all EMF's vectors such that if $e \in Q(\mathscr{D})$, there exists at least one current distribution i in the network corresponding to the excitation given by e.

The set mapping $A = \hat{X}W^{-}\hat{X}^* : Q(\mathscr{D}) \to \sigma(N_{\hat{a}} \cap \mathscr{D})$, appearing in the formula (1), will be called the admittance of $\mathfrak{N}$. Observe that in our interpretation, any network $\mathfrak{N}$ with $\mathfrak{K} \neq \emptyset$ possesses the admittance. This fact is in contrast to the "classical" admittance concept, where the "admittance" is usually thought of as an operator. Anyway, these circumstances are clarified by the following definition and theorem.

DEFINITION. A Hilbert network $\mathfrak{N}$ is called *regular on $\mathscr{D}$* if for each $e \in Q(\mathscr{D})$, $\mathfrak{N}$ possesses a unique solution corresponding to e.

Obviously, in this case, $\mathscr{T}$ given by (1) must be a singleton for every $e \in Q(\mathscr{D})$, i.e., the admittance A must be an operator. Actually, we have the following.

THEOREM B. *Let $\mathfrak{N} = (\hat{Z}, G)$ be a Hilbert network with $\mathfrak{K} \neq \emptyset$; then $\mathfrak{N}$ is regular on $\mathscr{D}$ iff the set mapping $W = \hat{X}^*\hat{Z}\hat{X}$ is simple on $\mathfrak{K}$.*

Finally, let us mention the following fact.

THEOREM C. *Let $\mathfrak{N}$ be a Hilbert network with $\mathfrak{K} \neq \emptyset$; then $Q(\mathscr{D}) = H^{c_2}$ iff $(W\mathfrak{K})^0 = H^{c_0}$.*

The meaning of this theorem is straightforward in view of the above comment on $Q(\mathscr{D})$.

The proofs of Theorems A and B are minor modifications of proofs of Theorems 1.1, 1.2 and 2.1 in [3]; the only difference is the fact that the quasi-inverse has a broader meaning in the present context. Consequently, we omit the details. Theorem C is an elementary consequence of relations (2.41) in [3].

We have completed the survey of earlier results. Let us now turn to the anticipated topic of the paper.

We will need the following proposition.

LEMMA 1. *Let $\mathfrak{N} = (\hat{Z}, G)$ be a Hilbert network, and let $\hat{Z}$ be an operator on $\mathscr{D} \subset H^{c_2}$. Assume that*

(i) *there exist $c > 0$ and $p > 1$ such that*

$$\operatorname{Re}\langle Wx_1 - Wx_2, x_1 - x_2\rangle_{c_0} \geqq c\|x_1 - x_2\|_{c_0}^{p} \quad \text{for all } x_1, x_2 \in \mathfrak{K}, \tag{2}$$

(ii) *there exist $d > 0$ and $q > 0$ such that*

$$\|Wx_1 - Wx_2\|_{c_0} \leqq d\|x_1 - x_2\|_{c_0}^{q} \quad \text{for all } x_1, x_2 \in \mathfrak{K}. \tag{3}$$

Then $\mathfrak{N}$ is regular on $\mathfrak{D}$ and the admittance A of $\mathfrak{N}$ (operator) satisfies the inequality

(4) $\operatorname{Re}\langle Ae_1 - Ae_2, e_1 - e_2\rangle_{c_0} \geqq cd^{-p/q}\|\hat{X}^*(e_1 - e_2)\|_{c_0}^{p/q}$ *for all* $e_1, e_2 \in Q(\mathfrak{D})$.

Proof. The regularity of $\mathfrak{N}$ on $\mathfrak{D}$ is guaranteed by Theorem 2.2 (c) in [3]. The inequality (4) follows easily from (2) and (3) by applying the Schwarz inequality; since the pattern of the proof is the same as that of Theorem 2 in [2], we omit the details.

For our purposes it will be convenient to introduce the following notation. Let $c \leqq \aleph_0$ and let n be an integer with $1 \leqq n \leqq c$. If $x = [x_k] \in H^c$, we put $(x)_n = [x_1, x_2, \cdots, x_n]^T \in H^n$; if $M \subset H^c$, we put $(N)_n = \{(x)_n : x \in M\}$. Analogously, for $y = [y_1, y_2, \cdots, y_n]^T \in H^n$, we let $(y)' = [y_1, y_2, \cdots, y_n, 0, 0, \cdots]^T \in H^c$; for $N \subset H^n$, we let $(N)' = \{y' : y \in N\}$.

Let $\mathfrak{N}$ be a Hilbert network with $\mathfrak{F} \neq \emptyset$, let A be the admittance of $\mathfrak{N}$ and let $n \leqq c_2$ be a positive integer; define the sets

(5) $$Q_n = \{e : e \in H^n, e' \in Q(\mathfrak{D})\},$$

(6) $$\mathfrak{D}_n = ((A(Q_n)')^0)_n.$$

Obviously, $\mathfrak{D}_n$ is the set of all n-vectors of currents existing in branches $b_1, b_2, \cdots, b_n$ provided the EMF excitations are present only in $b_1, b_2, \cdots, b_n$.

It is clear that if $Q_n \neq \emptyset$, then $\mathfrak{D}_n \neq \emptyset$; also note that $Q_n \subset H^n$ and $\mathfrak{D}_n \subset H^n$.

DEFINITION. Let $\mathfrak{N}$ be a Hilbert network with $\mathfrak{F} \neq \emptyset$, $\mathfrak{D}_n \neq \emptyset$, and let $A : Q(\mathfrak{D}) \to \sigma(N_{\hat{a}} \cap \mathfrak{D})$ be the admittance of $\mathfrak{N}$. Then the quasi-inverse R of the set mapping $A_n : Q_n \to \sigma(\mathfrak{D}_n)$, defined by

(7) $$A_n e = (A(e)')_n,$$

will be called the *driving point set impedance* of branches $b_1, b_2, \cdots, b_n$ (further DPSI).

A comment on this definition is in order. First note that, in the present context, the DPSI of $b_1, b_2, \cdots, b_n$ exists for any Hilbert network satisfying the requirements $\mathfrak{F} \neq \emptyset$ and $\mathfrak{D}_n \neq \emptyset$. Also observe that we do not lose any generality by focusing our attention to branches $b_1, b_2, \cdots, b_n$ only, since the enumeration of branches in G is immaterial.

The physical meaning of the DPSI is clarified by the following fact.

PROPOSITION. *Let $j \in \mathfrak{D}_n$; then for any $e \in Rj$, there exists a solution i of $\mathfrak{N}$ corresponding to $(e)'$ such that $(i)_n = j$.*

Proof. Choose some $e \in Rj$. Since $R : D_n \to \sigma(Q_n)$, we have $e \in Q_n$, and consequently, by the above propositions (i) and (iii), $j \in R^- e = (A_n^-)^- e = A_n e = (A(e)')_n$. This means that there exists $i \in A(e)'$ such that $j = (i)_n$. However, since $A(e)'$ is the set of all solutions corresponding to $(e)'$ by Theorem A, our proof is complete.

The most important situation occurs when both the admittance and DPSI are operators. In this case, the above proposition reads: "If $j \in \mathfrak{D}_n$, then the unique solution i of $\mathfrak{N}$ corresponding to $(Rj)'$ has the property that $(i)_n = j$." In order to investigate this case more closely, let us introduce several new concepts and carry out some auxiliary considerations.

Let G be a locally finite oriented graph; a nonzero vector $\xi = [\xi_k] \in R^{c_2}$ will be called a loop if $a \cdot \xi = 0$ and each element ξ_k attains one of the values $-1, 0, 1$.

Note that every loop is a simple vector, i.e., all but finitely many of its entries are zero.

We will say that a loop $\xi=[\xi_k]$ contains (does not contain) a branch b_j if $\xi_j \neq 0 (\xi_j = 0)$.

Remark. In the above definition we deviate from the standard terminology of the graph theory. This is done only for the sake of simplicity of the presentation and can hardly lead to a misunderstanding.

DEFINITION. Let G be a locally finite oriented graph, and let n be an integer with $1 \leqq n < c_0$. The set of branches $\{b_1, b_2, \cdots, b_n\}$ will be called *regular* if there exist loops $\xi^1, \xi^2, \cdots, \xi^n$ such that, for each $k = 1, 2, \cdots, n$, the loop ξ^k contains b_k and does not contain any other branch in the set $\{b_1, b_2, \cdots, b_n\}$.

From this definition it follows immediately that vectors $\xi^1, \xi^2, \cdots, \xi^n$ are linearly independent.

LEMMA 2. *Let the set of branches $\{b_1, b_2, \cdots, b_n\}$ be regular, let X be any $c_2 \times c_0$ matrix whose columns constitute an orthonormal basis in the solution space of the equation $a \cdot \xi = 0$, $\xi \in R^{c_2}$, and let H be a separable Hilbert space. Then there exists $\lambda > 0$ such that for the operator $\hat{X} : H^{c_0} \to H^{c_2}$ defined by $\hat{X}z = X \cdot z$ we have*

$$\|\hat{X}^*(z)'\|_{c_0} \geqq \lambda \|z\|_n \tag{8}$$

for every $z \in H^n$.

Proof. Let $\{\xi^1, \xi^2, \cdots, \xi^n\}$ be the linearly independent set of loops existing by the above definition of regularity, and let $\{\xi_0^1, \xi_0^2, \cdots, \xi_0^n\}$ be the orthonormal set in R^{c_2} obtained from the former by applying the Gram–Schmidt process. From the properties of ξ^k's, it follows that the $n \times n$-matrix M_{11} formed by the n first rows of $[\xi_0^1 \vdots \xi_0^2 \vdots \cdots \vdots \xi_0^n]$ is an upper triangular matrix such that each element in the main diagonal is nonzero; consequently, M_{11} is nonsingular. Thus, it is clear that there exists $\mu > 0$ such that

$$\|\bar{M}_{11}^T \cdot \nu\|_n \geqq \mu \|\nu\|_n \tag{9}$$

for every $\nu \in H^n$.

Next, choose vectors $\xi_0^k \in R^{c_2}$, $k = n+1, \quad n+2, \cdots,$ so that $\{\xi_0^1, \xi_0^2, \cdots, \xi_0^n, \xi_0^{n+1}, \cdots\}$ constitutes an orthonormal basis in the solution space of $a \cdot \xi = 0$, $\xi \in R^{c_0}$, and define the $c_2 \times c_0$ matrix X_0 by $X_0 = [\xi_0^1 \vdots \xi_0^2 \vdots \cdots]$. Thus, decomposing X_0 into blocks, we can write

$$X_0 = \left[\begin{array}{c|c} M_{11} & M_{12} \\ \hline M_{21} & M_{22} \end{array}\right]. \tag{10}$$

Now, choosing $z \in H^n$ we get by (10), (9) and proposition (c),

$$\begin{aligned} \|\hat{X}_0^*(z)'\|_{c_0}^2 = \|\bar{X}_0^T \cdot (z)'\|_{c_0}^2 &= \|\bar{M}_{11}^T \cdot z\|_n^2 + \|\bar{M}_{12}^T \cdot z\|_{c_0-n}^2 \\ &\geqq \|\bar{M}_{11}^T \cdot z\|_n^2 \geqq \mu^2 \|z\|_n^2. \end{aligned} \tag{11}$$

To conclude the proof, choose a $c_2 \times c_0$ matrix X whose columns constitute an orthonormal basis in the solution space of $a \cdot \xi = 0$, $\xi \in R^{c_2}$, and let $\hat{X} : H^{c_0} \to H^{c_2}$ be the operator generated by X. Define the operator $S : H^{c_0} \to H^{c_0}$ by

$$S = \hat{X}_0^* \hat{X}. \tag{12}$$

By proposition (a), S is a linear bounded operator. Moreover, S maps H^{c_0} onto H^{c_0}. Indeed, by (a), $\hat{X}H^{c_0} = N_{\hat{a}}$, and consequently, $SH^{c_0} = \hat{X}_0^*\hat{X}H^{c_0} = \hat{X}_0^*N_{\hat{a}}$; however, because also $\hat{X}_0H^{c_0} = N_{\hat{a}}$, it follows by (c) that $H^{c_0} = \hat{X}_0^*\hat{X}_0H^{c_0} = \hat{X}_0^*N_{\hat{a}}$. Hence $SH^{c_0} = H^{c_0}$.

Furthermore, it is easy to see that, on H^{c_0},

$$\hat{X}_0S = \hat{X}. \tag{13}$$

Indeed, choose an $x \in H^{c_0}$. Then $\hat{X}_0Sx = \hat{X}_0\hat{X}_0^*\hat{X}x$; however, since $\hat{X}x \in N_{\hat{a}}$ by (a) and $\hat{X}_0\hat{X}_0^*$ is the orthogonal projection onto $N_{\hat{a}}$ by (e), we have $\hat{X}_0\hat{X}_0^*\hat{X}x = \hat{X}x$, which confirms (13).

Finally, it follows that S is 1–1. Indeed, assume that $Sx = 0$ for some $x \in H^{c_0}$; then, by (13), $\hat{X}_0Sx = \hat{X}x = 0$, and consequently, by (a), $x = 0$.

Summarizing these facts, we have that S is a bounded bijection; thus, by the open mapping theorem, S^{-1} is bounded. Consequently, $(S^*)^{-1} = (S^{-1})^*$ is bounded; i.e., there exists a $\mu' > 0$ such that

$$\|S^*x\|_{c_0} \geqq \mu'\|x\|_{c_0} \tag{14}$$

for each $x \in H^{c_0}$.

To finish the proof, let $z \in H^n$; then (13), (14) and (11) yield

$$\|\hat{X}^*(z)'\|_{c_0} = \|S^*\hat{X}_0^*(z)'\|_{c_0} \geqq \mu'\|\hat{X}_0^*(z)'\|_{c_0} \geqq \mu'\mu\|z'\|_n,$$

which is the inequality (8).

Now we are ready to state conditions under which the DPSI of a network is an operator.

THEOREM 1. *Let $\hat{\mathfrak{N}} = (\hat{Z}, G)$ be a Hilbert network, let $\hat{Z}$ be an operator on $\mathscr{D} \subset H^{c_2}$, let $\mathfrak{F} \neq \varnothing$, $Q_n \neq \varnothing$ and let the set of branches $\{b_1, b_2, \cdots, b_n\}$ be regular. Furthermore, assume that*

(i) *there exist $c > 0$ and $p > 1$ such that*

$$\operatorname{Re}\langle Wx_1 - Wx_2, x_1 - x_2\rangle_{c_0} \geqq c\|x_1 - x_2\|_{c_0}^p \tag{15}$$

for all $x_1, x_2 \in \mathfrak{F}$, where $W = \hat{X}^*\hat{Z}\hat{X} : \mathfrak{F} \to W\mathfrak{F}$,

(ii) *there exist $d > 0$ and $0 < q \leqq 1$ such that*

$$\|Wx_1 - Wx_2\|_{c_0} \leqq d\|x_1 - x_2\|_{c_0}^q \quad \text{for all } x_1, x_2 \in \mathfrak{F}. \tag{16}$$

Then the DPSI $R : \mathscr{D}_n \to Q_n$ *of branches $b_1, b_2, \cdots, b_n$ is an operator, and R has the properties*:

$$\|Rj_1 - Rj_2\|_n \leqq \gamma\|j_1 - j_2\|_n^{q/(p-q)}, \qquad \gamma > 0 \quad \text{for all } j_1, j_2 \in \mathscr{D}_n, \tag{17}$$

$$\operatorname{Re}\langle Rj_1 - Rj_2, j_1 - j_2\rangle_n \geqq c\|j_1 - j_2\|_n^p \quad \text{for all } j_1, j_2 \in \mathscr{D}_n. \tag{18}$$

Proof. First, Lemma 1 implies that $\hat{\mathfrak{N}}$ is regular on $\mathscr{D}$ and its admittance (operator) $A : Q(\mathscr{D}) \to \mathscr{D}$ satisfies the inequality

$$\operatorname{Re}\langle Ax_1 - Ax_2, x_1 - x_2\rangle_{c_2} \geqq \alpha\|\hat{X}^*(x_1 - x_2)\|_{c_0}^\lambda \tag{19}$$

for all $x_1, x_2 \in Q(\mathscr{D})$, where $\lambda = p/q$ and $\alpha = cd^{-\lambda} > 0$. Observe that, by our hypothesis, $\lambda > 1$.

Let the operator $A_n : Q_n \to \mathscr{D}_n$ be defined by (7). From (5) and (6) it follows that A_n maps Q_n onto $\mathscr{D}_n$. We are going to show that A_n is 1–1 on Q_n. Indeed, if $e_1, e_2 \in Q_n$, we have by (7) and (19),

$$\begin{aligned} \operatorname{Re}\langle A_n e_1 - A_n e_2, e_1 - e_2\rangle_n &= \operatorname{Re}\langle A(e_1)' - A(e_2)', (e_1 - e_2)'\rangle_{c_2} \\ &\geqq \alpha \|\hat{X}^*(e_1 - e_2)'\|_{c_0}^{\lambda}. \end{aligned} \tag{20}$$

However, invoking Lemma 2, we get from (20),

$$\operatorname{Re}\langle A_n e_1 - A_n e_2, e_1 - e_2\rangle \geqq \alpha\mu^{\lambda}\|e_1 - e_2\|_n^{\lambda} \tag{21}$$

with $\mu > 0$. This inequality shows readily that A_n is 1–1. Consequently, the quasi-inverse of A_n coincides with A_n^{-1}, i.e., the DPSI $R = A_n^{-1} : \mathscr{D}_n \to Q_n$ is an operator.

To prove the inequality (17), we can proceed as follows. The Schwarz inequality and (21) imply that for $e_1, e_2 \in Q_n$,

$$\|A_n e_1 - A_n e_2\|_n \geqq \alpha\mu^{\lambda}\|e_1 - e_2\|_n^{\lambda - 1}. \tag{22}$$

Choose $j_1, j_2 \in \mathscr{D}_n$ and put $e_k = Rj_k = A_n^{-1} j_k$, $k = 1, 2$. Then (22) yields $\|j_1 - j_2\|_n \geqq \alpha\mu^{\lambda}\|Rj_1 - Rj_2\|_n^{\lambda - 1}$, i.e.,

$$\|Rj_1 - Rj_2\|_n \leqq \gamma\|j_1 - j_2\|_n^{1/(\lambda - 1)}, \qquad \gamma > 0.$$

This, however, is (17), since $1/(\lambda - 1) = q/(p - q)$.

Finally to prove (18), choose $j_1, j_2 \in \mathscr{D}_n$, and put $e_k = (Rj_k)' \in Q(\mathscr{D})$, $k = 1, 2$; also, let i_1 and i_2 be the solution of $\hat{\mathfrak{N}}$ corresponding to e_1 and e_2, respectively. Then we have

$$\operatorname{Re}\langle Rj_1 - Rj_2, j_1 - j_2\rangle_n = \operatorname{Re}\langle e_1 - e_2, i_1 - i_2\rangle_{c_2}. \tag{23}$$

On the other hand, by the definition of a solution we have (note that $\hat{Z}$ is an operator)

$$\begin{aligned} \langle e_1 - e_2, i_1 - i_2\rangle_{c_2} &= \langle e_1 - e_2, Ae_1 - Ae_2\rangle_{c_2} \\ &= \langle \hat{Z}i_1 - \hat{Z}i_2, i_1 - i_2\rangle_{c_2}. \end{aligned}$$

Hence by (23),

$$\operatorname{Re}\langle Rj_1 - Rj_2, j_1 - j_2\rangle_n = \operatorname{Re}\langle \hat{Z}i_1 - \hat{Z}i_2, i_1 - i_2\rangle_{c_2}. \tag{24}$$

However, (15) is equivalent to the condition (see [3])

$$\operatorname{Re}\langle \hat{Z}y_1 - \hat{Z}y_2, y_1 - y_2\rangle_{c_2} \geqq c\|y_1 - y_2\|_{c_2}^{p} \tag{25}$$

for all $y_1, y_2 \in N_{\hat{a}} \cap \mathscr{D}$. Since $i_1, i_2 \in N_{\hat{a}} \cap \mathscr{D}$, we get from (24),

$$\operatorname{Re}\langle Rj_1 - Rj_2, j_1 - j_2\rangle_n \geqq c\|i_1 - i_2\|_{c_2}^{p}. \tag{26}$$

Since $(i_1)_n = j_1$, $(i_2)_n = j_2$, it follows that $\|i_1 - i_2\|_{c_2} \geqq \|j_1 - j_2\|_n$; this together with (26) concludes the proof of (18).

From Theorem 1 we get readily the following result.

THEOREM 2. *Let H be a real Hilbert space, let $\hat{\mathfrak{N}} = (\hat{Z}, G)$ be a Hilbert network with $\hat{Z}$ being an operator on H^{c_2}, and let the set of branches $\{b_1, b_2, \cdots, b_n\}$ be regular. Assume that*

(i) *there exist* $c>0$ *and* $p>1$ *such that*

$$\langle Wx_1 - Wx_2, x_1 - x_2\rangle_{c_0} \geqq c\|x_1 - x_2\|_{c_0}^p \tag{27}$$

for all $x_1, x_2 \in H^{c_0}$, *where* $W = \hat{X}^*\hat{Z}\hat{X} : H^{c_0} \to H^{c_0}$,

(ii) *there exist* $d>0$ *and* $0<q\leqq 1$ *such that*

$$\|Wx_1 - Wx_2\|_{c_0} \leqq d\|x_1 - x_2\|_{c_0}^q \tag{28}$$

for all $x_1, x_2 \in H^{c_0}$.

Then the DPSI R *of branches* $b_1, b_2, \cdots, b_n$ *is an operator from* H^n *onto* H^n *and* (17), (18) *holds for all* $j_1, j_2 \in H^n$.

Proof. Since W is continuous by (ii), (i) shows that W is maximal monotone on H^{c_0}; hence, the corollary following Theorem 2.6 in [3] implies that $Q(H^{c_2}) = H^{c_2}$. Consequently, by (5), $Q_n = H^n$. Thus, by (21), the operator $A_n : H^n \to \mathcal{D}_n \subset H^n$ satisfies the condition

$$\langle A_n e_1 - A_n e_2, e_1 - e_2\rangle_n \geqq \beta\|e_1 - e_2\|_n^\lambda \tag{29}$$

for all $e_1, e_2 \in H^n$, where $\beta>0$ and $\lambda>1$.

On the other hand, if $e_1, e_2 \in H^n$, we have by the definition (7) and Theorem 2.2 (c) in [3] (we use the inequality (2.10) in [3]),

$$\begin{aligned}\|A_n e_1 - A_n e_2\|_n &\leqq \|A(e_1)' - A(e_2)'\|_{c_2} \leqq c^{-1/(p-1)}\|\hat{X}^*(e_1 - e_2)'\|_{c_0}^{1/(p-1)} \\ &\leqq c^{-1/(p-1)}\|(e_1 - e_2)'\|_{c_2}^{1/(p-1)} = c^{-1/(p-1)}\|e_1 - e_2\|_n^{1/(p-1)}.\end{aligned} \tag{30}$$

However, (30) shows that A_n is continuous on H^n. Moreover, (29) implies that A_n is coercive and monotone on H^n. Hence, by continuity, A_n is maximal monotone on H^n, and consequently, $A_n H^n = H^n$ (see [4]), i.e., $\mathcal{D}_n = H^n$. The rest of the proof follows from Theorem 1.

Note that Theorem 5 in [2], apart from the assumption that H is real, is a special case of our Theorem 2 with $n=1$.

Let us now prove our main result—a theorem on regularity of a linear Hilbert network that contains finitely many nonlinear elements.

THEOREM 3. *Let* $\mathfrak{N}' = (\hat{Z}', G)$ *be a Hilbert network with* $\hat{Z}'$ *being a linear (not necessarily bounded) operator defined on a linear subspace* $\mathcal{D}' \subset H^{c_2}$. *Assume that*

(i) $\mathfrak{N}'$ *is regular on* $\mathcal{D}'$ *and* $Q(\mathcal{D}') = H^{c_2}$,

(ii) *the* DPSI *of branches* $b_1, b_2, \cdots, b_n$ *is an operator* $R : \mathcal{D}_n \to H^n$, *where* $\mathcal{D}_n \subset H^n$ *is defined by* (6),

(iii) *for any solution* i' *of* $\mathfrak{N}'$ *we have* $(i')_n \in \mathcal{D}_n$.

Furthermore, let $Z^+ : \mathcal{D}^+ \to \sigma(H^n)$ *be a set mapping, where* $\mathcal{D}^+ \subset H^n$, *and let the set mapping* $\hat{Z}''$ *be defined on* $\mathcal{D}'' = \{x : x \in H^{c_2}, (x)_n \in \mathcal{D}^+\}$ *by*

$$\hat{Z}''x = (Z^+(x)_n)'. \tag{31}$$

Let $\mathfrak{N} = (\hat{Z}, G)$ *where* $\hat{Z} = \hat{Z}' + \hat{Z}''$. *Then*

(a) $\mathfrak{N}$ *is regular on* $\mathcal{D}' \cap \mathcal{D}'' \Leftrightarrow R + Z^+$ *is simple on* $\mathcal{D}_n \cap \mathcal{D}^+$,

(b) *for* $\mathfrak{N}$, $Q(\mathcal{D}' \cap \mathcal{D}'') = H^{c_2} \Leftrightarrow [(R + Z^+)(\mathcal{D}_n \cap \mathcal{D}^+)]^0 = H^n$.

A comment on the physical meaning of this theorem is in order. It is clear that, by definition of $\hat{Z}$, the network $\mathfrak{N}$ is obtained from the linear network $\mathfrak{N}'$ by inserting additional nonlinear elements into branches $b_1, b_2, \cdots, b_n$. The

behavior of these elements is described by the set mapping $\hat{Z}^+$. Since we do not make any particular assumption about Z^+, mutual couplings may exist between the additional nonlinear elements (but no couplings are allowed between the nonlinear elements and elements of the original network $\mathfrak{N}'$). In view of propositions (a) and (b), the regularity of the composite network $\mathfrak{N}$ and the existence of a solution of $\mathfrak{N}$ for a given excitation by EMF's is completely determined by the behavior of the set mapping $R+Z^+$, where R is the DPSI of those branches in $\mathfrak{N}'$ which contain the nonlinear elements.

Proof of Theorem 3. First of all, from the assumption $Q(\mathscr{D}')=H^{c_2}$ and (5) follows that $Q_n=H^n$. Moreover, since $\hat{Z}'$ is a linear operator and $\mathfrak{N}'$ is regular on $\mathscr{D}'$, it follows that the admittance $A': H^{c_2}\to\mathscr{D}'$ of $\mathfrak{N}'$ is a linear operator. (Witness Theorems A and B; $W=\hat{X}^*\hat{Z}'\hat{X}$ is a linear operator on H^{c_0} and so is $W^{-1}=W^-$). Thus, by (6), $\mathscr{D}_n$ is a linear subspace of H^n.

Since $n<c_2$, we will introduce the following notation: if $x=[x_k]\in H^{c_2}$, we let $(x)_{-n}=[x_{n+1}, x_{n+2}, \cdots]^T\in H^{c_2-n}$ (here we put $c_2-n=\aleph_0$ if $c_2=\aleph_0$). Then we clearly have $x=[(x)_n^T:(x)_{-n}^T]^T$ for any $x\in H^{c_2}$.

Due to linearity of A' it follows that there exist linear operators $A_{11}: H^n\to H^n$, $A_{12}: H^{c_2-n}\to H^n$, $A_{21}: H^n\to H^{c_2-n}$ and $A_{22}: H^{c_2-n}\to H^{c_2-n}$ such that

$$A'x=\left[\begin{array}{c:c}A_{11}&A_{12}\\\hdashline A_{21}&A_{22}\end{array}\right]\cdot\left[\begin{array}{c}(x)_n\\\hdashline (x)_{-n}\end{array}\right] \tag{32}$$

for every $x\in H^{c_2}$. Moreover, the assumption (iii) implies that we even have $A_{11}: H^n\to\mathscr{D}_n$ and $A_{12}: H^{c_2-n}\to\mathscr{D}_n$.

On the other hand, by (7) and (32), we have for any $z\in Q_n=H^n$,

$$A_nz=(A(z)')_n=A_{11}z. \tag{33}$$

Since, by (7) and (6), $A_nH^n=\mathscr{D}_n$, and by (ii) the DPSI $R:\mathscr{D}_n\to H^n$ is an operator, (33) shows that $A_n=A_{11}$, i.e., the inverse $A_{11}^{-1}:\mathscr{D}_n\to H^n$ exists and $A_{11}^{-1}=R$.

Now, we are going to show that the following equivalence ($\mathscr{E}$) is true:

($\mathscr{E}$) Let $i\in N_{\hat{a}}\cap(\mathscr{D}'\cap\mathscr{D}'')$ and let $e\in H^{c_2}$; then i is a solution of $\mathfrak{N}$ corresponding to $e\Leftrightarrow$ there exists an element $\tilde{\nu}\in Z^+(i)_n$ such that i is a solution of $\mathfrak{N}'$ corresponding to $e-(\tilde{\nu})'$. In this case,

$$(e)_n+RA_{12}(e)_{-n}\in(R+Z^+)(i)_n, \tag{34}$$

$(i)_n\in\mathscr{D}_n\cap\mathscr{D}^+$, and the element $\tilde{\nu}$ is determined uniquely.

Indeed, assume first that $i\in N_{\hat{a}}\cap(\mathscr{D}'\cap\mathscr{D}'')$ is a solution of $\mathfrak{N}$ corresponding to $e\in H^{c_2}$; then, since $i\in\mathscr{D}''$, we have $(i)_n\in\mathscr{D}^+$. Moreover, by condition K_1 in the definition of a solution, there exists $\nu\in\hat{Z}i$ such that

$$\nu-e\in N_{\hat{a}}^{\perp}. \tag{35}$$

However, since $\hat{Z}=\hat{Z}'+\hat{Z}''$, we have by (31), $\hat{Z}i=\hat{Z}'i+\hat{Z}''i=\hat{Z}'i+(Z^+(i)_n)'$; this amounts to saying that there exists $\tilde{\nu}\in Z^+(i)_n$ such that $\nu=\hat{Z}'i+(\tilde{\nu})'$. Consequently, (35) can be written as

$$\hat{Z}'i-(e-(\hat{\nu})')\in N_{\hat{a}}^{\perp}, \tag{36}$$

i.e., because $i\in N_{\hat{a}}\cap\mathscr{D}'$, i is a solution of $\mathfrak{N}'$ corresponding to $e-(\tilde{\nu})'$.

On the other hand, by (32) we have

$$(i)_n = A_{11}((e)_n - \tilde{\nu}) + A_{12}(e)_{-n}. \tag{37}$$

However, since both elements $A_{11}((e)_n - \tilde{\nu})$ and $A_{12}(e)_{-n}$ are in $\mathscr{D}_n$, (37) is equivalent to

$$R(i)_n + \tilde{\nu} = (e)_n + RA_{12}(e)_{-n}. \tag{38}$$

This relation shows that the element $\tilde{\nu}$ is determined uniquely. Moreover, since $\tilde{\nu} \in Z^+(i)_n$, we have $R(i)_n + \tilde{\nu} \in (R + Z^+)(i)_n$; thus, (38) yields immediately the relation (34).

Conversely, assume that $i \in N_{\hat{a}} \cap (\mathscr{D}' \cap \mathscr{D}'')$ and that there exists $\tilde{\nu} \in Z^+(i)_n$ such that i is a solution of $\mathfrak{N}'$ corresponding to $e - (\tilde{\nu})'$. Then (36) holds (it is K_1 for $\mathfrak{N}'$), and letting $\nu = \hat{Z}'i + (\tilde{\nu})'$, we see readily that $\nu \in \hat{Z}i$ and (35) is satisfied. Hence, i is a solution of $\mathfrak{N}$ corresponding to e, and our equivalence ($\mathscr{E}$) is proven.

(a) Assume now that $R + Z^+$ is simple on $\mathscr{D}_n \cap \mathscr{D}^+$. Suppose that, for some $e \in H^{c_2}$, $\mathfrak{N}$ has solutions i, $\tilde{i} \in N_{\hat{a}} \cap (\mathscr{D}' \cap \mathscr{D}'')$ which correspond to e. Denote $f = (e)_n + RA_{12}(e)_{-n}$; then we have by (34), $f \in (R + Z^+)(i)_n$ and $f \in (R + Z^+)(\tilde{i})_n$. Consequently, $((R + Z^+)(i)_n) \cap ((R + Z^+)(\tilde{i})_n) \neq \varnothing$, so that, by simplicity of $R + Z^+$, $(i)_n = (\tilde{i})_n$.

Moreover, for i there exists a unique element $\tilde{\nu} \in Z^+(i)_n$, and for $\tilde{i}$ a unique element $\tilde{\tilde{\nu}} \in Z^+(\tilde{i})_n$ such that $\tilde{\nu}$ and $\tilde{\tilde{\nu}}$ have properties described in ($\mathscr{E}$). Since $(i)_n = (\tilde{i})_n$, (38) shows that necessarily $\tilde{\nu} = \tilde{\tilde{\nu}}$. Now, again by ($\mathscr{E}$), both i and $\tilde{i}$ are solutions of $\mathfrak{N}'$ corresponding to the same element $e - (\tilde{\nu})'$. However, since $\mathfrak{N}'$ is regular on $\mathscr{D}'$ by (i), it is regular on $\mathscr{D}' \cap \mathscr{D}''$, too; hence, $i = \tilde{i}$, i.e., $\mathfrak{N}$ is regular on $\mathscr{D}' \cap \mathscr{D}''$.

Conversely, assume that $\mathfrak{N}$ is regular on $\mathscr{D}' \cap \mathscr{D}''$. Suppose that there exist $j, \tilde{j} \in \mathscr{D}_n \cap \mathscr{D}^+$ such that $((R + Z^+)j) \cap ((R + Z^+)\tilde{j}) \neq \varnothing$. Then there exists $f \in H^n$ such that $f \in (R + Z^+)j$ and $f \in (R + Z^+)\tilde{j}$. (Note that $R + Z^+$ maps $\mathscr{D}_n \cap \mathscr{D}^+$ into $\sigma(H^n)$). Stated differently, there exists $\tilde{\nu} \in Z^+ j$ and $\tilde{\tilde{\nu}} \in Z^+ \tilde{j}$ such that

$$f = Rj + \tilde{\nu}, \quad f = R\tilde{j} + \tilde{\tilde{\nu}}. \tag{39}$$

Let $e = (f - \tilde{\nu})' = (f)' - (\tilde{\nu})'$ and $\tilde{e} = (f - \tilde{\tilde{\nu}})' = (f)' - (\tilde{\tilde{\nu}})'$; since $\tilde{\nu}$, $\tilde{\tilde{\nu}} \in H^n$, we have $e, \tilde{e} \in H^{c_2}$. Now, since $\mathfrak{N}'$ is regular on $\mathscr{D}'$ and $Q(\mathscr{D}') = H^{c_2}$ by our hypothesis (i), then for e and $\tilde{e}$ there exists a unique solution i and $\tilde{i}$ of $\mathfrak{N}'$ corresponding to e and $\tilde{e}$, respectively. On the other hand, by (32) and (39) we have $(i)_n = A_{11}(f - \tilde{\nu}) = A_{11}Rj = j$ (clearly, $(e)_{-n} = ((f - \tilde{\nu})')_{-n} = 0$), and similarly, $(\tilde{i})_n = A_{11}(f - \tilde{\tilde{\nu}}) = A_{11}R\tilde{j} = \tilde{j}$. Also, note that since $i, \tilde{i} \in \mathscr{D}'$, and the elements $(i)_n = j$ and $(\tilde{i})_n = \tilde{j}$ belong to $\mathscr{D}^+$, we have $i, \tilde{i} \in N_{\hat{a}} \cap (\mathscr{D}' \cap \mathscr{D}'')$. Thus, invoking ($\mathscr{E}$), we conclude that i is a solution of $\mathfrak{N}$ corresponding to $(f)'$, and $\tilde{i}$ is a solution of $\mathfrak{N}$ corresponding to $(f)'$. Hence, by our hypothesis, $i = \tilde{i} \Rightarrow j = (i)_n = (\tilde{i})_n = \tilde{j}$. Thus, $R + Z^+$ is simple on $\mathscr{D}_n \cap \mathscr{D}^+$, which finishes the proof of the assertion (a).

(b) Assume first that $[(R + Z^+)(\mathscr{D}_n \cap \mathscr{D}^+)]^0 = H^n$. Choose arbitrarily $e \in H^{c_2}$ and construct the element $(e)_n + RA_{12}(e)_{-n} \in H^n$. (This is possible, since $A_{12}(e)_{-n} \in \mathscr{D}_n$ and $R : \mathscr{D}_n \to H^n$). By our hypothesis, there exists $j \in \mathscr{D}_n \cap \mathscr{D}^+$ such

that $(e)_n + RA_{12}(e)_{-n} \in (R + Z^+)j$. Thus, there exists $\tilde{\nu} \in Z^+ j$ such that

$$(e)_n + RA_{12}(e)_{-n} = Rj + \tilde{\nu}. \tag{40}$$

On the other hand, by the assumption (i) on $\hat{\mathfrak{N}}'$, there exists a unique solution $i \in \mathscr{D}'$ of $\hat{\mathfrak{N}}'$ corresponding to $e - (\tilde{\nu})' \in H^{c_2}$. Note that $(i)_n \in \mathscr{D}_n$. Moreover, by (32) and (40) we have

$$\begin{aligned}(i)_n &= A_{11}(e - (\tilde{\nu})')_n + A_{12}(e - (\tilde{\nu})')_{-n} \\ &= A_{11}(e)_n - A_{11}^n \tilde{\nu} + A_{12}(e)_{-n} = A_{11}(Rj - RA_{12}(e)_{-n}) + A_{12}(e)_{-n} = j.\end{aligned} \tag{41}$$

Since $(i)_n = j \in \mathscr{D}^+$, we have $i \in N_{\hat{a}} \cap (\mathscr{D}' \cap \mathscr{D}'')$, (see the definition of $\mathscr{D}''$). Thus by $(\mathscr{E})$, i is a solution of $\hat{\mathfrak{N}}$ corresponding to e. Hence, $Q(\mathscr{D}' \cap \mathscr{D}'') = H^{c_2}$.

Conversely, assume that $Q(\mathscr{D}' \cap \mathscr{D}'') = H^{c_2}$, and arbitrarily choose $f \in H^n$. Then, by our hypothesis, there exists a solution $i \in N_{\hat{a}} \cap (\mathscr{D}' \cap \mathscr{D}'')$ of $\hat{\mathfrak{N}}$ corresponding to $(f)' \in H^{c_2}$. Note that $(i)_n \in \mathscr{D}^+$. However, by $(\mathscr{E})$, i is also a solution of $\hat{\mathfrak{N}}'$ corresponding to $(f)' - (\tilde{\nu})'$, where $\tilde{\nu} \in Z^+(i)_n$. Consequently, by (iii), $(i)_n \in \mathscr{D}_n$, so that $(i)_n \in \mathscr{D}_n \cap \mathscr{D}^+$. Moreover, by (34),

$$((f)')_n + RA_{12}((f)')_{-n} = f \in (R + Z^+)(i)_n, \quad \text{i.e., } f \in [(R + Z^+)(\mathscr{D}_n \cap \mathscr{D}^+]^0.$$

Hence $[(R + Z^+)(\mathscr{D}_n \cap \mathscr{D}^+)]^0 = H^n$, which completes the proof of proposition (b).

Before proceeding further, let us make a few remarks.

Theorem 3 clearly solves completely the problem of the existence and uniqueness of a current distribution in a composite network, since our propositions (a), (b) give sufficient *and* necessary conditions.

On the other hand, it is worth pointing out the following fact: as we can see, Theorem 3 (and its proof, too) does not make any essential use of the topological properties of spaces involved, i.e., it has a purely set theoretic character. Consequently, Theorem 3 remains true without any change for finite networks whose underlying space is any linear space not necessarily equipped with any topology. As a matter of fact, we can introduce the "abstract network" in a slightly more general way than it is done in [3]. In essence, we replace the Hilbert space $\mathscr{H}$ by a linear space $\mathscr{L}$, and subspaces N_a, $N_a^\perp$ by some subspaces N, M of $\mathscr{L}$ such that $\mathscr{L} = N \oplus M$. Specifying then $\mathscr{L} = L^{c_2}$, where L is a linear space and $c_2 < \aleph_0$, (we can consider only finite networks, since a convergence concept is missing), we can prove the same results as Theorems A–C, define the DPSI as above, and get the same proposition as Theorem 3.

Returning to the Hilbert network, let us point out the fact that the concept of the DPSI can be extended without essential difficulties to the case that countably many branches are involved, and that results like Theorems 1–3 can be proved. Because this is more or less obvious, we omit the details.

Finally, it is easy to see that the generalization of the Shannon–Hagelbarger theorem given in [2] (Theorem 6) remains true, if the "driving point impedance of b_1" is replaced by the DPSI of $b_1, b_2, \cdots, b_n$.

Let us now discuss some simple applications of Theorem 3 in the case $n = 1$.

Example 1. Let G be a finite oriented graph having branches $b_1, b_2, \cdots, b_{c_2}$, and let $\hat{\mathfrak{N}}' = (\hat{Z}', G)$ be a (finite) network built up from constant (not necessarily

nonnegative) R, L, C elements, i.e., the operator $\hat{Z}'$ is described by a $c_2 \times c_2$ matrix $[Z'_{ik}]$, where

$$(Z_{ik}x)(t) = L_{ik}x'(t) + R_{ik}x(t) + S_{ik}\int_0^t x(\sigma)\, d\sigma, \tag{42}$$

and L_{ik}, R_{ik}, S_{ik} are real numbers.

For the underlying space H let us take the real space $L_2[0, \tau]$, $\tau > 0$; also, let $\mathscr{K}$ be the space of all absolutely continuous functions x on $[0, \tau]$ such that $x(0) = 0$ and $x' \in L_2[0, \tau]$. Clearly, $\mathscr{K} \subset L_2[0, \tau]$.

Now, let $\mathscr{D}' = M_1 \times M_2 \times \cdots \times M_{c_2}$, where $M_k = L_2[0, \tau]$, if the branch b_k does not contain any inductance, i.e., if $L_{kj} = 0$ for $j = 1, 2, \cdots, c_2$, and $M_k = \mathscr{K}$ in the opposite case.

It is clear that then $\hat{Z}'$ is well-defined on $\mathscr{D}'$, and maps $\mathscr{D}'$ into $L_2^{c_2}[0, \tau]$. We will assume that

(a) $M_1 = L_2[0, \tau]$,

(b) $\mathfrak{N}'$ is regular on $\mathscr{D}'$ and $Q(\mathscr{D}') = H^{c_2}$.

If the admittance $\hat{A}' : H^{c_2} \to \mathscr{D}'$ of $\mathfrak{N}'$ is described by a matrix $[A_{ik}]$ (of type $c_2 \times c_2$), then, as known from the elementary network analysis, each operator A_{ik} has necessarily the form (witness assumption (2))

$$(A_{ik}x)(t) = a_{ik}x(t) + \int_0^t K_{ik}(t-\sigma)x(\sigma)\, d\sigma \tag{43}$$

where a_{ik} is a real number and $K_{ik}(\sigma)$ is a continuous function on $[0, \infty)$ (in fact, K_{ik} is a linear combination of functions $e^{\lambda t}\{P(t)\cos \omega t + Q(t)\sin \omega t\}$, P, Q being polynomials.

Furthermore, we will assume that

(c) $a_{11} \neq 0$.

Then it is easy to see that A_{11} is 1–1 from $L_2[0, \tau]$ onto itself (it is a Volterra operator), and consequently, $R = A_{11}^{-1} : L_2[0, \tau] \to L_2[0, \tau]$ is the driving point impedance of the branch b_1. Also, R has the form

$$(Rx)(t) = rx(t) + \int_0^t K(t-\sigma)x(\sigma)\, d\sigma, \tag{44}$$

where $r \neq 0$ and $K(\sigma)$ is continuous on $[0, \infty)$. (Note that in the above development we assume that our network is at rest $t = 0$, i.e., currents and charges in capacitors are zero).

Thus, our assumptions (a)–(c) imply that conditions (i)–(iii) in Theorem 3 are satisfied.

Next, let us build a network $\mathfrak{N}$ from $\mathfrak{N}'$ by inserting a nonlinear resistor into the branch b_1. To be more specific, let φ be a real-valued continuous function on R^1, and, using the notation of Theorem 3, let $Z^+ : L_2[0, \tau] \to L_2[0, \tau]$ be defined by

$$(Z^+x)(t) = \varphi(x(t)). \tag{45}$$

(Thus, we have $\mathscr{D}^+ = L_2[0, \tau]$).

Finally, we will assume that the function $\Psi : R^1 \to R^1$, defined by

$$\Psi(\xi) = \varphi(\xi) + r\xi, \tag{46}$$

satisfies the following condition:

(d) there exists an $\alpha > 0$ such that, for all $\xi_1, \xi_2 \in R^1$,

$$|\Psi(\xi_1) - \Psi(\xi_2)| \geqq \alpha |\xi_1 - \xi_2|. \tag{47}$$

We are going to show that, under the assumptions (a)–(d), the network $\hat{\mathfrak{N}} = (\hat{Z}, G)$, with $\hat{Z}$ being defined as in Theorem 3, i.e., $\hat{Z}x = \hat{Z}'x + (Z^+(x)_1)'$, is regular on $\mathscr{D}'$, and $Q(\mathscr{D}') = H^{c_2}$. (Note that we have $\mathscr{D}' \cap \mathscr{D}'' = \mathscr{D}'$).

To this end, observe first that,due to continuity of Ψ and (47), Ψ is 1–1 and maps R^1 onto R^1. Thus if Ψ^{-1} is the inverse of ψ, we get readily from (47),

$$|\Psi^{-1}(\eta_1) - \Psi^{-1}(\eta_2)| \leqq \alpha^{-1} |\eta_1 - \eta_2| \tag{48}$$

for all $\eta_1, \eta_2 \in R^1$.

Referring to Theorem 3, consider the operator $(R + Z^+) : L_2[0, \tau] \to L_2[0, \tau]$. By (44), (45) and (46) we have

$$((R + Z^+)x)(t) = \Psi(x(t)) + \int_0^t K(t - \sigma)x(\sigma)\, d\sigma. \tag{49}$$

Arbitrarily choose $y \in L_2[0, \tau]$ and consider the equation $(R + Z^+)x = y$. By the above and (49), this equation is equivalent to

$$x(t) = \Psi^{-1}\left(-\int_0^t K(t - \sigma)x(\sigma)\, d\sigma + y(t)\right). \tag{50}$$

Define the operator S on $L_2[0, \tau]$ by

$$(Sx)(t) = \Psi^{-1}\left(-\int_0^t K(t - \sigma)x(\sigma)\, d\sigma + y(t)\right). \tag{51}$$

Since ψ^{-1} is continuous by (48), we see easily that S maps $L_2[0, \tau]$ into itself.

Now we are going to show that, for some integer $n \geqq 1$, S^n is a contraction on $L_2[0, \tau]$. Indeed, let $C = \sup_{\xi \in [0,\tau]} |K(\xi)|$. Then we have by (51) for $x_1, x_2 \in L_2[0, \tau]$ and $t \in [0, \tau]$ (see (48)),

$$\begin{aligned} |(Sx_1 - Sx_2)(t)| &\leqq \alpha^{-1} \left| \int_0^t K(t - \sigma)(x_2 - x_1)(\sigma)\, d\sigma) \right| \\ &\leqq \alpha^{-1} C \int_0^t |x_1 - x_2|(\sigma)\, d\sigma. \end{aligned} \tag{52}$$

Using the induction, we can easily confirm that, for any integer $n \geqq 1$,

$$|(S^n x_1 - S^n x_2)(t)| \leqq \frac{(\alpha^{-1} C)^n}{(n-1)!} \int_0^t (t - \sigma)^{n-1} |x_1 - x_2|(\sigma)\, d\sigma. \tag{53}$$

However, (53) yields

$$\begin{aligned} |(S^n x_1 - S^n x_2)(t)| &\leqq \frac{(\alpha^{-1} C)^n}{(n-1)!} t^{n-1} \sqrt{\int_0^t d\sigma} \sqrt{\int_0^t |x_1 - x_2|^2\, d\sigma} \\ &\leqq \frac{(\alpha^{-1} C)^n \tau^{n-(1/2)}}{(n-1)!} \|x_1 - x_2\|. \end{aligned}$$

Hence

$$\|S^n x_1 - S^n x_2\| \leqq \lambda_n \|x_1 - x_2\|, \tag{54}$$

where

$$\lambda_n = (\alpha^{-1} C\tau)^n/(n-1)!. \tag{55}$$

From (55) it follows that $\lambda_n < 1$ for n sufficiently large, i.e., S^n is a contraction for such n. Hence there exists a unique $x \in L_2[0, \tau]$ such that $x = Sx$, i.e., (50) holds; consequently, we have $(R + Z^+)x = y$. Thus, $R + Z^+$ is 1–1 on $L_2[0, \tau]$ and maps it onto itself. This means that, by (a) and (b) in Theorem 3, network $\mathfrak{N}$ is regular on $\mathscr{D}'$ and $Q(\mathscr{D}') = H^{c_2}$, which is what we wanted to show.

Example 2. Let $\mathfrak{N}'$ be exactly the same network as in Example 1, and assume that conditions (a)–(c) are satisfied. Now, we will consider a network $\mathfrak{N}$ obtained from $\mathfrak{N}'$ by inserting a nonlinear inductance into the branch b_1.

In more detail, let Φ be a real-valued function on R^1 which possesses a continuous derivative Φ' everywhere and satisfies the condition $\Phi(0) = 0$. Using the notation of Theorem 3, define the operator $Z^+ : \mathscr{D}^+ = \mathscr{H} \to L_2[0, \tau]$ by

$$(Z^+ x)(t) = [\Phi(x(t))]'. \tag{56}$$

It is clear that this definition is meaningful, and Z^+ truly maps $\mathscr{H}$ into $L_2[0, \tau]$.

Furthermore, we will assume that

(d)* there exists $\alpha > 0$ such that $|\Phi'(\xi)| \geqq \alpha$ for all $\xi \in R^1$.

We are going to show that, under conditions (a)–(c) and (d)*, the network $\mathfrak{N} = (\hat{Z}, G)$ is regular on $\mathscr{D}' \cap \mathscr{D}''$ and $Q(\mathscr{D}' \cap \mathscr{D}'') = H^{c_2}$. Here $\mathscr{D}' \cap \mathscr{D}'' = \mathscr{H} \times M_2 \times \cdots M_{c_2}$, where $M_2, M_3, \cdots, M_{c_2}$ are the same as in Example 1, and $\hat{Z}x = \hat{Z}'x + (Z^+(x)_1)'$ for every $x \in \mathscr{D}' \cap \mathscr{D}''$.

To prove this, note first that (d)* and the mean value theorem imply that

$$|\Phi(\xi_1) - \Phi(\xi_2)| \geqq \alpha |\xi_1 - \xi_2| \tag{57}$$

for all $\xi_1, \xi_2 \in R^1$. Since Φ is continuous, it is a 1–1 mapping from R^1 onto R^1, and for the inverse Φ^{-1} we have

$$|\Phi^{-1}(\eta_1) - \Phi^{-1}(\eta_2)| \leqq \alpha^{-1} |\eta_1 - \eta_2| \tag{58}$$

for all $\eta_1, \eta_2 \in R^1$; also $\Phi^{-1}(0) = 0$.

Consider now the operator $(R + Z^+) : \mathscr{D}_1 \cap \mathscr{D}^+ = \mathscr{H} \to L_2[0, \tau]$. By (44) and (56) we have

$$((R + Z^+)x(t) = [\Phi(x(t))]' + rx(t) + \int_0^t K(t - \sigma)x(\sigma)\, d\sigma. \tag{59}$$

Next, arbitrarily choose $y \in L_2[0, \tau]$, and consider the following two equations:

$$[\Phi(x(t))]' + rx(t) + \int_0^t K(t - \sigma)x(\sigma)\, d\sigma = y(t), \tag{*}$$

$$\Phi(x(t)) + r\int_0^t x(\sigma)\, d\sigma + \int_0^t \left(\int_0^\omega K(\omega - \sigma)x(\sigma)\, d\sigma \right) d\omega = \int_0^t y(\sigma)\, d\sigma. \tag{$*_*$}$$

Now if there exists $x \in \mathcal{K}$ which satisfies (*), then x satisfies ($\overset{*}{*}$), too; for seeing this it suffices to realize that $\Phi(x(0)) = \Phi(0) = 0$.

Conversely, let $x \in L_2[0, \tau]$ be an element satisfying ($\overset{*}{*}$); then $x \in \mathcal{K}$ and satisfies (*).

Indeed, since Φ is 1–1 from R^1 onto R^1, equation ($\overset{*}{*}$) is equivalent to

$$(+) \qquad x(t) = \Phi^{-1}\left[-r\int_0^t x(\sigma)\,d\sigma - \int_0^t \left(\int_0^\omega K(\omega-\sigma)x(\sigma)\,d\sigma\right)d\omega + \int_0^t y(\sigma)\,d\sigma\right].$$

However, all terms in $[\cdots]$ are absolutely continuous; since Φ^{-1} has a continuous derivative $(\Phi^{-1})'$ and, by (d)*, $|(\Phi^{-1})'(\xi)| \leqq \alpha^{-1}$ for all $\xi \in R^1$, it follows that $\Phi^{-1}[\cdots]$ is absolutely continuous, and consequently, so is $x(t)$. Moreover, differentiating (+), we get

$$(\ddagger) \qquad x'(t) = (\Phi^{-1})'[\cdots]\cdot\left\{-rx(t) - \int_0^t K(t-\sigma)x(\sigma)\,d\sigma + y(t)\right\}.$$

Since $(\Phi^{-1})'[\cdots]$ is continuous, ($\ddagger$) shows that $x'(t) \in L_2[0, \tau]$.

Finally, putting $t = 0$ into (+), we get $x(0) = \Phi^{-1}[0] = 0$; hence, $x \in \mathcal{K}$.

To conclude the proof of our assertion, it suffices to realize that (*) follows from ($\overset{*}{*}$) by differentiation.

Next, define the operator $S : L_2[0, \tau] \to L_2[0, \tau]$ by

$$(60) \qquad (Sx)(t) = \Phi^{-1}\left(-\int_0^t h(t-\sigma)x(\sigma)\,d\sigma + \int_0^t y(\sigma)\,d\sigma\right),$$

where

$$h(t-\sigma) = r + \int_0^t K(\xi-\sigma)\,d\xi.$$

Then it is clear that equation $x = Sx$ is equivalent to (+).

On the other hand, since h is continuous and Φ^{-1} satisfies (58), it follows in the same way as in Example 1 that S^n is a contraction for n sufficiently large (after all, our S is practically the same as operator S in Example 1; see (51)). Thus for the chosen $y \in L_2[0, \tau]$, there exists a unique $x \in L_2[0, \tau]$ such that $x = Sx \Rightarrow (+)$ holds $\Rightarrow$ ($\overset{*}{*}$) holds $\Rightarrow x \in \mathcal{K}$ and (*) holds $\Rightarrow (R + Z^+)x = y$.

Hence the operator $(R + Z^+)$ is 1–1 and maps $\mathcal{K}$ onto $L_2[0, \tau]$. Thus by Theorem 3, our network $\hat{\mathfrak{N}}$ is regular on $\mathcal{D}' \cap \mathcal{D}''$ and $Q(\mathcal{D}' \cap \mathcal{D}'') = H^{c_2}$, which is what we wanted to show.

Let us mention the fact that the above examples can be extended to the case that $\hat{\mathfrak{N}}'$ contains time-varying elements; the case of constant elements, however, is easier to analyze because the operators A_{ik} (see (43)) can be easily established by using methods of classical network analysis.

Example 3. Again let G be a finite oriented graph having branches $b_1, b_2, \cdots, b_{c_2}$, and let $\hat{\mathfrak{N}} = (\hat{Z}', G)$ be a DC-network built up from constant resistors, i.e., $\hat{Z}'$ is described by a matrix $\tilde{R} = \operatorname{diag}(r_1, r_2, \cdots, r_{c_2})$ where the r_j's are real (not necessarily positive) numbers. For H we take the real line R^1.

We will assume that:

(a) The set of branches $\{b_1, b_2, \cdots, b_n\}$, $n < c_2$, is regular.

(b) The smallest eigenvalue λ of the $c_0 \times c_0$ matrix $X^T \cdot \tilde{R} \cdot X$ is positive. (Here we take X real; clearly, λ is independent of the choice of X. Also note that $\lambda \leqq \min_{1 \leqq k \leqq c_2} r_k$.)

Now we will construct $\hat{\mathfrak{N}}$ by inserting nonlinear resistors into branches $b_1, b_2, \cdots, b_n$; note that mutual couplings may exist between these nonlinear resistors. In more detail, we let $Z^+ : R^n \to R^n$ be defined by

$$Z^+ x = \varphi(x), \tag{61}$$

where $\varphi : R^n \to R^n$ is a continuous function.

Finally, we will assume that

(c) there exists $\mu < \lambda$ such that

$$(\varphi(\xi_1) - \varphi(\xi_2))^T \cdot (\xi_1 - \xi_2) \geqq -\mu \|\xi_1 - \xi_2\|^2 \quad \text{for all } \xi_1, \xi_2 \in R^n. \tag{62}$$

We are going to show that, under assumptions (a)–(c), $\hat{\mathfrak{N}}$ is regular on R^{c_2} and $Q(R^{c_2}) = R^{c_2}$.

Indeed, since the matrix $X^T \cdot \tilde{R} \cdot X$ is positive definite by (2), the operator $W = \hat{X}^* \hat{Z}' \hat{X}$ is a bijection between R^{c_0} and itself; thus, by Theorems B and C it follows that $\hat{\mathfrak{N}}'$ is regular on R^{c_2} and $Q(R^{c_2}) = R^{c_2}$. Hence the condition (i) in Theorem 3 is met.

Moreover, (b) implies readily that

$$\langle Wx, x \rangle_{c_0} \geqq \lambda \|x\|_{c_0}^2 \quad \text{and} \quad \|Wx\|_{c_0} \leqq d \|x\|_{c_0}$$

for all $x \in R^{c_0}$ and some $d > 0$; hence, by (a) and Theorem 2, the DPSI $R_0 : R^n \to R^n$ of $b_1, b_2, \cdots, b_n$ is a continuous operator, and we have by (18),

$$\langle R_0 j_1 - R_0 j_2, j_1 - j_2 \rangle_n \geqq \lambda \|j_1 - j_2\|_n^2 \tag{63}$$

for all $j_1, j_2 \in R^n$.

On the other hand, (c) yields $\langle Z^+ j_1 - Z^+ j_2, j_1 - j_2 \rangle_n \geqq -\mu \|j_1 - j_2\|_n^2$, and consequently, by (63),

$$\langle (R_0 + Z^+) j_1 - (R_0 + Z^+) j_2, j_1 - j_2 \rangle_n \geqq (\lambda - \mu) \|j_1 - j_2\|_n^2 \tag{64}$$

for all $j_1, j_2 \in R^n$.

However, (64) shows that $R_0 + Z^+$ is 1–1 on R^n, and that $R_0 + Z^+$ is a coercive, maximal monotone operator because it is continuous; hence, by Rockafellar's theorem [4], $(R_0 + Z^+) R^n = R^n$. Theorem 3 completes the proof of our claim.

As for the existence of a solution of $\hat{\mathfrak{N}}$, our conditions (a)–(c) can be modified. Indeed, assume that the following requirements (a)* and (b)* are met:

(a)* The network $\hat{\mathfrak{N}}'$ is regular on R^{c_2} and the DPSI R_0 of branches $b_1, b_2, \cdots, b_n$ is an operator.

(b)* There exist constants $q > 0$ and $M > 0$ such that

$$\|\varphi(\xi)\| \leqq q \|\xi\| \quad \text{for all } \xi \in R^n \text{ with } \|\xi\| > M. \tag{65}$$

We are going to show that, if q is sufficiently small, the network $\hat{\mathfrak{N}}$ has the property $Q(R^{c_2}) = R^{c_2}$.

To this end, observe first that, due to Theorems B, C and finite-dimensionality of R^{c_0}, we have $Q(R^{c_2}) = R^{c_2}$ for $\hat{\mathfrak{N}}'$. Moreover, since R_0 is a

homeomorphism between R^n and itself (witness (7)), it follows that $\mathscr{D}_n = R^n$, and consequently, all conditions (i)–(iii) in Theorem 3 are satisfied. Also, there exists $\mu > 0$ such that

$$\mu\|x\|_n \leqq \|R_0 x\|_n \quad \text{for all } x \in R^n. \tag{66}$$

On the other hand, (61) and (65) yield $\|Z^+ x\|_n \leqq q\|x\|_n$ for all $x \in R^n$ with $\|x\|_n > M$; thus, for any such x, we have by (66),

$$\|Z^+ x\|_n \leqq q\mu^{-1}\|R_0 x\|_n. \tag{67}$$

Now, if q is so small that $q\mu^{-1} < 1$, then the Sandberg–Willson's theorem [5] shows that $(R_0 + Z^+)R^n = R^n$; hence, Theorem 3 concludes the proof of our claim.

REFERENCES

[1] V. DOLEZAL, *Hilbert networks I*, this Journal, 12 (1974), pp. 755–778.

[2] V. DOLEZAL AND A. H. ZEMANIAN, *Hilbert networks. II: Some qualitative properties*, this Journal, 13 (1975), pp. 153–162.

[3] V. DOLEZAL, *Generalized Hilbert networks*, this Journal, 14 (1976), pp. 26–41.

[4] R. T. ROCKAFELLAR, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75–88.

[5] I. W. SANDBERG AND A. N. WILLSON, JR., *Some theorems on properties of* DC *equations of nonlinear networks*, SIAM J. Appl. Math., 22 (1972), pp. 173–186.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

MULTISTAGE STOCHASTIC PROGRAMMING WITH RECOURSE: THE EQUIVALENT DETERMINISTIC PROBLEM*

PAUL OLSEN†

Abstract. Multistage stochastic programming with recourse is defined recursively as a natural extension of two-stage stochastic programming with recourse. Some existing results for two-stage problems are extended to problems with $K+1$ stages, where $1 \leqq K < +\infty$, in the special case of a fixed technology matrix. These results generally involve characterization of the "equivalent deterministic problem"—showing, for example, that it has a convex objective function, a lower-semicontinuous convex objective function, a Lipschitzian objective function, or linear induced constraints.

1. Introduction. The theory of two-stage stochastic programming with recourse has been extensively developed by Roger Wets and D. W. Walkup (see, for example, Walkup and Wets (1967), (1969b), Wets (1966b), (1966c)). They specialized Dantzig's "linear programming under uncertainty" (Dantzig (1955), also see Dantzig (1963, Chap. 25)) to the two-stage case, but also generalized it by allowing a random technology matrix. Wets (1966c), (1972) has explored the generalization to more than two stages under the assumption that the random variables in any stage are independent of the random variables in the preceding stages. This paper extends some of the results for two-stage problems to problems with $K+1$ stages, where $1 \leqq K < +\infty$, in the special case of linear constraints and a fixed technology matrix. These results usually involve characterization of the "equivalent deterministic problem"—showing, for example, that it has a convex objective function, a lower-semicontinuous convex objective function, a Lipschitzian objective function, or linear induced constraints.

2. Statement of the problem. The stochastic element in multistage stochastic programming with recourse enters via dependence of problem data on random vectors—$X_0, \cdots, X_K$. X_k $(0 \leqq k \leqq K)$ represents the state of the world at stage k. A realization of X_k is denoted "x_k"; $x_k \in R^{s_k}$. Let $\underline{X}_k$ be the random vector $(X_0, \cdots, X_k)$; a realization is denoted "$\underline{x}_k$". Let $S_k = \sum_{i=0}^{k} s_i$. For each $1 \leqq k \leqq K$, a regular conditional distribution function for X_k given $\underline{X}_{k-1}$, $F_{X_k|\underline{X}_{k-1}}$, is specified as part of the problem framework. By definition (Ash (1972, p. 263)), for each $y \in R^{s_k}$

$$F_{X_k|\underline{X}_{k-1}}(y|\underline{x}_{k-1}) = P\{X_k \leqq y | \underline{X}_{k-1} = \underline{x}_{k-1}\}$$

for almost every (a.e.) $\underline{x}_{k-1}$. Thus, the regular conditional distribution functions link the random vectors $X_0, \cdots, X_K$.

The stochastic programming problem to be studied is really a recursive sequence of problems, one for each stage. The decision made at stage k $(0 \leqq k \leqq K)$ is given by a vector $u_k \in R^{n_k}$. The sequence of decisions made up to and including stage k is given by the vector $\underline{u}_k = (u_0, \cdots, u_k) \in R^{N_k}$, where $N_k = \sum_{i=0}^{k} n_i$. To begin the recursion, let $\bar{p}_{K+1}(\underline{u}_k\,;\,\underline{x}_k) \equiv 0$. Now let $1 \leqq k \leqq K$, and

* Received by the editors July 2, 1974, and in revised form June 1, 1975.
† Institute for Defense Analyses, Arlington, Virginia 22202.

suppose that $\bar{p}_{k+1} : R^{N_k} \times R^{S_k} \to [-\infty, +\infty]$ has been defined. Define the stage k objective function

$$f_k(\underline{u}_k ; \underline{x}_k) \equiv c_k(\underline{u}_k ; \underline{x}_k) + \bar{p}_{k+1}(\underline{u}_k ; \underline{x}_k)$$

and return function

$$r_k(\underline{u}_k ; \underline{x}_k) \equiv \begin{cases} f_k(\underline{u}_k ; \underline{x}_k), & \sum_{j=0}^{k} A_{kj}\underline{u}_j = b_k(\underline{x}_k), \quad \underline{u}_k \geqq 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

The cost function $c_k : R^{N_k} \times R^{S_k} \to [-\infty, +\infty]$, the $m_k \times n_j$ technology matrix A_{kj} $(0 \leqq j \leqq k)$, and the right-hand side $b_k : R^{S_k} \to R^{m_k}$ are all problem data. For $z \in R^{N_{k-1}}$ define the parameterized problem

$$P_k(z ; \underline{x}_k) : \quad \underset{u_k}{\text{minimize}}\ r_k(z, u_k ; \underline{x}_k)$$

or, equivalently,

$$\begin{aligned} &\underset{\underline{u}_k}{\text{minimize}}\ f_k(\underline{u}_k ; \underline{x}_k) \\ &\text{subject to } A_k\underline{u}_k = b_k(\underline{x}_k), \\ &\qquad \underline{u}_{k-1} = z, \qquad \underline{u}_k \geqq 0, \end{aligned}$$

where $A_k \triangleq [A_{k0} \mid \cdots \mid A_{kk}]$. The stage k perturbation function, p_k, is given by

$$p_k(z ; \underline{x}_k) \equiv \inf (P_k(z ; \underline{x}_k)),$$

the optimal value of the problem $P_k(z ; \underline{x}_k)$ ($+\infty$ if the problem is inconsistent). The stage k expected optimal return function, $\bar{p}_k$, is given by

$$\bar{p}_k(z ; \underline{x}_{k-1}) \equiv \int p_k(z ; \underline{x}_k)\, dF_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1}).^{1}$$

The last definition completes the recursion step.

Three of the preceding definitions give rise to an important technicality. The definitions of f_k and r_k may involve adding $-\infty$ and $+\infty$. So may the definition of $\bar{p}_k$: for any extended-real-valued, measurable function f on a measure space (X, μ),

$$\int f\, d\mu \triangleq \int f^+\, d\mu - \int f^-\, d\mu,$$

where

$$f^+(x) \equiv \max\{f(x), 0\}, \qquad f^-(x) \equiv \max\{-f(x), 0\}.$$

Following the convention adopted in Walkup and Wets (1967), let $+\infty + (-\infty) = -\infty + (+\infty) = +\infty$.[2]

---

[1] A more precise way of writing the integral is

$$\bar{p}_k(z ; \underline{x}_{k-1}) = \int p_k(z ; \underline{x}_{k-1}, x'_k)\, dF_{X_k|\underline{X}_{k-1}}(x'_k|\underline{x}_{k-1}).$$

The shorter notation will be used often.

[2] See Walkup and Wets (1967) for the properties of the integral under this extended definition of integration.

For stage 0 define

$$f_0(u_0; x_0) \equiv c_0(u_0; x_0) + \bar{p}_1(u_0; x_0),$$

$$r_0(u_0; x_0) \equiv \begin{cases} f_0(u_0; x_0), & A_{00}u_0 = b_0(x_0), \quad u_0 \geqq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

The problem at stage 0 is

$$P_0(x_0): \quad \underset{u_0}{\text{minimize}}\, f_0(u_0; x_0)$$

$$\text{subject to } A_{00}u_0 = b_0(x_0), \qquad u_0 \geqq 0.$$

Define $p_0(x_0) \equiv \inf (P_0(x_0))$.

The stochastic programming problem is specified by the sequence of parameterized problems $P_0, \cdots, P_K$. One seeks solutions to the family of problems $\{P_0(x_0); x_0 \in R^{n_0}\}$, which amounts to seeking a function $\bar{u}_0 : R^{s_0} \to R^{n_0}$ such that

$$A_{00}\bar{u}_0(x_0) = b_0(x_0),$$

$$\bar{u}_0(x_0) \geqq 0$$

and

$$f_0(\bar{u}_0(x_0); x_0) = p_0(x_0)$$

for almost every (a.e.) x_0. The decision that must be made here-and-now is u_0; it is made knowing x_0, the realization of X_0. After X_1 is observed, a decision u_1 will be made, taking into account x_0 and the decision u_0 already made. At stage k, X_k will be observed, and then a decision u_k will be made, taking into account the realizations of $X_0, \cdots, X_k$ and the past decisions $u_0, \cdots, u_{k-1}$. The past decisions cannot be amended, but u_k provides a "recourse".

Since X_0 represents the initial state of the world and its realization is known when the initial decision, u_0, is made, it might as well be a constant random variable—i.e., $P\{X_0 = \bar{x}_0\} = 1$ for some $\bar{x}_0 \in R^{s_0}$. Then finding the function $\bar{u}_0$ described above reduces to finding a *vector* $u_0 \in R^{n_0}$ such that

$$A_{00}u_0 = b_0(\bar{x}_0), \quad u_0 \geqq 0 \quad \text{and} \quad f_0(u_0; \bar{x}_0) = p_0(\bar{x}_0).$$

It is convenient to identify $\bar{p}_1, c_0, b_0, f_0, r_0$ and p_0 with their values at $\bar{x}_0$ and to write the stage 0 problem as

$$P_0 : \text{minimize } f_0(u_0), \qquad u_0 \in R^{n_0},$$

$$\text{subject to } A_{00}u_0 = b_0, \qquad u_0 \geqq 0.$$

This is the stochastic programming problem's *equivalent deterministic problem.*

According to Olsen (1975a),

$$\bar{p}_k(z; \underline{x}_{k-1}) = E[p_k(z; \underline{X}_k) | \underline{X}_{k-1} = \underline{x}_{k-1}]$$

for almost every (a.e.) $\underline{x}_{k-1}$ whenever $E[p_k^+(z; \underline{X}_k)] < +\infty$ or $E[p_k^-(z; \underline{X}_k)] < +\infty$; that is, if the positive part or the negative part of $p_k(z; \cdot)$ is summable, the

conditional expectation $E[p_k(z; \underline{X}_k)|\underline{X}_{k-1}]$ exists and $\bar{p}_k(z; \cdot)$ is one version of it. To avoid the nuisance of events with probability 0 in some proofs below, use the version

$$E[g(\underline{X}_k)|\underline{X}_{k-1} = \underline{x}_{k-1}] = \int g(\underline{x}_{k-1}, x'_k)\, dF_{X_k|\underline{X}_{k-1}}(x'_k|\underline{x}_{k-1})$$

whenever the conditional expectation exists (g is any Borel measurable extended-real-valued function). $E[\bar{p}_k(z; \underline{X}_k)] < +\infty$ if, for example, $c_i \geqq 0$ for every $i \geqq k$. Proposition 2.2 gives another condition.

The existence of $E[p_k(z; \underline{X}_k)|\underline{X}_{k-1}]$ is not necessary to any of the results obtained below although when it does not exist, the stochastic programming problem's economic meaningfulness is dubious. If the conditional expectations all exist, $\bar{p}_{k+1}(\underline{u}_k; \underline{x}_k)$ is the expected optimal return from future stages given that realizations $\underline{x}_k$ and decisions $\underline{u}_k$ have occurred. If the decisions $\underline{u}_k$ create a positive probability, given $\underline{X}_k = \underline{x}_k$, of inconsistency in a later stage problem, then $\bar{p}_{k+1}(\underline{u}_k; \underline{x}_k) = +\infty$.

The recursion defining $P_0, \cdots, P_K$ tacitly assumes a measurability property of $p_k(\underline{u}_{k-1}; \cdot)$ for $1 \leqq k \leqq K$. It also assumes the existence of a regular conditional distribution function for X_k given $\underline{X}_{k-1}$. And because this function, if it exists, is ordinarily not unique, one might think that $\inf(P_0)$ can depend on the (perhaps completely arbitrary) choice of $F_{X_k|\underline{X}_{k-1}}$. The existence of $F_{X_k|\underline{X}_{k-1}}$ requires no additional assumptions; it follows from the fact that the values of X_k lie in a complete separable metric space, R^{s_k} (Ash (1972, pp. 263–66)). Perhaps less obvious is that $\inf(P_0)$ is independent of the choice of the regular conditional distribution functions (given the same joint distribution of $X_0, \cdots, X_K$); Theorem 2.1 of Olsen (1975a) shows that it is. The same paper gives sufficient conditions for the requisite measurability property of $\bar{p}_k(\underline{u}_{k-1}, \cdot)$.

Notation. Let $F_{\underline{X}_k}$ be the distribution function of $\underline{X}_k$. If $w \in R^n$, $|w| \triangleq (\sum_{i=1}^n |w_i|^2)^{1/2}$. For $p \in [1, +\infty)$, $L_p(R^{s_k})$ is the space of Borel measurable functions $g : R^{s_k} \to [-\infty, +\infty]$ such that $\|g\|_p < +\infty$;

$$\|g\|_p \triangleq \left(\int |g(\underline{x}_k)|^p\, dF_{\underline{X}_k}(\underline{x}_k) \right)^{1/p}.$$

Let μ_k be the Borel probability measure on R^{s_k} determined by $F_{\underline{X}_k}$. Let $\mu_k(\cdot\,|\underline{x}_{k-1})$ be the Borel probability measure on R^{s_k} determined by $F_{X_k|\underline{X}_{k-1}}(\cdot\,|\underline{x}_{k-1})$.

The proof of Proposition 2.2 uses the following lemma. It is almost certainly well-known, and is a direct consequence of Walkup and Wets (1969a, Lemma 2); a brief proof is given for the sake of completeness.

LEMMA 2.1. *Let A be a real $m \times n$ matrix such that $\{x : Ax = 0, x \geqq 0\} = \{0\}$. Then there is a positive number ρ such that for any $b \in R^m$, $Ax = b$ and $x \geqq 0$ implies $|x| \leqq \rho|b|$.*

Proof. The hypothesis implies that $C \triangleq \{x : Ax = b, x \geqq 0\}$ is the convex hull of its extreme points. The conclusion follows if it is true for every extreme point of C. But an extreme point corresponds to a basic feasible solution of the system $Ax = b$. Let $B_1, \cdots, B_l$ be the basis matrices of A. These matrices have left-

inverses $E_1, \cdots, E_l$. If x is an extreme point of C,

$$|x| \leqq \max_{1 \leqq i \leqq l} |E_i b| \leqq \left(\max_{1 \leqq i \leqq l} \|E_i\| \right) |b|$$

Take $\rho = \max_{1 \leqq i \leqq l} \|E_i\|$.

PROPOSITION 2.2. *Assume for each* $0 \leqq k \leqq K$:

(a) *There are summable functions* $\beta_k : R^{S_k} \to (-\infty, 0]$ *and* $\alpha_k : R^{S_k} \to [-\infty, 0]$ *such that*

$$c_k(\underline{u}_k; \underline{x}_k) \geqq \beta_k(\underline{x}_k)|\underline{u}_k| + \alpha_k(\underline{x}_k)$$

for every $\underline{u}_k$ *and every* $\underline{x}_k$.

(b) $E|\beta_k(\underline{X}_k) b_i(\underline{X}_i)| < +\infty$ *if* $0 \leqq i \leqq k$ ($\underline{X}_i$ *being a subvector of* $\underline{X}_k$).

(c) $\{w : A_{kk} w = 0,\ w \geqq 0\} = \{0\}$.

Then for each $0 \leqq k \leqq K$ *there are summable functions* $\delta_k : R^{S_k} \to (-\infty, 0]$ *and* $\gamma_k : R^{S_k} \to [-\infty, 0]$ *such that*

$$p_k(\underline{u}_{k-1}; \underline{x}_k) \geqq \delta_k(\underline{x}_k)|\underline{u}_{k-1}| + \gamma_k(\underline{x}_k)$$

for every $\underline{u}_{k-1}$ *and every* $\underline{x}_k$. *There are also summable functions* $\bar{\delta}_{k+1} : R^{S_k} \to (-\infty, 0]$ *and* $\bar{\gamma}_{k+1} : R^{S_k} \to [-\infty, 0]$ *such that*

$$\bar{p}_{k+1}(\underline{u}_k; \underline{x}_k) \geqq \bar{\delta}_{k+1}(\underline{x}_k)|\underline{u}_k| + \bar{\gamma}_{k+1}(\underline{x}_k)$$

for every $\underline{u}_k$ *and every* $\underline{x}_k$.

Proof. Let $1 \leqq k \leqq K$. Assume that

$$\bar{p}_{k+1}(\underline{u}_k; \underline{x}_k) \geqq \bar{\delta}(\underline{x}_k)|\underline{u}_k| + \bar{\gamma}(\underline{x}_k)$$

for every $\underline{u}_k$ and every $\underline{x}_k$, where $\bar{\delta}$ and $\bar{\gamma}$ have the nonpositivity, mean and covariance properties of β_k and α_k, respectively, in (a) and (b). Let $\nu = \beta_k + \bar{\delta}$ and $\eta = \alpha_k + \bar{\gamma}$.

Let $B = [A_{k0} \mid \cdots \mid A_{k,k-1}]$. By (c) and Lemma 2.1, there is a positive number ρ such that

$$\begin{pmatrix} B & A_{kk} \\ I & 0 \end{pmatrix} \underline{u}_k = \begin{pmatrix} b_k(\underline{x}_k) \\ z \end{pmatrix}, \qquad \underline{u}_k \geqq 0,$$

implies

$$|\underline{u}_k| \leqq \rho |(b_k(\underline{x}_k), z)|$$

$$\leqq \rho |b_k(\underline{x}_k)| + \rho |z|.$$

Therefore,

$$p_k(z; \underline{x}_k) \geqq \rho \nu(\underline{x}_k)|z| + \rho \nu(\underline{x}_k)|b_k(\underline{x}_k)| + \eta(\underline{x}_k)$$

for every z and every $\underline{x}_k$. Define

$$\delta_k(\underline{x}_k) \equiv \rho \nu(\underline{x}_k)$$

and

$$\gamma_k(\underline{x}_k) \equiv \rho \nu(\underline{x}_k)|b_k(\underline{x}_k)| + \eta(\underline{x}_k).$$

Clearly, $\delta_k(\underline{X}_k)$ is nonpositive, has finite mean, and has finite covariance with $b_1(\underline{X}_1), \cdots, b_k(\underline{X}_k)$. Also, γ_k is nonpositive; since η is summable and $E|\nu(\underline{X}_k)b_k(\underline{X}_k)| < +\infty$, γ_k is summable.

It remains to verify the induction hypothesis for stage $k-1$. Since $E[p_k^-(\underline{u}_{k-1}; \underline{X}_k)] < +\infty$ for any $\underline{u}_{k-1}$, $E[p_k(\underline{u}_{k-1}; \underline{X}_k)|\underline{X}_{k-1}]$ exists, and

$$\bar{p}_k(\underline{u}_{k-1}; \underline{x}_{k-1}) = E[p_k(\underline{u}_{k-1}; \underline{X}_k)|\underline{X}_{k-1} = \underline{x}_{k-1}]$$

for every $\underline{x}_{k-1}$. (The rule on choosing a version of the conditional expectation avoids the qualification "almost every.") Therefore,

$$\bar{p}_k(\underline{u}_{k-1}; \underline{x}_{k-1}) \geqq \hat{\delta}(\underline{x}_{k-1})|\underline{u}_{k-1}| + \hat{\gamma}(\underline{x}_{k-1})$$

for every $\underline{u}_{k-1}$ and every $\underline{x}_{k-1}$, where

$$\hat{\delta}(\underline{x}_{k-1}) \equiv E[\delta_k(\underline{X}_k)|\underline{X}_{k-1} = \underline{x}_{k-1}]$$

and

$$\hat{\gamma}(\underline{x}_{k-1}) \equiv E[\gamma_k(\underline{X}_k)|\underline{X}_{k-1} = \underline{x}_{k-1}].$$

Now

$$E[\hat{\delta}(\underline{X}_{k-1})] = E[E[\delta_k(\underline{X}_k)|\underline{X}_{k-1}]] = E[\delta_k(\underline{X}_k)] < +\infty,$$

and similarly for $\hat{\gamma}$. Nonpositivity is preserved since the selected version of conditional expectation is defined in terms of an integral over a positive measure. For $i \leqq k-1$,

$$\begin{aligned} +\infty &> E|\delta_k(\underline{X}_k)b_i(\underline{X}_i)| \\ &= E[E[|\delta_k(\underline{X}_k)b_i(\underline{X}_i)|\,|\underline{X}_{k-1}]] \\ &= E|\hat{\delta}(\underline{X}_{k-1})b_i(\underline{X}_i)|. \end{aligned}$$

That completes the induction step. The induction hypothesis holds trivially if $k = K$. □

The full notation of the multistage stochastic programming problem is an unnecessary burden in §§ 3 and 4, where properties of p_k and $\bar{p}_k$ are deduced from properties of P_k. To simplify the notation, everything in those sections is stated as though $k = 1$. The convention that X_0 is constant is abrogated, so that for any $0 \leqq k \leqq K$, P_k has the same form as P_1 and $\bar{p}_k$ has the same form as $\bar{p}_1$; thus, there is no loss of generality. To simplify the notation further, let $F_{10} = F_{X_1|X_0}$, $F_1 = F_{\underline{X}_1}$ and $F_0 = F_{X_0}$.

Sections 3 and 4 give conditions under which $\bar{p}_k$ inherits key properties of $\bar{p}_{k+1}$. Section 5 uses the propositions in §§ 3 and 4 inductively to derive properties of $\bar{p}_1$ and hence P_0, the equivalent deterministic problem. Section 3 examines $\bar{p}_k(\cdot\,; \underline{x}_{k-1})$ in terms of convexity and lower-semicontinuity. Section 4 examines it in terms of a Lipschitz property and polyhedrality of $\{z : \bar{p}_k(z; \underline{x}_{k-1}) < +\infty\}$.

3. Convexity and lower-semicontinuity. This section gives sufficient conditions for $\bar{p}_1$ to inherit convexity and lower-semicontinuity properties from $\bar{p}_2$.[3]

[3] See Rockafellar (1970) for the definitions of convexity and lower-semicontinuity of extended-real-valued functions.

The next proposition is used often in the sequel, sometimes without being cited. It specializes the product measure theorem (Ash (1972, p. 98)).

PROPOSITION 3.1. *Let T be a Borel subset of R^{S_1}. Let $T(x_0) \equiv \{x_1' : (x_0, x_1') \in T\}$. Then*

$$\int \int_{T(x_0)} dF_{10}(\cdot \mid x_0)\, dF_0(x_0) = \int_T dF_1.$$

Consequently, if $\mu_1(T) = 1$, $\mu_1(T(x_0)|x_0) = 1$ for a.e. x_0.

PROPOSITION 3.2. *Fix $\underline{x}_1$. If $c_1(\cdot\,;\underline{x}_1)$ and $\bar{p}_2(\cdot\,;\underline{x}_1)$ are convex, so is $p_1(\cdot\,;\underline{x}_1)$.*

Proof. The hypothesis implies that $r_1(\cdot\,;\underline{x}_1)$ is convex. Therefore, by a straightforward argument,

$$p_1(z\,;\underline{x}_1) \equiv \inf_{u_1} r_1(z, u_1\,;\underline{x}_1)$$

is convex in z. □

PROPOSITION 3.3. *Fix x_0. Assume that $p_1(\cdot\,;\underline{x}_1)$ is convex for a.e. x_1 given x_0—i.e., there is a Borel set $T \subset \{x_1' : p_1(\cdot\,; x_0, x_1')$ is convex$\}$ with $\mu_1(T|x_0) = 1$. Then $\bar{p}_1(\cdot\,; x_0)$ is convex.*

Proof. Let $\lambda \in (0, 1)$. Let $z', z'' \in R^{N_0}$.

$$\begin{aligned}
&\bar{p}_1(\lambda z' + (1-\lambda)z''; x_0) \\
&\quad \leqq \int_T (\lambda p_1(z'; \underline{x}_1) + (1-\lambda) p_1(z''; \underline{x}_1))\, dF_{10}(x_1|x_0) \\
&\quad \leqq \lambda \int_T p_1(z'; \underline{x}_1)\, dF_{10}(x_1|x_0) \\
&\qquad + (1-\lambda) \int_T p_1(z''; \underline{x}_1)\, dF_{10}(x_1|x_0) \\
&\quad = \lambda \bar{p}_1(z'; x_0) + (1-\lambda)\bar{p}_1(z''; x_0).
\end{aligned}$$

Since $+\infty + (-\infty) = +\infty$, that demonstrates convexity. □

PROPOSITION 3.4. *Assume that $\{w : A_{11}w = 0,\ w \geqq 0\} = \{0\}$. Fix $\underline{x}_1$. Assume that $c_1(\cdot\,;\underline{x}_1)$ and $\bar{p}_2(\cdot\,;\underline{x}_1)$ are lower-semicontinuous convex functions. Then $p_1(\cdot\,;\underline{x}_1)$ is lower-semicontinuous and convex, or it is identically $-\infty$ on its effective domain* (or both).[4]

A proof is based upon Lemma 2.1 of Walkup and Wets (1969a, Thm. 2), and the fact that the sum of two convex, lower-semicontinuous (l.s.c.) functions is l.s.c. convex or nowhere finite (Rockafellar (1970, p. 77)).

If, in addition to the hypothesis of Proposition 3.4, $c_1(\cdot\,;\underline{x}_1)$ is finite everywhere, $p_1(\cdot\,;\underline{x}_1)$ is l.s.c. and convex, for $\bar{p}_2$ is l.s.c. and convex (perhaps

[4] Let X be an arbitrary set, and let $f : X \to [-\infty, +\infty]$. *The effective domain* of f is $\operatorname{dom} f \triangleq \{x \in X : f(x) < +\infty\}$ (Rockafellar (1970)).

improper).[5] If improper, $\bar{p}_2(\cdot\,;\underline{x}_1)$ is identically $-\infty$ on its effective domain, and its effective domain is closed. But dom $f_1(\cdot\,;\underline{x}_1) = \text{dom}\ \bar{p}_2(\cdot\,;\underline{x}_1)$.

PROPOSITION 3.5. *Fix x_0. If $p_1(\cdot\,;\underline{x}_1)$ is l.s.c. and convex for a.e. x_1 given x_0, and if $\bar{p}_1(u_0; x_0) > -\infty$ for every u_0, then $\bar{p}_1(\cdot\,; x_0)$ is l.s.c. and convex.*

Proof. Let T be a Borel set in R^{S_1} such that $p_1(\cdot\,; x_0, x_1')$ is l.s.c. and convex for every $x_1' \in T$ and $\mu_1(T|x_0) = 1$.

$$\bar{p}_1(u_0; x_0) = \int_T p_1(u_0; \underline{x}_1)\, dF_{10}(x_1|x_0).$$

Apply the lemma in Walkup and Wets (1969b) to $p_1(\cdot\,; x_0, \cdot)$: $R^{N_0} \times T \to [-\infty, +\infty]$. □

4. Lipschitz properties and polyhedrality. This section's propositions characterize $\bar{p}_k$ (by analogy with $\bar{p}_1$) in terms of a Lipschitz property, a boundedness (more accurately, a growth) property, and descriptions of its effective domain—assuming certain properties of P_k. Among the assumed properties of P_k are, of course, properties of $\bar{p}_{k+1}$. Typically, the propositions fail to assert that $\bar{p}_k$ inherits all the properties of $\bar{p}_{k+1}$, and therefore, in contrast to the previous section's propositions, it is difficult to apply them repeatedly to obtain characterizations of P_0. The next section invokes additional assumptions, leading to Lipschitz and polyhedrality characterizations of P_0 in some important special cases.

DEFINITION 4.1. Let $g : R^n \to [-\infty, +\infty]$. g is *Lipschitzian-β* if and only if $\beta \in [0, +\infty)$ and either:

(i) $g(z) = -\infty$ for every $z \in \text{dom}\ g$ or

(ii) g is finite everywhere on dom g and

$$|g(z') - g(z'')| \leqq \beta |z' - z''| \quad \forall\, z', z'' \in \text{dom}\ g.$$

PROPOSITION 4.2. *Fix $\underline{x}_1$. Assume that $c_1(\cdot\,;\underline{x}_1)$ is Lipschitzian-α' and $\bar{p}_2(\cdot\,;\underline{x}_1)$ is Lipschitzian-α''. Assume that* dom $c_1(\cdot\,;\underline{x}_1)$ *and* dom $\bar{p}_2(\cdot\,;\underline{x}_1)$ *are polyhedral convex sets.*[6] *Then $p_1(\cdot\,;\underline{x}_1)$ is Lipschitzian-β, for some β, and its effective domain is a polyhedral convex set.*

A proof relies upon Walkup and Wets (1969a, Thm. 2 (iii)). The proposition is given only for the sake of comparison with Proposition 4.3, which is stronger. The stronger conclusion is necessary to achieve the Lipschitz characterization of $\bar{p}_1$ in Proposition 4.5—one of this section's aims.

PROPOSITION 4.3. *Assume that for every $\underline{x}_1 \in S$, a subset of R^{S_1}, $c_1(\cdot\,;\underline{x}_1)$ is Lipschitzian-$\alpha'(\underline{x}_1)$ and $\bar{p}_2(\cdot\,;\underline{x}_1)$ is Lipschitzian-$\alpha''(\underline{x}_1)$, with α' and α'' in $L_1(R^{S_1})$. Assume also that there are an $l \times N_1$ matrix D and a function $d : R^{S_1} \to R^l$ such that for every $\underline{x}_1 \in S$*

$$\text{dom}\ r_1(\cdot\,;\underline{x}_1) = \{\underline{u}_1 : D\underline{u}_1 \geqq d(\underline{x}_1),\ \underline{u}_1 \geqq 0\}.$$

(*Measurability of d is not assumed.*)

---

[5] An extended-real-valued function is *proper* if it is $-\infty$ nowhere and finite somewhere (Rockafellar (1970)).

[6] See Rockafellar (1970, p. 170) for the definition of "polyhedral convex set."

Then there is a function $\delta \in L_1(R^{S_1})$ *such that, for every* $\underline{x}_1 \in S$, $p_1(\cdot\,;\underline{x}_1)$ *is Lipschitzian-*$\delta(\underline{x}_1)$; *in fact, there is such a function proportional to* $\alpha' + \alpha''$. *Moreover,* dom $p_1(\cdot\,;\underline{x}_1)$ *is a polyhedral convex set for every* $\underline{x}_1 \in S$.

The proof uses the following lemma.

LEMMA 4.4. *Let A be a real* $m \times n$ *matrix. For any b and z in* R^m, *let*

$$C_b(z) = \{x : Ax = b - z,\ x \geqq 0\}.$$

Then there is a positive number β *such that*

$$d(C_b(z'), C_b(z'')) \leqq \beta |z' - z''|$$

for every z', z'' *and b such that* $C_b(z')$ *and* $C_b(z'')$ *are nonempty, where* $d(\cdot\,,\cdot)$ *denotes the Hausdorff distance with respect to the Euclidean norm* (see Berge (1963, p. 126)).

Proof. Lemma 2 of Walkup and Wets (1969a) says: for any given b, there is a $\beta > 0$ such that

$$d(C_b(z'), C_b(z'')) \leqq \beta |z' - z''|$$

for every z' and z'' such that $C_b(z')$ and $C_b(z'')$ are nonempty. Then for any $b' \in R^m$ such that $C_{b'}(z')$ and $C_{b'}(z'')$ are nonempty,

$$\begin{aligned} d(C_{b'}(z'), C_{b'}(z'')) &= d(C_b(z' + b - b'), C_b(z'' + b - b')) \\ &\leqq \beta |z' - z''|. \end{aligned}$$

□

Proof of Proposition 4.3. Let

$$C(z\,;\underline{x}_1) = \{\underline{u}_1 : D\underline{u}_1 \geqq d(\underline{x}_1),\ \underline{u}_{k-1} = z,\ \underline{u}_1 \geqq 0\}.$$

It follows from Lemma 4.4 that there is a $\rho > 0$ such that

$$d(C(z'\,;\underline{x}_1), C(z''\,;\underline{x}_1)) \leqq \rho |z' - z''|$$

for every z', z'' and $\underline{x}_1$ such that both sets are nonempty.

Let $\underline{x}_1 \in S$.

$$\begin{aligned} \text{dom } p_1(\cdot\,;\underline{x}_1) &= \{z : \text{dom } r_1(z, \cdot\,;\underline{x}_1) \neq \varnothing\} \\ &= \{z : C(z\,;\underline{x}_1) \neq \varnothing\}, \end{aligned}$$

a polyhedral convex set. The assumptions on c_1 and $\bar{p}_2$ imply that $f_1(\cdot\,;\underline{x}_1)$ is Lipschitzian-$\beta(\underline{x}_1)$, where $\beta = \alpha' + \alpha''$. There are two possible cases: (i) $f_1(\cdot\,;\underline{x}_1)$ is identically $-\infty$ on its effective domain; or (ii) $f_1(\cdot\,;\underline{x}_1)$ is finite and Lipschitzian with Lipschitz constant $\beta(\underline{x}_1)$ on dom $f_1(\cdot\,;\underline{x}_1)$. In case (i), $p_1(\cdot\,;\underline{x}_1)$ is identically $-\infty$ on its effective domain. In case (ii), the argument parallels the proof of Lemma 3 in Walkup and Wets (1969a). □

The next proposition says that if p_1 has the Lipschitz property described in the conclusion of Proposition 4.3, $\bar{p}_1$ inherits the property. It does *not* say that there are a matrix D and a function d such that

$$\text{dom } r_0(\cdot\,;x_0) = \{u_0 : Du_0 \geqq d(x_0),\ u_0 \geqq 0\}$$

for a.e. x_0, or even that dom $r_0(\cdot\,;x_0)$ is a polyhedral convex set for a.e. x_0.

PROPOSITION 4.5. *Assume that $p_1(\cdot\,;\underline{x}_1)$ is Lipschitzian-$\beta(\underline{x}_1)$ for a.e. $\underline{x}_1$, with $\beta \in L_1(R^{S_1})$. Then there is a function $\delta \in L_1(R^{S_0})$ such that $\bar{p}_1(\cdot\,;x_0)$ is Lipschitzian-$\delta(x_0)$ for a.e. x_0. In fact, $E[\beta(\underline{X}_1)|X_0]$ is such a function.*

Proof. Let S be a Borel subset of R^{S_1} of measure 1 contained in the set of every $\underline{x}_1$ such that $p_1(\cdot\,;\underline{x}_1)$ is Lipschitzian-$\beta(\underline{x}_1)$. Let

$$S(x_0) \equiv \{x_1' : (x_0, x_1') \in S\}.$$

Let T be a Borel subset of R^{S_0} such that $\mu_0(T) = 1$ and $\mu_1(S(x_0)|x_0) = 1$ for every $x_0 \in T$.

Now let $x_0 \in T$. If dom $\bar{p}_1(\cdot\,;x_0) = \varnothing$ or $\bar{p}_1(\cdot\,;x_0)$ is identically $-\infty$ on its effective domain, $\bar{p}_1(\cdot\,;x_0)$ is Lipschitzian-γ for any γ. Therefore, suppose the contrary. Choose u_0' such that $-\infty < \bar{p}_1(u_0';x_0) < +\infty$. Let $u_0'' \in$ dom $\bar{p}_1(\cdot\,;x_0)$.

$$\begin{aligned}
&|\bar{p}_1(u_0';x_0) - \bar{p}_1(u_0'';x_0)| \\
&\qquad = \left| \int (p_1(u_0';\underline{x}_1) - p_1(u_0'';\underline{x}_1))\, dF_{10}(x_1|x_0) \right| \\
&\qquad \leqq \int_{S(x_0)} |p_1(u_0';\underline{x}_1) - p_1(u_0'';\underline{x}_1)|\, dF_{10}(x_1|x_0) \\
&\qquad \leqq \int_{S(x_0)} \beta(\underline{x}_1')|u_0' - u_0''|\, dF_{10}(x_1|x_0) \\
&\qquad = |u_0' - u_0''| \int \beta(\underline{x}_1)\, dF_{10}(x_1|x_0).
\end{aligned}$$

Since $E|\beta(\underline{X}_1)| < +\infty$,

$$E[\beta(\underline{X}_1)|X_0 = x_0] \equiv \int \beta(\underline{x}_1)\, dF_{10}(x_1|x_0).$$

Take $\delta(x_0) \equiv E[\beta(\underline{X}_1)|X_0 = x_0]$. Then $\bar{p}_1(\cdot\,;x_0)$ is Lipschitzian-$\delta(x_0)$ for a.e. x_0.

$$E|\delta(X_0)| = E[\delta(X_0)] = E[E[\beta(\underline{X}_1)|X_0]] = E[\beta(\underline{X}_1)] < +\infty. \qquad \square$$

Proposition 4.5 (in conjunction with 4.3) fails to characterize the effective domain of $\bar{p}_1(\cdot\,;x_0)$; consequently, an attempt to show that f_0 is Lipschitzian by using Propositions 4.3 and 4.5 inductively, starting at stage K, breaks down after one step. The remainder of this section is devoted to identifying conditions on P_1 which imply that dom $\bar{p}_1(\cdot\,;x_0)$ is a polyhedral convex set for a.e. x_0 and hence that $\bar{p}_1$ induces linear constraints at stage 0. The conclusion of Proposition 4.7 is the hypothesis of Proposition 4.8(i). If the conclusion of Proposition 4.8(i) holds and the conclusion of 4.8(ii) holds for a.e. x_0, Proposition 4.10 reveals that dom $\bar{p}_1(\cdot\,;x_0)$ is a polyhedral convex set for a.e. x_0.

DEFINITION 4.6. Let $g : R^{N_j} \times R^{S_k} \to [-\infty, +\infty]$. Suppose that for every $\underline{x}_k \in R^{S_k}$,

$$g(\underline{u}_j;\underline{x}_k) \leqq \beta(\underline{x}_k)|\underline{u}_j| + \alpha(\underline{x}_k)$$

for every $\underline{u}_j \in$ dom $g(\cdot\,;\underline{x}_k)$, where $\beta : R^{S_k} \to [-\infty, +\infty]$, $\alpha : R^{S_k} \to [-\infty, +\infty]$, and β and α belong to $L_1(R^{S_k})$. Then g is *upper-bounded-*(β, α).

Although in the preceding definition $\beta < +\infty$ a.e., complete rigor requires a convention on the value of $(+\infty)\,(0)$. Define $(+\infty)\,(0) = 0$.

PROPOSITION 4.7. *Assume*:

(a) f_1 *is upper-bounded-*(β, α).

(b) *There are an* $l \times N_1$ *matrix* D *and Borel measurable function* $d : R^{s_1} \to R^l$ *such that for a.e.* $\underline{x}_1$,

$$\operatorname{dom} r_1(\cdot\,; \underline{x}_1) = \{\underline{u}_1 : D\underline{u}_1 \geqq d(\underline{x}_1), \underline{u}_1 \geqq 0\}.$$

(c) $E|\beta(\underline{X}_1)d(\underline{X}_1)| < +\infty$.

Then:

(i) p_1 *is upper-bounded-*(δ, γ) *for some* (δ, γ). *Moreover, there is such a* δ *proportional to* β *a.e.* (μ_1).

(ii) $\bar{p}_1$ *is upper-bounded-*$(\bar{\delta}, \bar{\gamma})$ *for some* $(\bar{\delta}, \bar{\gamma})$. *In fact, one can take*

$$\bar{\delta}(x_0) \equiv E[\delta(\underline{X}_1)|X_0 = x_0],$$

$$\bar{\gamma}(x_0) \equiv E[\gamma(\underline{X}_1)|X_0 = x_0].$$

(iii) *For a given* u_0, *let*

$$S(x_0) \equiv \{x_1' : p_1(u_0; x_0, x_1') < +\infty\}.$$

For a.e. x_0, *if* $\mu_1(S(x_0)|x_0) = 1$, $\bar{p}_1(u_0; x_0) < +\infty$.

Proof of (i). Let

$$C(z; \underline{x}_1) \equiv \{\underline{u}_1 : D\underline{u}_1 \geqq d(\underline{x}_1), u_0 = z, \underline{u}_1 \geqq 0\}.$$

Suppose $C(z'; \underline{x}_1')) \neq \varnothing$; if $(z', \underline{x}_1')$ does not exist, the conclusions hold trivially. Choose $\underline{u}_1' \in C(z'; \underline{x}_1')$. By Lemma 4.4, there is a positive number ρ such that

$$d(C(z; \underline{x}_1), C(z'; \underline{x}_1')) \leqq \rho|d(\underline{x}_1) - d(\underline{x}_1')| + \rho|z - z'|$$

for any $(z, \underline{x}_1)$ such that $C(z\,; \underline{x}_1) \neq \varnothing$. Given any such $(z, \underline{x}_1)$, choose $\underline{u}_1$ to a point in $C(z\,; \underline{x}_1)$ closest to $\underline{u}_1'$ in order to verify the following proposition: there are positive numbers ρ and π such that

$$\inf\{|\underline{u}_1| : \underline{u}_1 \in C(z; \underline{x}_1)\} \leqq \rho|z| + \rho|d(\underline{x}_1)| + \pi$$

for every $(z, \underline{x}_1)$ such that $C(z; \underline{x}_1) \neq \varnothing$.

Suppose $\underline{x}_1$ is such that the equation in assumption (b) holds, and suppose $z \in \operatorname{dom} p_1(\cdot\,; \underline{x}_1)$. Let $\underline{u}_1$ be the minimum-norm point in $C(z; \underline{x}_1)$. Then

$$\begin{aligned} p_1(z; \underline{x}_1) &\leqq f_1(\underline{u}_1; \underline{x}_1) \\ &\leqq \beta(\underline{x}_1)|\underline{u}_1| + \alpha(\underline{x}_1) \\ &\leqq \rho\beta(\underline{x}_1)|z| + \rho\beta(\underline{x}_1)|d(\underline{x}_1)| + \beta(\underline{x}_1)\pi + \alpha(\underline{x}_1). \end{aligned}$$

Take $\delta(\underline{x}_1) \equiv \rho\beta(\underline{x}_1)$. Take

$$\gamma(\underline{x}_1) = \rho\beta(\underline{x}_1)|d(\underline{x}_1)| + \pi\beta(\underline{x}_1) + \alpha(\underline{x}_1)$$

if the equation in assumption (b) holds, and take $\gamma(\underline{x}_1) = +\infty$ if not. $E|\gamma(\underline{X}_1)| < +\infty$ because $E|\beta(\underline{X}_1)d(\underline{X}_1)| < +\infty$, $E|\beta(\underline{X}_1)| < +\infty$ and $E|\alpha(\underline{X}_1)| < +\infty$. Of course, δ inherits the properties of β.

Proof of (ii). Suppose $z \in \operatorname{dom} \bar{p}_1(\cdot\,; x_0)$. Then $z \in \operatorname{dom} p_1(\cdot\,; x_0, x_1')$ for a.e. x_1' given x_0.

$$\begin{aligned}\bar{p}_1(z; x_0) &= \int p_1(z; x_0, x_1')\, dF_{10}(x_1'|x_0) \\ &\leqq \int [\delta(x_0, x_1')|z| + \gamma(x_0, x_1')]\, dF_{10}(x_1'|x_0) \\ &\leqq \left[\int \delta(x_0, x_1')\, dF_{10}(x_1'|x_0)\right]|z| \\ &\qquad + \int \gamma(x_0, x_1')\, dF_{10}(x_1'|x_0) \\ &= E[\delta(\underset{\sim}{X}_1)|X_0 = x_0]|z| + E[\gamma(\underset{\sim}{X}_1)|X_0 = x_0].\end{aligned}$$

The last inequality follows directly from the subadditivity of the extended integral used in this paper. (See Walkup and Wets (1967).)

Proof of (iii). Suppose $\mu_1(S(x_0')|x_0') = 1$. Then

$$\begin{aligned}\bar{p}_1(u_0; x_0') &= \int_{S(x_0')} p_1(u_0; x_0', x_1)\, dF_{10}(x_1|x_0') \\ &\leqq \int_{S(x_0')} [\delta(x_0', x_1)|u_0| + \gamma(x_0', x_1)]\, dF_{10}(x_1|x_0') \\ &\leqq \bar{\delta}(x_0')|u_0| + \bar{\gamma}(x_0').\end{aligned}$$

Therefore, the set of every x_0 such that $\mu_1(S(x_0)|x_0) = 1$ and $\bar{p}_1(u_0; x_0) = +\infty$ has measure 0. □

PROPOSITION 4.8.[7] *Define*

$K^s(x_0) \equiv \operatorname{dom} \bar{p}_1(\cdot\,; x_0)$,

$K^\mu(x_0) \equiv \{u_0 : P\{p_1(u_0; \underset{\sim}{X}_1) < +\infty | X_0 = x_0\} = 1\}$,

$K^p(x_0) \equiv \{u_0 : p_1(u_0\,; \underset{\sim}{x}_1) < +\infty \ \forall\, x_1 \in \tilde{\Xi}_1(x_0)\}$,

where $\tilde{\Xi}_1(x_0)$ *is the support of* $\mu_1(\cdot\,|x_0)$*—i.e., the smallest closed set* T *such that* $\mu_1(T|x_0) = 1$.

(i) *If the hypothesis of Proposition* 4.7 *holds,*

$$K^s(x_0) = K^\mu(x_0) \quad \textit{for a.e. } x_0.$$

(ii) *For a given* x_0, *assume the existence of a matrix* $D = (D_1 \mid D_2)$ *and a continuous function* $d : R^{S_1} \to R^l$ *such that for every* $x_1 \in \tilde{\Xi}_1(x_0)$,

$$\operatorname{dom} r_1(\cdot\,; \underset{\sim}{x}_1) = \{\underset{\sim}{u}_1 : D_1 u_1 + D_2 u_0 \geqq d(\underset{\sim}{x}_1),\ \underset{\sim}{u}_1 \geqq 0\}.$$

Then $K^\mu(x_0) = K^p(x_0)$.

Proof of (i). Proposition 4.7(iii).

Proof of (ii). Let $R = (D_1 \mid -I)$.

$$\begin{aligned}K^p(x_0) &= \{u_0 : \forall\, x_1' \in \tilde{\Xi}_1(x_0)\ \exists\, u_1' \ni r_1(u_0, u_1'; x_0, x_1') < +\infty\} \\ &= \{u_0 : d(x_0, x_1') - D_2 u_0 \in \operatorname{cone} R\ \forall\, x_1' \in \tilde{\Xi}_1(x_0),\ u_0 \geqq 0\},\end{aligned}$$

[7] The definitions of K^p, K^μ and K^s and the proof of (ii) parallel Wets (1974, Thm. 4.1). In the case under consideration, where A is fixed, that theorem is a special case of Proposition 4.8.

where cone $R \triangleq \{y : y = Rw, \text{ some } w \geqq 0\}$.

$$K^{\mu}(x_0) = \{u_0 : d(x_0, x_1') - D_2 u_0 \in \text{cone } R \text{ for a.e. } x_1' \text{ given } x_0,\ u_0 \geqq 0\}.$$

Clearly, $K^{p}(x_0) \subset K^{\mu}(x_0)$. Let $u_0 \in K^{\mu}(x_0)$. The continuity of d implies that

$$S \triangleq \{x_1 : d(x_0, x_1) - D_2 u_0 \in \text{cone } R\}$$

is closed. Since $u_0 \in K^{\mu}(x_0)$, $\mu_1(S|x_0) = 1$. Therefore, $\tilde{\Xi}_1(x_0) \cap S$ is closed, and

$$\mu_1(\tilde{\Xi}_1(x_0) \cap S | x_0) = 1.$$

But $\tilde{\Xi}_1(x_0)$ is the smallest such set. Hence, $S \supset \tilde{\Xi}_1(x_0)$, and $u_0 \in K^{p}(x_0)$. □

Example 4.9 (Eisner (1970, p. 65)). ($K^{\mu} = K^{s} \neq K^{p}$). Let X_1 be a random variable uniformly distributed on [0, 1]. Let

$$A = \begin{pmatrix} -1 & 0 & -1 & 0 \\ -1 & -1 & 0 & -1 \end{pmatrix},$$

$$b_0 = -1, \qquad b_1(x_1) = \begin{cases} 0, & x_1 = 0, \\ -1, & x_1 \neq 0. \end{cases}$$

Notice that b_1 is not continuous. Let

$$c_0(u_0) \equiv -u_0, \qquad c_1(u_1; x_1) \equiv 0.$$

Thus, the equivalent deterministic problem, P_0, is

$$\underset{0 \leqq u_0 \leqq 1}{\text{minimize}} \; -u_0 + E[\inf\{0; 0 \leqq u_1 \leqq -b_1(X_1) - u_0\}].$$

Since X_0 is constant in this example, $\tilde{\Xi}_1(x_0)$ is identified with $\Xi = [0, 1]$, the support of X_1.

$$K^{\mu} \equiv \{u_0 : 0 \leqq u_0 \leqq -b_1(x_1) \text{ for a.e. } x_1\} = [0, 1].$$

$$K^{s} \equiv \{u_0 : E[p_1(u_0, X_1)] < +\infty\} = [0, 1].$$

$$K^{p} \equiv \{u_0 : 0 \leqq u_0 \leqq -b_1(x_1)\ \forall\, x_1 \in [0, 1]\} = \{0\}.$$

The unique solution to P_0 is $u_0 = 1$, which yields the stage 1 constraint $u_1 \leqq -b_1(x_1) - 1$, which can be satisfied (with $u_1 \geqq 0$) almost everywhere but not for every possible realization of X_1—i.e., for every $x_1 \in \Xi$.

The failure of K^{p} to agree with K^{μ} in the example is closely related to critical anomalies in stochastic programming duality and discretization theories. Typically, these theories formulate the SP problem as mathematical programming in an L_p space, the variables being functions of $X_0, \cdots, X_K$. The constraints, which involve functions of $X_0, \cdots, X_K$, are required to hold a.e.(μ), where μ is the Borel probability measure determined by X_K. A major nuisance is the possible nonexistence of an optimal program that satisfies the constraints everywhere on the support of μ instead of just almost everywhere. (See Rockafellar (1975) and Olsen (1975b).)

PROPOSITION 4.10.[8] *Define* $K^p(x_0)$ *as in Proposition* 4.8. *Assume the hypothesis of Proposition* 4.8(ii) *except for continuity of* d. *Then* $K^p(x_0)$ *is a polyhedral convex set.*

Proof. Using the definitions in the proof of Proposition 4.8, let

$$\tilde{K}^p(\underset{\sim}{x}_1) = \operatorname{dom} p_1(\,\cdot\,;\underset{\sim}{x}_1)$$
$$= \{u_0 : d(\underset{\sim}{x}_1) - D_2 u_0 \in \operatorname{cone} R,\ u_0 \geqq 0\}.$$

Now

$$d(\underset{\sim}{x}_1) - D_2 u_0 \in \operatorname{cone} R \quad \Leftrightarrow \quad d(\underset{\sim}{x}_1) - S_2 u_0 \in (\operatorname{cone} R)^{**},$$

by Rockafellar (1970, Thm. 14.1).[9] But

$$d(\underset{\sim}{x}_1) - D_2 u_0 \in (\operatorname{cone} R)^{**}$$
$$\Leftrightarrow \quad \langle d(\underset{\sim}{x}_1) - D_2 u_0,\ w_i^* \rangle \geqq 0$$

for $i = 1, \cdots, n$, where $w_1^*, \cdots, w_n^*$ generate $(\operatorname{cone} R)^*$. The last statement is equivalent to

$$\langle u_0, D_2^t w_i^* \rangle \leqq \langle d(\underset{\sim}{x}_1), w_i^* \rangle \quad \text{for } i = 1, \cdots, n.$$

Therefore,

$$K^p(x_0) = [\cap_{[x_1' \in \tilde{\Xi}_1(x_0)]} \tilde{K}^p(x_0, x_1')]$$
$$= \{u_0 : \langle u_0, D_2^t w_i^* \rangle \leqq \inf \{\langle d(x_0, x_1'), w_i^* \rangle : x_1' \in \tilde{\Xi}_1(x_0)\}$$
$$\text{for } i = 1, \cdots, n\} \cap \{u_0 : u_0 \geqq 0\},$$

which is a polyhedral convex set (empty unless the infimum is finite for $i = 1, \cdots, n$). □

5. Characterizing the equivalent deterministic problem. This section inductively applies the propositions in §§ 3 and 4 to obtain characterizations of the equivalent deterministic problem, given properties of c_k, b_k and $A_{k0}, \cdots, A_{kk}$ for $0 \leqq k \leqq K$. (The convention that X_0 is constant is restored, so that P_0 is identical to the equivalent deterministic problem.)

The first major result is Theorem 5.2, which gives a sufficient condition for $\bar{p}_1$ to be l.s.c. and convex. An argument based on Theorem 1(i) of Walkup and Wets (1969a) shows that if $\bar{p}_1$ and c_0 are l.s.c. convex functions, if $\{u_0 : A_{00} u_0 = b_0,\ u_0 \geqq 0\} = \{0\}$, and if inf (P_0) is finite, then P_0 is solvable ($r_0(u_0) = \inf (P_0) < +\infty$ for some u_0) and dualizable (its dual problem, defined as in Rockafellar (1967) or Van Slyke and Wets (1968), has the same optimal value).

Corollaries 5.7 and 5.10 and Theorem 5.12 give sufficient conditions for P_0's feasible region to be a polyhedral convex set and for f_0 to be either finite and Lipschitzian or identically $-\infty$ on it. Therefore, with the additional assumption that inf (P_0) is finite, they give sufficient conditions for P_0 to be stable (Walkup and

---

[8] Conceptually identical results have appeared in several places—e.g., Walkup and Wets (1969a), Wets (1966c), Wets (1974).

[9] If $S \subset R^n$, the *dual cone*, S^*, is $\{w^* \in R^n : \langle w, w^* \rangle \geqq 0\ \forall\ w \in S\}$. $S^{**} \triangleq (S^*)^*$. Rockafellar (1970) uses "S*" to denote the *polar cone*, $\{w^* \in R^n : \langle w, w^* \rangle \leqq 0\ \forall\ w \in S\}$.

Wets (1969a, Thm. 1(ii))). (P_0 is *stable* if its dual has the same optimal value and is solvable (Rockafellar (1967), Van Slyke and Wets (1968)).)

THEOREM 5.1. *Assume that for each* $0 \leqq k \leqq K$ $c_k(\cdot\,;\underline{x}_k)$ *is convex for a.e.* $\underline{x}_k$. *Then for each* $1 \leqq k \leqq K$, $p_k(\cdot\,;\underline{x}_k)$ *is convex for a.e.* $\underline{x}_k$, $\bar{p}_k(\cdot\,;\underline{x}_{k-1})$ *is convex for a.e.* $\underline{x}_{k-1}$, *and* P_0 *is a convex programming problem.*

Proof. Let $0 \leqq k \leqq K$. Assume that $\bar{p}_{k+1}(\cdot\,;\underline{x}_k)$ is convex for a.e. $\underline{x}_k$. Then by Proposition 3.2, $p_k(\cdot\,;\underline{x}_k)$ is convex for a.e. $\underline{x}_k$. This implies by Proposition 3.1 that for a.e. $\underline{x}_{k-1}$ $p_k(\cdot\,;\underline{x}_k)$ is convex for a.e. x_k given $\underline{x}_{k-1}$. By Proposition 3.3, $\bar{p}_k(\cdot\,;\underline{x}_{k-1})$ is convex for a.e. $\underline{x}_{k-1}$, to complete the induction step. The induction hypothesis holds trivially if $k = K$. □

THEOREM 5.2. *Assume that for each* $1 \leqq k \leqq K$ $\{w : A_{kk}w = 0,\ w \geqq 0\} = \{0\}$, $c_k(\cdot\,;\underline{x}_k)$ *is a l.s.c. convex function for a.e.* $\underline{x}_k$, *and* $\bar{p}_k(\cdot\,;\underline{x}_{k-1}) > -\infty$ *for a.e.* $\underline{x}_{k-1}$. *Then for each* $1 \leqq k \leqq K$: (i) $p_k(\cdot\,;\underline{x}_k)$ *is l.s.c. and convex and greater than* $-\infty$ *for a.e.* $\underline{x}_k$; *and* (ii) $\bar{p}_k(\cdot\,;\underline{x}_{k-1})$ *is l.s.c. and convex for a.e.* $\underline{x}_{k-1}$. *In particular,* $\bar{p}_1$ *is l.s.c. and convex.*

Proposition 2.2 gives a condition guaranteeing the theorem's assumption that for each $0 \leqq k < K$ $\bar{p}_{k+1}(\cdot\,;\underline{x}_k) > -\infty$ for a.e. $\underline{x}_k$; $c_k \geqq 0$ for each $1 \leqq k \leqq K$ also suffices. If the assumption is not satisfied, but the hypothesis of Theorem 5.1 is satisfied, there is a set T of positive measure such that $\underline{x}_k \in T$ implies $\bar{p}_{k+1}(\cdot\,;\underline{x}_k)$ is a convex function having the value $-\infty$ somewhere. Let $\underline{x}_k \in T$. Then $\bar{p}_{k+1}(\cdot\,;\underline{x}_k)$ is $-\infty$ everywhere on the relative interior of its effective domain (Rockafellar (1970, p. 53)). In that case, if $\bar{p}_{k+1}(\underline{u}'_k;\underline{x}_k)$ is finite, an arbitrarily small perturbation of $\underline{u}'_k$, say $\underline{u}''_k$, yields $\bar{p}_{k+1}(\underline{u}''_k;\underline{x}_k) = -\infty$—a situation atypical of well formulated physical-world problems.

Proof of Theorem 5.2. Convexity is immediate from Theorem 5.1.

Let $1 \leqq k \leqq K$. Assume that $\bar{p}_{k+1}(\cdot\,;\underline{x}_k)$ is l.s.c. and convex for a.e. $\underline{x}_k$. By Proposition 3.4, for a.e. $\underline{x}_k$, $p_k(\cdot\,;\underline{x}_k)$ is l.s.c. convex or identically $-\infty$ on its effective domain.

Suppose the latter alternative holds on a set $S \subset R^{S_k}$ with $\mu_k(S) > 0$. Let $S(\underline{x}_{k-1}) \equiv \{x'_k : (\underline{x}_{k-1}, x'_k) \in S\}$. By Proposition 3.1, $\mu_k(S(\underline{x}_{k-1})|\underline{x}_{k-1}) > 0$ on a set $T \subset R^{S_{k-1}}$ with $\mu_{k-1}(T) > 0$. It follows that for every $\underline{x}_{k-1} \in T$, $\bar{p}_k(\cdot\,;\underline{x}_{k-1})$ is identically $-\infty$ on its effective domain. Since $\bar{p}_k(\cdot\,;\underline{x}_{k-1}) > -\infty$ for a.e. $\underline{x}_{k-1}$ by assumption, dom $p_k(\cdot\,;\underline{x}_k) = \varnothing$ for a.e. $\underline{x}_k \in S$. Thus, for a.e. $\underline{x}_k$, $p_k(\cdot\,;\underline{x}_k)$ is l.s.c. and convex (perhaps identically $+\infty$). If, for a particular $\underline{x}_k$, $p_k(\cdot\,;\underline{x}_k)$ is l.s.c. convex and is $-\infty$ somewhere, it is $-\infty$ everywhere on its effective domain. Consequently, $p_k(\cdot\,;\underline{x}_k) > -\infty$ for a.e. $\underline{x}_k$.

By Proposition 3.5 together with Proposition 3.1, $\bar{p}_k(\cdot\,;\underline{x}_{k-1})$ is l.s.c. and convex for a.e. $\underline{x}_{k-1}$, which completes the induction step. The induction hypothesis holds trivially if $k = K$. □

Walkup and Wets (1969b) obtain the following result, which applies only in the case of $K = 1$ and $c_1(\cdot\,;\underline{x}_1)$ linear, but which allows A_{10} and A_{11} to depend on x_1.

THEOREM 5.3. *Let* $K = 1$. *Assume that* c_0 *is linear and* $c_1(\cdot\,;x_1)$ *is linear for a.e.* x_1. *Then* f_0 *is l.s.c. and convex if it is nowhere* $-\infty$.

The next theorem ties together many of the propositions from §§ 3 and 4. It applies only in the case of $K = 1$. The theorems following it achieve similar conclusions in the multistage case, at the cost of more stringent assumptions.

THEOREM 5.4. *Let* $K=1$ *and* $c_1(\underline{u}_1;\underline{x}_1)\equiv\langle u_1, q(x_1)\rangle$, *where* $q: R^{s_1}\to R^{n_1}$ *belongs to* $L_1^{n_1}(R^{s_1})$.

(i) *If* c_0 *is convex, then* f_0 *is convex and* P_0 *is a convex programming problem.*

(ii) *If* c_0 *is Lipschitzian-*β_0, f_0 *is Lipschitzian-*α_0 *for some* α_0.

(iii) *If* $E|q_j(X_1)b_{1i}(X_1)|<+\infty$ *for each* $1\leqq j\leqq n_1$ *and* $1\leqq i\leqq m_1$, *if* b_1 *is continuous, and if* dom c_0 *is a polyhedral convex set, then* dom f_0 *is a polyhedral convex set. In that case* P_0*'s constraints (implicit as well as explicit) are linear.*

Proof of (i). This follows immediately from Theorem 5.1.

Proof of (ii). Take $\beta(x_1)\equiv|q(x_1)|$. $c_1(\cdot\,;x_1)$ is Lipschitzian-$\beta(x_1)$ and dom $c_1(\cdot\,;x_1)=R^{n_1}$ for every x_1. By Proposition 4.3, $p_1(\cdot\,;x_1)$ is Lipschitzian-$\rho\beta(x_1)$ for every x_1, for some $\rho>0$. By Proposition 4.5, $\bar{p}_1$ is Lipschitzian-$\rho E[\beta(X_1)]$.

Proof of (iii). Take $\beta(x_1)\equiv|q(x_1)|$, $\alpha(x_1)\equiv 0$. f_1 is upper-bounded-(β,α). The covariance condition in (iii) implies assumption (c) of Proposition 4.7 by a straightforward argument. Thus, the hypothesis of Proposition 4.7 holds at stage 1. Then by Proposition 4.8 together with Proposition 4.10, dom $\bar{p}_1$ is a polyhedral convex set. Since the intersection of two polyhedral convex sets is a polyhedral convex set, (iii) is proved. □

The preceding theorem resembles Theorem 2.2 of Walkup and Wets (1970). Part (i) is essentially a special case of Theorem 4.4 of Walkup and Wets (1967), which requires b_1 to be linear but allows A_{10} and A_{11} to depend upon x_1. Parts (ii) and (iii), combined, differ from Walkup and Wets' (1967) Proposition 3.16 and Theorem 4.5, combined, principally in substituting a covariance condition for the requirement that $q\in L_2^{n_1}(R^{s_1})$ and $b_1\in L_2^{m_1}(R^{s_1})$. The proof of part (iii) basically consists of showing that dom $\bar{p}_1=K^P$ and that K^P is a polyhedral convex set; cf. Corollary 4.5 and Theorem 4.10 of Wets (1974).

THEOREM 5.5. *Let* Ξ *be the support of* μ_K, *the Borel probability measure determined by* $\underline{X}_K$. (Ξ *is the smallest closed set of measure* 1.) *For* $1\leqq k\leqq K$, *let* Ξ_k *be the projection of* Ξ *on* R^{S_k}, *let* $\Xi_k(\underline{x}_{k-1})$ *be the section* $\{x'_k:(\underline{x}_{k-1},x'_k)\in\Xi\}$, *and let* $\tilde{\Xi}_k(\underline{x}_{k-1})$ *be the support of* $\mu_k(\cdot\,|\underline{x}_{k-1})$.

Assume that Ξ *is compact. For each* $1\leqq k\leqq K$ *assume*:

(a) *For a.e.* $\underline{x}_k$,

$$\text{dom } c_k(\cdot\,;\underline{x}_k)\supset\{\underline{u}_k : A_k\underline{u}_k=b_k(\underline{x}_k),\ \underline{u}_k\geqq 0\}.$$

(b) c_k *is upper-bounded-*(β_k,α_k).

(c) b_k *is continuous on* Ξ_k.

(d) $\tilde{\Xi}_k$ *is a continuous mapping from* Ξ_{k-1} *into the space of closed, nonempty subsets of* R^{s_k}, *equipped with the Hausdorff metric with respect to the Euclidean norm.*

(e) $\tilde{\Xi}_k(\underline{x}_{k-1})\subset\Xi_k(\underline{x}_{k-1})$ *for every* $\underline{x}_{k-1}\in\Xi_{k-1}$.

Also assume dom $c_0\supset\{u_0 : A_{00}u_0=b_0,\ u_0\geqq 0\}$.

Then for each $0\leqq k\leqq K$, *there are a matrix* D *and a function* d *continuous on* Ξ_k *such that*

$$\text{dom } r_k(\cdot\,;\underline{x}_k)=\{\underline{u}_k : D\underline{u}_k\geqq d(\underline{x}_k),\ \underline{u}_k\geqq 0\}$$

for a.e. $\underline{x}_k$. *In particular,* dom r_0 *is a polyhedral convex set.*

If dom $c_k(\cdot\,;\underline{x}_k)=\{\underline{u}_k : D_k\underline{u}_k \geqq d_k(\underline{x}_k)\}$ for a.e. $\underline{x}_k$, for some matrix D_k and continuous function d_k, the constraints $D_k\underline{u}_k \geqq d_k(\underline{x}_k)$ can be adjoined to the constraints $A_k\underline{u}_k = b_k(\underline{x}_k)$ at stage k (possibly requiring expansion of $\underline{u}_k$ to include slacks), and assumption (a) will hold. Also, there is no harm in redefining $c_k(\cdot\,;\underline{x}_k)$ outside $\{\underline{u}_k : A_k\underline{u}_k = b_k(\underline{x}_k),\ \underline{u}_k \geqq 0\}$ in order to get (b) to hold.

Since Ξ has measure 1, so has Ξ_k, and therefore, $\mu_k(\Xi_k(\underline{x}_{k-1})\,|\,\underline{x}_{k-1})=1$ for a.e. $\underline{x}_{k-1}$. Since $\Xi_k(\underline{x}_{k-1})$ is closed for every $\underline{x}_{k-1}$, this implies that $\tilde{\Xi}_k(\underline{x}_{k-1})\subset \Xi_k(\underline{x}_{k-1})$ for a.e. $\underline{x}_{k-1}$. Thus, $F_{X_k|\underline{X}_{k-1}}(\cdot\,|\underline{x}_{k-1})$ can be redefined on a subset of $R^{S_{k-1}}$ of measure 0 in such a way that it is still a regular conditional distribution function for X_k given $\underline{X}_{k-1}$ and $\tilde{\Xi}_k(\underline{x}_{k-1})\subset \Xi_k(\underline{x}_{k-1})$ for every $\underline{x}_{k-1}\in\Xi_{k-1}$. Substituting the redefined $F_{X_k|\underline{X}_{k-1}}$ in the definition of $\bar{p}_k$ would alter the function $r_i(\cdot\,;\underline{x}_i)$, for $0\leqq i<k$, only on a set of $\underline{x}_i$'s of measure 0, which would be irrelevant to the theorem's conclusion. (Note, however, that (d) must hold after the redefinition in order to apply the theorem.)

The proof of Theorem 5.5 uses the following lemma.

LEMMA 5.6. *Let X and Y be metric spaces. Let $\mathcal{Y}$ be the collection of nonempty, compact sets in Y; equip $\mathcal{Y}$ with the Hausdorff metric* (see Berge (1963, p. 126)). *Let Γ be a continuous mapping from X into $\mathcal{Y}$, and let $\phi : X\times Y\to R$ be continuous. Then the function $\psi : X\to R$ defined as*

$$\psi(x)\equiv \inf\{\phi(x,y) : y\in\Gamma(x)\}$$

is continuous.

Proof. Apply Theorem 1 of Berge (1963, p. 126) and then Theorems 1 and 2 of Berge (1963, pp. 115–16). □

Proof of Theorem 5.5. Let $1\leqq k\leqq K$. Assume:

(i) $\bar{p}_{k+1}$ is upper-bounded-$(\hat{\beta},\hat{\alpha})$;

(ii) dom $r_k(\cdot\,;\underline{x}_k)=\{\underline{u}_k : D\underline{u}_k\geqq d(\underline{x}_k),\ \underline{u}_k\geqq 0\}$ for every $\underline{x}_k\in\Xi_k$, where D is a matrix and d is a continuous function on Ξ_k.

Ξ_k is compact, and $\mu_k(\Xi_k)=1$. Hence, d is bounded on Ξ_k, and, since $E|\hat{\beta}(\underline{X}_k)|<+\infty$, $E|\hat{\beta}(\underline{X}_k)d(\underline{X}_k)|<+\infty$. This together with assumption (b) implies by Proposition 4.7 that $\bar{p}_k$ is upper-bounded-$(\bar{\delta},\bar{\gamma})$ for some $(\bar{\delta},\bar{\gamma})$, so that induction assumption (i) holds at stage $k-1$.

The next step involves a minor modification of Proposition 4.8(ii), precipitated by a modification of the set S that occurs in its proof. Let $\underline{x}_{k-1}\in\Xi_{k-1}$. In the proof of 4.8(ii), let

$$S=\{x'_k : d(\underline{x}_{k-1},x'_k)-D_2\underline{u}_{k-1}\in \text{cone } R\}\cap\Xi_k(\underline{x}_{k-1}).$$

S is closed since d is continuous on $\Xi_k(\underline{x}_{k-1})$. Moreover, since $\mu_k(\Xi_k(\underline{x}_{k-1})|\underline{x}_{k-1})=1$, $\mu_k(S|\underline{x}_{k-1})=1$. The rest of the proof of 4.8(ii) is unchanged. It follows that $K^{\mu}(\underline{x}_{k-1})=K^{p}(\underline{x}_{k-1})$ for every $\underline{x}_{k-1}\in\Xi_{k-1}$.

By Proposition 4.8(i), $K^{\mu}(\underline{x}_{k-1})=K^{s}(\underline{x}_{k-1})$ for a.e. $\underline{x}_{k-1}$. Redefine $\bar{p}_k$ on the exceptional set according to

$$\bar{p}_k(\underline{u}_{k-1};\underline{x}_{k-1})=\begin{cases}0, & \underline{u}_{k-1}\in K^{\mu}(\underline{x}_{k-1}),\\ +\infty & \text{otherwise.}\end{cases}$$

(Redefine $\bar{\gamma}$ to be identically 0, for example, on the exceptional set.) This redefinition does not alter $\bar{p}_1$ since it occurs on a set of measure 0. After the redefinition, dom $\bar{p}_k(\cdot\,;\underline{x}_{k-1})=K^{p}(\underline{x}_{k-1})$ for every $\underline{x}_{k-1}\in\Xi_{k-1}$.

The proof of Proposition 4.10 reveals the existence of vectors $\hat{w}_1, \cdots, \hat{w}_n$, a matrix $\hat{R}$ and a function d continuous on Ξ_k such that for every $\underline{x}_{k-1} \in \Xi_{k-1}$,

$$K^p(\underline{x}_{k-1}) = \{\underline{u}_{k-1} : \langle \underline{u}_{k-1}, \hat{R}\hat{w}_i \rangle \leqq \hat{d}_i(\underline{x}_{k-1}) \text{ for } i = 1, \cdots, n, \underline{u}_{k-1} \geqq 0\},$$

where $\hat{d}_i(\underline{x}_{k-1}) = \inf\{\langle d(\underline{x}_k), \hat{w}_i \rangle : x_k \in \tilde{\Xi}_k(\underline{x}_{k-1})\}$. Now apply Lemma 5.6: take $X = \Xi_{k-1}$, $Y = R^{s_k}$, $\phi = \langle d(\cdot), \hat{w}_i \rangle$ (some $1 \leqq i \leqq n$), $\Gamma = \tilde{\Xi}_k$. ϕ is continuous on Ξ_k, which contains

$$\{(\underline{x}_{k-1}, x'_k) : \underline{x}_{k-1} \in \Xi_{k-1}, x'_k \in \tilde{\Xi}_k(\underline{x}_{k-1})\};$$

Γ is continuous by assumption (d). By the lemma, $\hat{d}_i$ is continuous on Ξ_{k-1} for each i. Consequently, there is a matrix $\hat{D}$ such that

$$\operatorname{dom} \bar{p}_k(\cdot\,; \underline{x}_{k-1}) = \{\underline{u}_{k-1} : \hat{D}\underline{u}_{k-1} \geqq -\hat{d}(\underline{x}_{k-1}), \underline{u}_{k-1} \geqq 0\}$$

for every $\underline{x}_{k-1} \in \Xi_{k-1}$, where $\hat{d} = (\hat{d}_1, \cdots, \hat{d}_n)^t$ is continuous on Ξ_{k-1}. Given assumption (c), the only remaining obstacle to verifying the induction hypothesis for stage $k-1$ is the qualification "*almost* every" in assumption (a). But just as $\bar{p}_k$ was redefined innocuously, c_{k-1} can be redefined to get the desired conclusion: take $c_{k-1}(\cdot\,; \underline{x}_{k-1}) \equiv 0$ for every $\underline{x}_{k-1}$ in the set of measure 0 where the inclusion in (a) fails to hold. After the redefinition, the induction hypothesis holds at stage $k-1$. It clearly holds at stage K once c_K is suitably redefined. □

COROLLARY 5.7. *In addition to the hypothesis of Theorem* 5.5, *assume that for each* $0 \leqq k \leqq K$, $c_k(\cdot\,; \underline{x}_k)$ *is Lipschitzian-*$\lambda_k(\underline{x}_k)$ *for a.e.* x_k, *with* $\lambda_k \in L_1(R^{S_k})$. *Then* P_0*'s feasible region is a polyhedral convex set, and on it* f_0 *is either finite and Lipschitzian or identically* $-\infty$.

Proof. Use the conclusion of Theorem 5.5 in an inductive application of Propositions 4.3 and 4.5. □

The conclusion of Theorem 5.5 implies that for each $0 \leqq k \leqq K$, dom $r_k(\cdot\,; \underline{x}_k)$ is a polyhedral convex set for a.e. $\underline{x}_k$. The next theorem gives a sufficient condition for polyhedrality to hold for every (not just almost every) $\underline{x}_k \in \Xi_k$. The stronger conclusion is useful in the study of discretizations of stochastic programming problems (see Olsen (1975b)).

THEOREM 5.8. *In addition to the hypothesis of Theorem* 5.5., *assume for each* $1 \leqq k \leqq K$:

(a′) dom $c_k(\cdot\,; \underline{x}_k) \supset \{\underline{u}_k : A_k\underline{u}_k = b_k(\underline{x}_k), \underline{u}_k \geqq 0\}$ $\forall$ $\underline{x}_k \in \Xi_k$.

(b′) c_k *is upper-bounded-*(β_k, α_k), *and* β_k *and* α_k *are bounded on* Ξ_k.

Then for each $0 \leqq k \leqq K$ *there are a matrix* D *and a function* d *continuous on* Ξ_k *such that*

$$\operatorname{dom} r_k(\cdot\,; \underline{x}_k) = \{\underline{u}_k : D\underline{u}_k \geqq d(\underline{x}_k), \underline{u}_k \geqq 0\}$$

for every $\underline{x}_k \in \Xi_k$.

Proof. In light of Theorem 5.5 and its proof, it suffices to show that the redefinitions of $\bar{p}_k$ and c_{k-1} in the induction step are unnecessary. Assumption (a′) removes any need to redefine c_{k-1}. Let $1 \leqq k \leqq K$. In addition to (i) and (ii) of the proof of Theorem 5.5, assume that $\hat{\beta}$ and $\hat{\alpha}$ are bounded on Ξ_k. Then by Proposition 4.7(ii) and assumption (e) of Theorem 5.5 (and the way the conditional expectations are defined in terms of integrals over a positive measure), $\bar{\delta}$ and $\bar{\gamma}$ (defined as in 4.7(ii)) are bounded on Ξ_{k-1}. Now consider the proof of

Proposition 4.7(iii). It says that if $\bar{\delta}$ and $\bar{\gamma}$ are finite everywhere on Ξ_{k-1}, $\operatorname{dom} \bar{p}_k(\cdot\,; \underline{x}_{k-1}) = K^{\mu}(\underline{x}_{k-1})$ for every $\underline{x}_{k-1} \in \Xi_{k-1}$, and consequently, the redefinition of $\bar{p}_k$ in the proof of 5.5 is unnecessary. The induction hypothesis is true by implication at stage $k-1$. It is clearly true at stage K. □

All that was needed in the induction step was $\bar{\delta}$ and $\bar{\gamma}$ finite on Ξ_{k-1}, but that in itself would not assure finiteness at the next stage in the induction. Boundedness is a convenient property for that purpose; in a particular problem, other properties with the same effect might exist.

The next theorem and its corollary deal with the important special case in which the random variables in any stage are independent of the random variables in earlier stages. They are fairly straightforward consequences of propositions in § 4. Wets (1972, Corollary 4.1) outlines a direct proof of a similar result.

THEOREM 5.9. *Assume for each* $1 \leqq k \leqq K$:

(a) X_k *is independent of* $\underline{X}_{k-1}$.

(b) c_k *and* b_k *do not depend on* $\underline{x}_{k-1}$; *thus, write* $c_k(\underline{u}_k\,; x_k)$ *and* $b_k(x_k)$.

(c) c_k *is upper-bounded-*(β_k, α_k).

(d) $\operatorname{dom} c_k(\cdot\,; x_k) \supset \{\underline{u}_k : A_k\underline{u}_k = b_k(x_k),\ \underline{u}_k \geqq 0\}$ *for a.e.* x_k.

(e) $E|\beta_k(X_k)b_k(X_k)| < +\infty$.

(f) b_k *is continuous, and* $E|b_k(X_k)| < +\infty$.

Also assume that $\operatorname{dom} c_0$ *is a polyhedral convex set. Take* $F_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1}) \equiv F_{X_k}(x_k)$, *the distribution function of* X_k.

Then for each $0 \leqq k \leqq K$ *there are a matrix* D *and vector* d *such that*

$$\operatorname{dom} r_k(\cdot\,; \underline{x}_k) = \{\underline{u}_k : A_k\underline{u}_k = b_k(\underline{x}_k),\ D\underline{u}_k \geqq d,\ \underline{u}_k \geqq 0\}$$

for a.e. $\underline{x}_k$. *In particular,* $\operatorname{dom} r_0$ *is a polyhedral convex set.*

Proof. Assumption (b) and the choice of the regular conditional distribution function imply that at any stage r_k and p_k can be written as functions $r_k(\underline{u}_k\,; x_k)$ and $p_k(\underline{u}_k\,; x_k)$, which do not depend on $\underline{x}_{k-1}$. Thus,

$$\bar{p}_k(\underline{u}_{k-1}) \equiv \int p_k(\underline{u}_{k-1}\,; x_k)\, dF_{X_k}(x_k).$$

For each $1 \leqq k \leqq K$, redefine c_k on the exceptional set of measure 0 in (d) to be identically 0. This redefinition can alter $\bar{p}_k$, for any $1 \leqq k \leqq K$, only on a set of measure 0, which is to say, not at all. After the redefinition, the inclusion in (d) holds for every x_k.

Let $1 \leqq k \leqq K$. Assume:

(i) $\bar{p}_{k+1}$ is upper-bounded-$(\hat{\beta}, \hat{\alpha})$ ($\hat{\beta}$ and $\hat{\alpha}$ are constants);

(ii) $\operatorname{dom} \bar{p}_{k+1}$ is a polyhedral convex set.

Then by (d), there are a matrix D and a vector d such that

$$\operatorname{dom} r_k(\cdot\,; x_k) = \{\underline{u}_k : A_k\underline{u}_k = b_k(x_k),\ D\underline{u}_k \geqq d,\ \underline{u}_k \geqq 0\}$$

for every x_k. By Proposition 4.7(ii), $\bar{p}_k$ is upper-bounded-$(\bar{\delta}, \bar{\gamma})$ for some $(\bar{\delta}, \bar{\gamma})$. By Proposition 4.8, $\operatorname{dom} \bar{p}_k = K^p$, which is the set of every $\underline{u}_{k-1}$ such that $p_k(\underline{u}_{k-1}\,; x_k) < +\infty$ for every x_k in the support of F_{X_k}, and by Proposition 4.10, K^p is a polyhedral convex set. That completes the induction step. The induction hypothesis is trivially true if $k = K$. □

COROLLARY 5.10. *Assume the hypothesis of Theorem* 5.9. *Also assume that for each* $0 \leqq k \leqq K$, $c_k(\cdot\,; x_k)$ *is Lipschitzian-*$\delta_k(x_k)$ *for a.e.* x_k, *with* $\delta_k \in L_1(R^{s_k})$.

Then P_0's feasible region is a polyhedral convex set, and on it f_0 is either finite and Lipschitzian or identically $-\infty$.

Proof. The proof of Theorem 5.9 shows that for each $0 \leqq k \leqq K$

$$\operatorname{dom} r_k(\cdot\,; x_k) = \{\underline{u}_k : A_k \underline{u}_k = b_k(x_k),\ D_k \underline{u}_k \geqq d_k,\ \underline{u}_k \geqq 0\}$$

for a.e. x_k. Inductive application of Propositions 4.3 and 4.5 reveals that $\bar{p}_1$ is Lipschitzian-γ for some γ. The polyhedrality assertion follows directly from the conclusion of Theorem 5.9. □

DEFINITION 5.11. A stochastic programming problem has *complete (almost complete) recourse at stage $k+1$ relative to stage k* if and only if, for every (almost every) $\underline{x}_k$,

$$A_k \underline{u}_k = b_k(\underline{x}_k), \quad \underline{u}_k \geqq 0 \quad \Rightarrow \quad \bar{p}_{k+1}(\underline{u}_k\,; \underline{x}_k) < +\infty.$$

A stochastic programming problem has *complete (almost complete) recourse at stage $k+1$ relative to stages* 0 through k (all prior stages) if and only if, for every (almost every) $\underline{x}_k$,

$$A_i \underline{u}_i = b_i(\underline{x}_i) \quad \text{for } i = 0, \cdots, k, \quad \underline{u}_k \geqq 0 \quad \Rightarrow \quad \bar{p}_{k+1}(\underline{u}_k\,; \underline{x}_k) < +\infty.$$

Of course, if $k = 0$, almost-complete and complete recourse are the same.

In the two-stage problem, Wets uses the term "relatively complete course" for what Definition 5.11 would call "complete recourse at stage 1 relative to stage 0" (Wets (1974)). Relatively complete recourse and related notions have often been assumed in stochastic programming problems (Dantzig (1963, p. 510), Madansky (1960), Charnes et al. (1965), Williams (1965), Wets (1966a)). When a problem has relatively complete recourse, many difficulties vanish, as the next theorem shows.

THEOREM 5.12. *Assume for each $0 \leqq k \leqq K$:*

(a) *There is almost complete recourse at stage $k+1$ relative to stage k.*

(b) *For a.e. $\underline{x}_k$,*

$$\operatorname{dom} c_k(\cdot\,; \underline{x}_k) \supset \{\underline{u}_k : A_k \underline{u}_k = b_k(\underline{x}_k),\ \underline{u}_k \geqq 0\}.$$

(c) *$c_k(\cdot\,; \underline{x}_k)$ is Lipschitzian-$\beta_k(\underline{x}_k)$ for a.e. $\underline{x}_k$, for some $\beta_k \in L_1(R^{s_k})$.*

Then P_0's feasible region is a polyhedral convex set, and on it f_0 is either finite and Lipschitzian or identically $-\infty$.

Proof. For any $0 \leqq k \leqq K$,

$$\operatorname{dom} r_k(\cdot\,; \underline{x}_k) = \{\underline{u}_k : A_k \underline{u}_k = b_k(\underline{x}_k),\ \underline{u}_k \geqq 0\}$$

for a.e. $\underline{x}_k$. Apply Propositions 4.3 and 4.5 inductively. □

PROPOSITION 5.13. *Assume that* $\inf(P_0) < +\infty$. *Assume for each $1 \leqq k \leqq K$:*

(a) *c_k is upper-bounded-(β_k, α_k).*

(b) *$\operatorname{dom} c_k(\cdot\,; \underline{x}_k) \supset \{\underline{u}_k : A_k \underline{u}_k = b_k(\underline{x}_k),\ \underline{u}_k \geqq 0\}$ for a.e. $\underline{x}_k$.*

(c) *$E|\beta_k(\underline{X}_k) b_i(\underline{X}_i)| < +\infty$ for each $1 \leqq i \leqq k$.*

(d) *If $P\{b_k(\underline{X}_k) \in S\} = 1$ and S is closed, $S = R^{m_k}$.*

(e) *b_k is continuous.*

Then there is almost complete recourse at stage $k+1$ relative to stage k for each $0 \leqq k \leqq K$.

Proof. Let $1 \leqq k \leqq K$. *Assume*:

(i) $\bar{p}_{k+1}$ is upper-bounded-(δ, γ), and $E|\delta(\underline{X}_k)b_i(\underline{X}_i)| < +\infty$ for each $1 \leqq i \leqq k$;

(ii) $\operatorname{dom} \bar{p}_{k+1}(\cdot\,; \underline{x}_k) \supset \{\underline{u}_k : \underline{u}_k \geqq 0\}$ for every $\underline{x}_k$.

Redefine c_k on the exceptional set of measure 0 in (b) to be identically 0. For any $1 \leqq i \leqq k$, the redefinition leaves $\bar{p}_i(\cdot\,; \underline{x}_i)$ unchanged for every $\underline{x}_i$ in some set of measure1. After the redefinition,

$$\operatorname{dom} r_k(\cdot\,; \underline{x}_k) = \{\underline{u}_k : A_k\underline{u}_k = b_k(\underline{x}_k),\ \underline{u}_k \geqq 0\}$$

for every $\underline{x}_k$. By Proposition 4.7(ii), $\bar{p}_k$ is bounded-(ξ, η), where

$$\xi(\underline{x}_{k-1}) \equiv \lambda E[\beta_k(\underline{X}_k) + \delta(\underline{X}_k) | \underline{X}_{k-1} = \underline{x}_{k-1}],$$

λ being a nonnegative number. For $1 \leqq i \leqq k-1$,

$$\begin{aligned} +\infty &> \lambda E|\beta_k(\underline{X}_k)b_i(\underline{X}_i)| + \lambda E|\delta(\underline{X}_k)b_i(\underline{X}_i)| \\ &\geqq \lambda EE[|(\beta_k(\underline{X}_k) + \delta(\underline{X}_k))b_i(\underline{X}_i)| \,|\underline{X}_{k-1}] \\ &\geqq \lambda E\,|\,E[(\beta_k(\underline{X}_k) + \delta(\underline{X}_k))b_i(\underline{X}_i)\,|\,\underline{X}_{k-1}]| \\ &= \lambda E|E[(\beta_k(\underline{X}_k) + \delta(\underline{X}_k))|\underline{X}_{k-1}]b_i(\underline{X}_i)| \\ &= E|\xi(\underline{X}_{k-1})b_i(\underline{X}_i)|. \end{aligned}$$

By Proposition 4.8,

$$\operatorname{dom} \bar{p}_k(\cdot\,; \underline{x}_{k-1}) = \{\underline{u}_{k-1} : p_k(\underline{u}_{k-1}; \underline{x}_k) < +\infty \ \forall\, x_k \in \tilde{\Xi}_k(\underline{x}_{k-1})\}$$

for a.e. $\underline{x}_{k-1}$. ($\tilde{\Xi}_k(\underline{x}_{k-1})$ is the support of $\mu_k(\cdot\,|\underline{x}_{k-1})$.) Since $\inf(P_0) < +\infty$, assumption (d) implies that

$$\operatorname{cone} A_{kk} \triangleq \{y : A_{kk}w = y, \text{ some } w \geqq 0\} = R^{m_k}.$$

Consequently, $p_k(\underline{u}_{k-1}; \underline{x}_k) < +\infty$ whenever $\underline{u}_{k-1} \geqq 0$, and $\operatorname{dom} \bar{p}_k(\cdot\,; \underline{x}_{k-1}) = R_+^{N_{k-1}}$ for a.e. $\underline{x}_{k-1}$. For every $\underline{x}_{k-1}$ in the exceptional set, redefine $\bar{p}_k(\cdot\,; \underline{x}_{k-1})$ to be identically 0.

That comples the induction step. The induction hypothesis is trivially true if $k = K$. □

The preceding proposition's conclusion can be strengthened, as the induction hypothesis reveals. The stronger condition—dom $\bar{p}_{k+1}(\cdot\,; \underline{x}_k) \supset \{\underline{u}_k : \underline{u}_k \geqq 0\}$ for a.e. $\underline{x}_k$—might be termed *almost complete recourse at stage* $k+1$.[10]

Assumption (d) of Proposition 5.13 holds if, for example, b_k is the identity mapping and $(X_1, \cdots, X_K)$ is a multivariate normal random vector with a positive definite covariance matrix.

Complete recourse relative to all prior stages is a weaker condition than complete recourse relative to the preceding stage alone. This section concludes by presenting a simple device that obviates the need for restatement and reproof of Theorem 5.12 in the case of complete recourse relative to all prior stages.

---

[10] Wets says that a two-stage problem ($K = 1$) has "complete recourse" if and only if cone $A_{11} = R^{m_1}$ (Wets, (1974)).

Recall the definitions of the problems $P_0, \cdots, P_K$, which fully specify the stochastic programming problem. Form a new sequence of problems $P'_0, \cdots, P'_K$ as follows:

$$P'_k(z; \underline{x}_k): \underset{\underline{u}_k}{\text{minimize}}\ f'_k(\underline{u}_k; \underline{x}_k)$$
$$\text{subject to} \quad A_i\underline{u}_i = b_i(\underline{x}_i) \quad \forall\, 0 \leqq i \leqq k,$$
$$\underline{u}_{k-1} = z, \qquad \underline{u}_k \geqq 0.$$

The notation is meant to indicate that $\underline{x}_i$ is a subvector of $\underline{x}_k$ and $\underline{u}_i$ is a subvector of $\underline{u}_k$. All that has been done is to adjoin the explicit constraints from previous stages to the explicit constraints at stage k. Then

$$f'_k(\underline{u}_k; \underline{x}_k) = \begin{cases} f_k(\underline{u}_k; \underline{x}_k), & A_i\underline{u}_i = b_i(\underline{x}_i) \quad \text{for } i = 0, \cdots, k, \\ +\infty_0 & \text{otherwise.} \end{cases}$$

P_0 and P'_0 have the same explicit constraints—namely, $A_{00}u_0 = b_0$, $u_0 \geqq 0$—and f'_0 differs from f_0 only outside $\{u_0 : A_{00}u_0 = b_0, u_0 \geqq 0\}$, so that P_0 and P'_0 are effectively the same problem.

But $P_0, \cdots, P_K$ have complete (almost complete) recourse at stage $k+1$ relative to stages $0, \cdots, k$ for each $0 \leqq k \leqq K$ if and only if $P'_0, \cdots, P'_K$ have complete (almost complete) recourse at stage $k+1$ relative to stage k for each $0 \leqq k \leqq K$. Furthermore, several propositions (and the theorems that use them) have polyhedrality assumptions that are weaker for $P'_0, \cdots, P'_K$ than for $P_0, \cdots, P_K$. For example, Proposition 4.3 requires

$$\operatorname{dom} f_k(\cdot\,; \underline{x}_k) \cap \{\underline{u}_k : A_k\underline{u}_k = b_k(\underline{x}_k), \underline{u}_k \geqq 0\}$$
$$= \{\underline{u}_k : A_k\underline{u}_k = b_k(\underline{x}_k), \hat{D}\underline{u}_k \geqq \hat{d}(\underline{x}_k), \underline{u}_k \geqq 0\}.$$

The same assumption for P'_k is weaker:

$$\operatorname{dom} f_k(\cdot\,; \underline{x}_k) \cap \{\underline{u}_k : A_i\underline{u}_i = b_i(\underline{x}_i)\ \forall\, 0 \leqq i \leqq k, \underline{u}_k \geqq 0\}$$
$$= \{\underline{u}_k : A_i\underline{u}_i = b_i(\underline{x}_i)\ \forall\, 0 \leqq i \leqq k, \hat{D}\underline{u}_k \geqq \hat{d}(\underline{x}_k), \underline{u}_k \geqq 0\}.$$

Other examples are Propositions 4.7, 4.8 and 4.10 and every theorem following Theorem 5.4.

Acknowledgment. This paper incorporates numerous improvements suggested by Mark J. Eisner and a referee.

REFERENCES

R. B. ASH (1972), *Real Analysis and Probability*, Academic Press, New York.

C. BERGE (1963), *Topological Spaces*, Macmillan, New York.

A. CHARNES, W. W. COOPER AND G. L. THOMPSON (1965), *Constrained generalized medians and hypermedians as deterministic equivalents of two-stage linear programs under uncertainty*, Management Sci., 12, pp. 83–112.

G. B. DANTZIG (1955), *Linear programming under uncertainty*, Ibid., 1, pp. 197–206; Reprinted in Mathematical Studies in Management Sci., A. F. Veinott, Jr., ed., Macmillan, New York, 1965.

——— (1963), *Linear Programming and Extensions*, Princeton University Press, Princeton, N.J.

M. J. EISNER (1970), *On duality in infinite-player games and sequential chance-constrained programming*, Ph.D. thesis, Cornell Univ., Ithaca, N.Y.

A. MADANSKY (1960), *Inequalities for stochastic linear programming problems*, Management Sci., 6, pp. 197–204.

P. OLSEN (1975a), *When is a multistage stochastic programming problem well-defined?*, this Journal, 14 (1976), pp. 518–527.

——— (1975b), *Discretizations of multistage stochastic programming problems*, Proc. of the Conf. on Stochastic Systems, June 10–14, 1975, Lexington, Kentucky; Mathematical Programming Studies, to appear.

R. T. ROCKAFELLAR (1967), *Duality and stability in extremum problems involving convex functions*, Pacific J. Math., 21, pp. 167–87.

——— (1970), *Convex Analysis*, Princeton University Press, Princeton, N.J.

——— (1975), *Lagrange multipliers for an N-stage model in stochastic convex programming*, Colloque d'Analyse Convexe, St. Pierre-Chartreuse, J. P. Aubin, ed., Springer-Verlag, New York.

R. M. VAN SLYKE AND R. WETS (1968), *A duality theory for abstract mathematical programs with applications to optimal control theory*, J. Math. Anal. Appl., 22, pp. 679–706.

D. W. WALKUP AND R. WETS (1967), *Stochastic programs with recourse*, SIAM J. Appl. Math., 15, pp. 1299–1314.

——— (1969a), *Some practical regularity conditions for nonlinear programs*, this Journal, 7, pp. 430–36.

——— (1969b), *Stochastic programs with recourse II: On the continuity of the objective*, SIAM J. Appl. Math., 17, pp. 98–103.

——— (1970), *Stochastic programs with recourse: Special forms*. Proc. on the Princeton Symp. on Math. Programming, H. W. Kuhn, ed., Princeton University Press, Princeton, N.J., pp. 139–62.

R. WETS (1966a), *Programming under uncertainty: The complete problem*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 4, pp. 316–39.

——— (1966b) *Programming under uncertainty: The equivalent convex problem*, SIAM J. Appl. Math., 14, pp. 89–105.

——— (1966c), *Programming under uncertainty: The solution set*, Ibid., 14, pp. 1143–51.

——— (1972), *Stochastic programs with recourse: A basic theorem for multistage problems*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 21, pp. 201–206.

——— (1974), *Stochastic programs with fixed recourse: The equivalent deterministic program*, SIAM Rev., 16, pp. 309–39.

A. WILLIAMS (1965), *On stochastic linear programming*, SIAM J. Appl. Math., 13, pp. 927–40.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

WHEN IS A MULTISTAGE STOCHASTIC PROGRAMMING PROBLEM WELL-DEFINED?*

PAUL OLSEN†

Abstract. Certain measure-theoretic issues raise the possibility that (a) the optimal value of a multistage stochastic programming problem may be ill-defined and (b) the recursion defining the problem may fail, so the problem itself is not even defined. The first difficulty is illustrated by example. A rigorous definition of multistage stochastic programming with fixed linear recourse is shown to avoid this difficulty. In the context of the new definition, certain measurability, convexity, and lower-semicontinuity assumptions on the objective function preclude the second possibility.

1. A reformulation of the problem. Multistage stochastic programming with recourse corresponds to a situation in which information is revealed in stages and a decision is made at each stage based on the information revealed up to and including that stage and on the decisions already made. Number the stages $0, \cdots, K$ $(K < +\infty)$. Assume that the information revealed at stage k $(0 \leqq k \leqq K)$ is fully represented by the realization of a random vector X_k with values in R^{s_k}, where $X_0, \cdots, X_K$ are defined on the same sample space and have known joint distribution; assume that the decision at stage k is represented by a vector $u_k \in R^{n_k}$.

Let $S_k = \sum_{i=0}^{k} s_i$, let $S = S_K$, and let $\underline{X}_k$ denote the random vector $(X_0, \cdots, X_k)$. Let $F_{\underline{X}_k}$ denote the distribution function of $\underline{X}_k$. For each $1 \leqq k \leqq K$, there is a regular conditional distribution function (r.c.d.f.) for X_k given $\underline{X}_{k-1}$—i.e., a function $F_{X_k|\underline{X}_{k-1}} : R^{s_k} \times R^{S_{k-1}} \to [0, 1]$ such that:

(a) for each $\underline{x}_{k-1} \in R^{S_{k-1}}$, $F_{X_k|\underline{X}_{k-1}}(\cdot\,|\underline{x}_{k-1})$ is a proper distribution function on R^{s_k};

(b) for each $x_k \in R^{s_k}$,

$$F_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1}) = P\{X_k \leqq x_k | \underline{X}_{k-1} = \underline{x}_{k-1}\}$$

for almost every (a.e.) $\underline{x}_{k-1}$.

For convenience, also require that $F_{X_k|\underline{X}_{k-1}}(x_k|\cdot)$ be Borel measurable for each x_k. The existence of such a function follows from the fact that the values of X_k lie in a complete separable metric space—R^{s_k} [1, pp. 263–66].

For $0 \leqq i \leqq K$ and $0 \leqq j \leqq K$, let A_{ij} be a real $m_i \times n_j$ matrix (recall that n_j is the dimensionality of the stage j decision vector, u_j). Require that $A_{ij} = 0$ if $j > i$. Let $N_k = \sum_{j=0}^{k} n_j$, let $N = N_K$, and let $\underline{u}_k = (u_0, \cdots, u_k)$.

For $0 \leqq k \leqq K$, let

$$b_k : R^{S_k} \to R^{m_k}, \qquad c_k : R^{N_k} \times R^{S_k} \to [-\infty, +\infty].$$

The stochastic programming problem is defined recursively. Let $\bar{p}_{K+1}(\underline{u}_K; \underline{x}_K) \equiv 0$. Now let $1 \leqq k \leqq K$ and suppose that $\bar{p}_{k+1} : R^{N_k} \times R^{S_k} \to [-\infty, +\infty]$ has been defined. Let.

$$r_k(\underline{u}_k; \underline{x}_k) \equiv c_k(\underline{u}_k; \underline{x}_k) + \bar{p}_{k+1}(\underline{u}_k; \underline{x}_k) + \psi(\underline{u}_k; \underline{x}_k),$$

* Received by the editors July 2, 1974, and in revised form March 6, 1975.

† Institute for Defense Analyses, Arlington, Virginia 22202.

where

$$\psi(\underline{u}_k; \underline{x}_k) = \begin{cases} 0, & \sum_{j=0}^{k} A_{kj}u_j = b_k(\underline{x}_k),\ \underline{u}_k \geqq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

The convention $+\infty+(-\infty)=(-\infty)+(+\infty)=+\infty$ applies throughout.

Define the parameterized problem

$$P_k(\underline{u}_{k-1}; \underline{x}_k) : \underset{v \in R^{n_k}}{\text{minimize}}\ r_k(\underline{u}_{k-1}, v; \underline{x}_k).$$

Let $p_k(\underline{u}_{k-1}; \underline{x}_k) \equiv \inf (P_k(\underline{u}_{k-1}; \underline{x}_k))$, the problem's optimal value. Let

$$\bar{p}_k(\underline{u}_{k-1}; \underline{x}_{k-1}) \equiv \int p_k(\underline{u}_{k-1}; \underline{x}_k)\, dF_{\underline{X}_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1}).^1 \tag{1}$$

That completes the induction step. So that the integral is defined whenever the integrand is measurable, define (as in [9])

$$\int f\, d\mu = \int f^+\, d\mu - \int f^-\, d\mu$$

for any measure μ and measurable, extended-real-valued function f, with the convention $+\infty-(+\infty)=-\infty+(+\infty)=+\infty$.

For stage 0 define

$$P_0(x_0)\text{: minimize } c_0(u_0; x_0) + \bar{p}_1(u_0; x_0)$$
$$\text{subject to } A_{00}u_0 = b_0(x_0),$$
$$u_0 \geqq 0.$$

The preceding formulation of multistage stochastic programming with (fixed linear) recourse resembles that of Dantzig [3], who introduced the problem in [2]. Wets [10] presents results for a special case of the problem in which X_k is independent of $\underline{X}_{k-1}$ for each $1 \leqq k \leqq K$, but he also formulates the problem in the absence of stagewise independence. The Dantzig and Wets formulations differ from the one given here essentially in replacing the definition (1) above with

$$\bar{p}_k(\underline{u}_{k-1}; \underline{x}_{k-1}) \equiv E[p_k(\underline{u}_{k-1}; \underline{X}_k)|\underline{X}_{k-1} = \underline{x}_{k-1}]. \tag{1'}$$

Section 2 shows that $\bar{p}_k(\underline{u}_{k-1}; \underline{x}_{k-1})$ is indeed a version of the above conditional expectation, but Example 2.3 reveals that using (1') instead of (1)—i.e., defining $\bar{p}_k(\underline{u}_{k-1}; \underline{X}_{k-1})$ to be any version of $E[p_k(\underline{u}_{k-1}; \underline{X}_k)|\underline{X}_{k-1}]$—can result in $E[P_0(X_0)]$ (the optimal value of the stochastic programming problem) being ill-defined. According to Theorem 2.1, this cannot happen with the formulation presented here.

A separate issue is whether the function $p_k(\underline{u}_{k-1}; \cdot)$ has the measurability property that permits the integration in equation (1). (The same property must

[1] A more precise way of writing the integral is

$$\int p_k(\underline{u}_{k-1}; \underline{x}_{k-1}, y)\, dF_{X_k|\underline{X}_{k-1}}(y|\underline{x}_{k-1}).$$

The abbreviated notation will be used often.

hold or the conditional expectation in (1′) is undefined.) Section 3 presents conditions under which the integrand in (1) has the requisite measurability property at each stage; it relies upon Theorem 2.2 (below) and Rockafellar's theory of normal convex integrands [6].

2. Uniqueness of the problem's optimal value. For now, measurability is assumed while attention focuses on the issue of whether P_0 is well-defined. Since the r.c.d.f. in (1) is not necessarily unique (there may be more than one function having properties (a) and (b) above), the definition of $\bar{p}_k$ in terms of an integral over "the" r.c.d.f. for X_k given $\underline{X}_{k-1}$ seems to raise the possibility that $\bar{p}_1$ can be essentially altered simply by using a different r.c.d.f. at each stage. In fact, that is not the case.

Suppose that for $1 \leqq k \leqq K$, $F_{X_k|\underline{X}_{k-1}}$ and $G_{X_k|\underline{X}_{k-1}}$ are both regular conditional distribution functions for X_k given $\underline{X}_{k-1}$. Using $F_{X_k|\underline{X}_{k-1}}$ at stage k for each k generates a sequence of problems $P'_K, \cdots, P'_0$ via the recursion

$$\bar{p}'_k(z; \underline{x}_{k-1}) = \int p'_k(z; \underline{x}_k)\, dF_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1}).$$

The functions $G_{X_k|\underline{X}_{k-1}}$, $1 \leqq k \leqq K$, generate a parallel sequence $P''_K, \cdots, P''_0$. (The data b_k and c_k are the same for P'_k and P''_k.)

THEOREM 2.1. *For each* $1 \leqq k \leqq K$, $\bar{p}'_k(\cdot\,; \underline{x}_k) = \bar{p}''_k(\cdot\,; \underline{x}_k)$ *for almost every* $\underline{x}_k$.[2] *Thus* P'_0 *is essentially the same as* P''_0.

Proof. Let $1 \leqq k \leqq K$. Assume that

$$\bar{p}'_{k+1}(\cdot\,; \underline{x}_k) = \bar{p}''_{k+1}(\cdot\,; \underline{x}_k) \quad \text{for a.e. } \underline{x}_k.$$

Then

$$p'_k(\cdot\,; \underline{x}_k) = p''_k(\cdot\,; \underline{x}_k) \quad \forall \underline{x}_k \in T,$$

where $P\{\underline{X}_k \in T\} = 1$. By the product measure theorem [1, p. 97],

$$\int \chi_T(\underline{x}_{k-1}, x_k)\, dG_{X_k|\underline{X}_{k-1}}(\underline{x}_k|\underline{x}_{k-1}) = 1 \quad \forall \underline{x}_{k-1} \in \mathcal{S},$$

where $P\{\underline{X}_{k-1} \in \mathcal{S}\} = 1$.

Property (b) of the r.c.d.f. implies that for each fixed x_k,

$$F_{X_k|\underline{X}_{k-1}}(x_k|\cdot) = G_{X_k|\underline{X}_{k-1}}(x_k|\cdot) \quad \text{a.e.}$$

Since R^{s_k} is separable and distribution functions are right-continuous, there is a set $\mathcal{S}'$ of measure 1 such that if $\underline{x}_{k-1} \in \mathcal{S}'$,

$$F_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1}) = G_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1}) \quad \forall x_k.$$

If $\underline{x}_{k-1} \in \mathcal{S} \cap \mathcal{S}'$,

$$\int p'_k(z; \underline{x}_k)\, dF_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1}) = \int p'_k(z; \underline{x}_k)\, dG_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1})$$

$$= \int p''_k(z; \underline{x}_k)\, dG_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1})$$

for every $z \in R^{N_{k-1}}$. That completes the induction step. The induction hypothesis is trivially true if $k = K$. □

---

[2] If $f\colon R^n \to R$ and $g\colon R^n \to R$, $f = g$ means $f(x) = g(x) \forall x \in R^n$.

THEOREM 2.2. *Let* g: $R^n \times R^{S_k} \to [-\infty, +\infty]$ *be Borel measurable. Let* $F_{\underline{X}_k|\underline{X}_{k-1}}$ *be any regular conditional distribution function for* $\underline{X}_k$ *given* $\underline{X}_{k-1}$ *such that* $F_{\underline{X}_k|\underline{X}_{k-1}}(x_k|\cdot)$ *is Borel measurable for each fixed* x_k. *Let*

$$h(u, \underline{x}_{k-1}) = \int g(u, \underline{x}_{k-1}, y)\, dF_{\underline{X}_k|\underline{X}_{k-1}}(y|\underline{x}_{k-1})$$

for every $u \in R^n$ *and every* $\underline{x}_{k-1}$. *Then* h *is Borel measurable, and*

$$\int h(u, \underline{x}_{k-1})\, dF_{\underline{X}_{k-1}}(\underline{x}_{k-1}) \leqq \int g(u, \underline{x}_k)\, dF_{\underline{X}_k}(\underline{x}_k) \tag{2}$$

for each $u \in R^n$. *If, for a given* u,

$$\int g^+(u, \underline{x}_k)\, dF_{\underline{X}_k}(\underline{x}_k) < +\infty \text{ or } \int g^-(u, \underline{x}_k)\, dF_{\underline{X}_k}(\underline{x}_k) < +\infty,$$

then (2) *holds as an equality, and*

$$h(u, \underline{x}_{k-1}) = E[g(u, \underline{X}_k)|\underline{X}_{k-1} = \underline{x}_{k-1}]$$

for a.e. $\underline{x}_{k-1}$.

The theorem extends Fubini's theorem [1, p. 101] to agree with the extended definition of integration given after equation (1). A proof appears in [4].

The first part of the theorem is used below to demonstrate measurability of $\bar{p}_k$ (whose role is played by h in the theorem) while the second part justifies the usual economic interpretation of the stochastic programming problem. The second part implies that if $E\bar{p}_{k+1}^-(\underline{u}_k\,; \underline{X}_{k+1}) < +\infty$—for which the Appendix's Proposition A.1 gives a verifiable sufficient condition—then $E[p_{k+1}(\underline{u}_k\,; \underline{X}_{k+1})|\underline{X}_k = \underline{x}_k]$ exists and equals $\bar{p}_{k+1}(\underline{u}_k\,; \underline{x}_k)$ for almost every $\underline{x}_k$. Thus, $\bar{p}_{k+1}(\underline{u}_k\,; \underline{x}_k)$ is the expected (minimum) cost of operations in stages $k+1$ through K given past decisions $u_0, \cdots, u_k$ and states of nature $x_0, \cdots, x_k$. The problem $P_k(\underline{u}_{k-1}\,; \underline{x}_k)$ involves trying to choose u_k to minimize the sum of current costs, $c_k(\underline{u}_k\,; \underline{x}_k)$, and expected future costs, $\bar{p}_{k+1}(\underline{u}_k\,; \underline{x}_k)$. To be precise, at stage k the decision-maker observes x_k; knowing $\underline{x}_k$ and his past decisions, $\underline{u}_{k-1}$, he seeks a stage k decision u_k that minimizes current costs plus expected future costs subject to the constraints

$$A_{kk}u_k = b_k(\underline{x}_k) - \sum_{j=0}^{k-1} A_{kj}u_j$$

$$u_k \geqq 0.$$

Theorem 2.2 implies $\bar{p}_k(\underline{u}_{k-1}\,; \underline{X}_{k-1})$ is a version of $E[p_k(\underline{u}_{k-1}\,; \underline{X}_k)|\underline{X}_{k-1}]$. It does not follow that one could simply take $\bar{p}_k(\underline{u}_{k-1}\,; \cdot)$ to be any version of the conditional expectation and still conclude (as in Theorem 2.1) that P_0 is well-defined. The example below illustrates how choosing a different version of the conditional expectation could make P_0 essentially different.

Example 2.3. Let $K = 2$, let $X_0 = 0$ almost surely, and let X_1 and X_2 be uniformly distributed on [0, 1]. Let

$$c_2(\underline{u}_2\,; \underline{x}_2) \equiv 0, \quad c_1(\underline{u}_1\,; \underline{x}_1) \equiv x_1u_1, \quad c_0(u_0\,; 0) \equiv 0.$$

Let the constraints be

$$\begin{aligned} u_0 &\leqq 1, \\ u_1 &\leqq 1, \\ u_2 &\leqq 1, \\ u_0, u_1, u_2 &\geqq 0. \end{aligned}$$

Clearly, $\bar{p}_2$ is everywhere 0 or $+\infty$. Redefine it for each $\underline{u}_1$ on a set of measure 0:

$$\bar{p}_2'(\underline{u}_1; \underline{x}_1) = \begin{cases} 0, & \underline{u}_1 \geqq 0, x_1 \neq u_1, \\ -1, & \underline{u}_1 \geqq 0, x_1 = u_1, \\ +\infty, & \text{otherwise.} \end{cases}$$

For any fixed $\underline{u}_1$, $\bar{p}_2'(\underline{u}_1; \underline{x}_1) = \bar{p}_2(\underline{u}_1; \underline{x}_1)$ for a.e. $\underline{x}_1$, so $\bar{p}_2(\underline{u}_1; \underline{X}_1)$ and $\bar{p}_2'(\underline{u}_1; \underline{X}_1)$ are two different versions of $E[p_2(\underline{u}_1; \underline{X}_2)|\underline{X}_1]$. Notice, however, that there is no regular conditional distribution function $F'_{X_2|\underline{X}_1}$ such that

$$\bar{p}_2'(\underline{u}_1; \underline{x}_1) = \int p_2(\underline{u}_1; \underline{x}_2)\, dF'_{X_2|\underline{X}_1}(x_2|\underline{x}_1)$$

for every $\underline{u}_1$ and $\underline{x}_1$.

Use of $\bar{p}_2'$ in place of $\bar{p}_2$ in the stage 1 objective function results in the sequence of problems

$$\begin{aligned} P_1'(u_0; \underline{x}_1)\text{: minimize } & x_1 u_1 - \chi_{\{x_1\}}(u_1) \\ \text{subject to } & u_1 \leqq 1 \\ & u_0, u_1 \geqq 0 \\ P_0'(x_0)\text{: minimize } & -\tfrac{2}{3} \\ \text{subject to } & 0 \leqq u_0 \leqq 1. \end{aligned}$$

Thus $\inf(P_0'(0)) = -\frac{2}{3} < 0 = \inf(P_0(0))$.

Although $\bar{p}_2'(\underline{u}_1; \cdot)$ was created by redefining $\bar{p}_2(\underline{u}_1; \cdot)$ on a set of measure 0 for each fixed $\underline{u}_1$, the set of every $\underline{x}_1$ such that $\bar{p}_2'(\underline{u}_1; \underline{x}_1) \neq \bar{p}_2(\underline{u}_1; \underline{x}_1)$ for *some* $\underline{u}_1$ has positive measure; that is the source of the discrepancy. In general, $p_k(\cdot\,; \underline{x}_k)$ depends (at least potentially) on the whole of the function $\bar{p}_{k+1}(\cdot\,; \underline{x}_k)$. Altering $\bar{p}_{k+1}(\cdot\,; \underline{x}_k)$ on a set of measure 0 alters $p_k(\cdot\,; \underline{x}_k)$ on a set of measure 0 and therefore leaves $\bar{p}_1$ essentially unchanged. Altering the function on a set of positive measure can essentially alter P_0 and destroy the original problem.

3. Measurability of each stage's objective function. Let $\mu_k(\cdot\,|\underline{x}_{k-1})$ denote the Lebesgue–Stieltjes measure determined by $F_{X_k|\underline{X}_{k-1}}(\cdot\,|\underline{x}_{k-1})$. The recursive definition of the stochastic programming problem tacitly assumes that at each stage, for a.e. $\underline{x}_{k-1}$, $p_k(\underline{u}_{k-1}; \underline{x}_{k-1}, \cdot)$ is $\mathcal{T}$-measurable for every $\underline{u}_{k-1}$, where $\mathcal{T}$ is the completion of the Borel sets in R^{s_k} with respect to $\mu_k(\cdot\,|\underline{x}_{k-1})$; if that is not the case, the function $\bar{p}_k(\cdot\,; \underline{x}_{k-1})$ is undefined on a set of $\underline{x}_{k-1}$'s of positive measure, and the recursion cannot continue.

DEFINITION 3.1. $\bar{p}_{K+1}$ is said to be essentially defined. Let $1 \leqq k \leqq K$. One says that $\bar{p}_k$ is *essentially defined* if and only if: (i) $\bar{p}_{k+1}$ is essentially defined; and

(ii) for a.e. $\underline{x}_{k-1}$, $p_k(\underline{u}_{k-1}; \underline{x}_{k-1}, \cdot)$ is measurable with respect to the completion of the σ-algebra of Borel sets in R^{s_k} under the measure $\mu_k(\cdot \mid \underline{x}_{k-1})$ for each $\underline{u}_{k-1}$.

PROPOSITION 3.2. *Let* $1 \leqq k \leqq K$. *Assume*:

(a) $\bar{p}_{k+1}$ *is essentially defined, and there is a Borel measurable function* $\bar{p}^0_{k+1}$ *such that* $\bar{p}_{k+1}(\cdot\,; \underline{x}_k) = \bar{p}^0_{k+1}(\cdot\,; \underline{x}_k)$ *for a.e.* $\underline{x}_k$.

(b) $\bar{p}^0_{k+1}(\cdot\,; \underline{x}_k)$ *is convex for a.e.* $\underline{x}_k$.

(c) c_k *and* b_k *are Borel measurable.*

(d) $c_k(\cdot\,; \underline{x}_k)$ *is convex for a.e.* $\underline{x}_k$.

(e) *there is a countable collection* V *of Borel measurable functions* $v : R^{N_{k-1}} \times R^{S_k} \to R^{n_k}$ *such that for a.e.* $\underline{x}_k$, $V(\underline{u}_{k-1}, \underline{x}_k) \cap \operatorname{dom} r_k(\underline{u}_{k-1}, \cdot\,; \underline{x}_k)$ *is dense in* $\operatorname{dom} r_k(\underline{u}_{k-1}, \cdot\,; \underline{x}_k)$ *for every* $\underline{u}_{k-1}$, *where*

$$V(\underline{u}_{k-1}, \underline{x}_k) = \{v(\underline{u}_{k-1}, \underline{x}_k);\ v \in V\}.^3$$

Then $\bar{p}_k$ *is essentially defined, and there is a Borel measurable function* $\bar{p}^0_k$ *such that*: (i) $\bar{p}^0_k(\cdot\,; \underline{x}_{k-1}) = \bar{p}_k(\cdot\,; \underline{x}_{k-1})$ *for a.e.* $\underline{x}_{k-1}$; *and* (ii) $\bar{p}^0_k(\cdot\,; \underline{x}_{k-1})$ *is convex for a.e.* $\underline{x}_{k-1}$.

Proof. Let S be a Borel set of measure 1 such that $\underline{x}_k \in S$ implies:

(i) $c_k(\cdot\,; \underline{x}_k)$ and $\bar{p}^0_{k+1}(\cdot\,; \underline{x}_k)$ are convex;

(ii) $V(\underline{u}_{k-1}, \underline{x}_k) \cap \operatorname{dom} r_k(\underline{u}_{k-1}, \cdot\,; \underline{x}_k)$ is dense in $\operatorname{dom} r_k(\underline{u}_{k-1}, \cdot\,; \underline{x}_k)$ for every $\underline{u}_{k-1}$;

(iii) $\bar{p}_{k+1}(\cdot\,; \underline{x}_k) = \bar{p}^0_{k+1}(\cdot\,; \underline{x}_k)$.

Redefine $\bar{p}^0_{k+1}(\cdot\,; \underline{x}_k)$ outside S to be identically $+\infty$. Let P^0_k be the (parameterized) problem formed from P_k by using $\bar{p}^0_{k+1}$ in place of $\bar{p}_{k+1}$. Denote its return function by r^0_k and its perturbation function by p^0_k. By [1, 1.5.6], r^0_k is Borel measurable.[4]

The first step in the proof is to show that

$$p^0_k(z; \underline{x}_k) \equiv \inf\{r^0_k(z, v(z, \underline{x}_k); \underline{x}_k) : v \in V\} \quad . \tag{3}$$

Let $z \in R^{N_{k-1}}$, $\underline{x}_k \in R^{S_k}$.

If $p^0_k(z; \underline{x}_k) = +\infty$, $r^0_k(z, u_k; \underline{x}_k) = +\infty$ for every $u_k \in R^{n_k}$, and (3) holds trivially.

Alternatively, suppose $p^0_k(z; \underline{x}_k) < +\infty$.

Choose a sequence $\{u^n_k\}$ such that

$$r^0_k(z, u^n_k; \underline{x}_k) < +\infty \quad \text{for every } n$$

and

$$r^0_k(z, u^n_k; \underline{x}_k) \searrow p^0_k(z; \underline{x}_k).$$

Choose $\{\alpha_n\} \subset R$ such that $r^0_k(z, u^n_k; \underline{x}_k) < \alpha_n$ for every n and $\alpha_n \searrow p^0_k(z; \underline{x}_k)$.

For any given n, suppose that u^n_k lies in the relative interior of $\operatorname{dom} r^0_k(z, \cdot\,; \underline{x}_k) \triangleq \{u_k : r^0_k(z, u_k; x_k) < +\infty\}$.[5] Then by [7, Thm. 10.1] and property (ii), there is a function $v \in V$ such that

$$r^0_k(z, v(z, \underline{x}_k); \underline{x}_k) < \alpha_n + (1/n).$$

---

[3] Cf. the definition of "normal convex integrand" in [6].

[4] Note that Ash does not use the term "Borel measurable" in the same sense as this paper, where it means "measurable with respect to the σ-algebra of Borel sets." Also, the theorem in [1] must be extended slightly since $r_k(\underline{u}_k; \underline{x}_k)$ may involve a sum of the form $+\infty + (-\infty)$, which Ash does not define.

[5] See [7] for the definitions of "relative interior" and "relative boundary".

Alternatively, if u_k^n lies in the relative boundary of dom $r_k^0(z, \cdot\,; \underline{x}_k)$, Corollary 7.3.1 of [7] reveals the existence of a point w in the relative interior of dom $r_k^0(z, \cdot\,; \underline{x}_k)$ such that $r_k^0(z, w; \underline{x}_k) < \alpha_n$, and the preceding argument can be applied to w.

Thus, in any case, there is a sequence $\{v^n\} \subset V$ such that

$$r_k^0(z, v^n(z, \underline{x}_k); \underline{x}_k) \to p_k^0(z; \underline{x}_k).$$

That demonstrates (3).

Since r_k^0 and v are Borel measurable, $r_k^0(z, v(z, \underline{x}_k); \underline{x}_k)$ is a Borel measurable function of $(z, \underline{x}_k)$ for each $v \in V$. As the infimum of a countable collection of Borel measurable functions, p_k^0 is Borel measurable. Then by Theorem 2.2,

$$\bar{p}_k^0(\underline{u}_{k-1}; \underline{x}_{k-1}) \triangleq \int p_k^0(\underline{u}_{k-1}; \underline{x}_k)\, dF_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1})$$

is defined for each $\underline{u}_{k-1}$ and $\underline{x}_{k-1}$, and $\bar{p}_k^0$ is Borel measurable. By Proposition A.2 (Appendix), $\bar{p}_k^0(\cdot\,; \underline{x}_{k-1})$ is convex for every $\underline{x}_{k-1}$. Since $p_k(\cdot\,; \underline{x}_k) = p_k^0(\cdot\,; \underline{x}_k)$ for a.e. $\underline{x}_k$, Theorem 2.2 reveals that for a.e. $\underline{x}_{k-1}$, there is a Borel set B in R^{S_k} such that $\mu_k(B|\underline{x}_{k-1}) = 1$ and $p_k(\cdot\,; \underline{x}_{k-1}, y) = p_k^0(\cdot\,; \underline{x}_{k-1}, y)$ if $y \in B$. Consequently, $\bar{p}_k$ is essentially defined. □

The preceding proposition's assumption (e) is appealing but not directly verifiable. The next proposition gives a sufficient condition for it. The hypothesis of Proposition 3.3 is the conclusion of Theorems 5.5, 5.9, and 5.12 of [5]. Consequently, one can merge the hypothesis of any of those theorems with assumptions (c) and (d) of Proposition 3.2 and prove inductively that $\bar{p}_1$ is essentially defined.

PROPOSITION 3.3. *Let $1 \leqq k \leqq K$. Suppose there are matrices D_1 and D_2 and a Borel measurable function d: $R^{S_k} \to R^m$ such that*

$$\text{dom } r_k(\cdot\,; \underline{x}_k) = \{\underline{u}_k : D_1 \underline{u}_{k-1} + D_2 u_k \geqq d(\underline{x}_k),\ \underline{u}_k \geqq 0\}$$

for a.e. $\underline{x}_k$. Then assumption (e) *of Proposition* 3.2 *holds.*[6]

Proof. Let $W = [D_2 \mid -I]$. There are square matrices $M_1, \cdots, M_m$ such that, for any given $a \in R^m$, w is an extreme point of $\{w : Ww = a,\ w \geqq 0\}$ if and only if $w \geqq 0$ and $w = M_i a$ for some i. Let $a_1, \cdots, a_s$ be the extreme directions of $\{w : Ww = 0,\ w \geqq 0\}$. Now redefine $M_1, \cdots, M_m$ by deleting all but the first n_k rows of each matrix, and redefine $a_1, \cdots, a_s$ by deleting all but the first n_k components of each vector. For any $z \in R^{N_{k-1}}$ and any $\underline{x}_k$,

$$\text{dom } r_k(z, \cdot\,; \underline{x}_k) = \{y : y = \sum_{i=1}^{m} \lambda_i M_i(d(\underline{x}_k) - D_1 z) + \sum_{i=1}^{s} \gamma_i a_i,$$

$$\sum_{i=1}^{m} \lambda_i = 1,\ \lambda_i \geqq 0 \forall i,\ \gamma_i \geqq 0 \forall i,\ y \geqq 0\}.$$

Let V be the collection of functions v such that

$$v(z, \underline{x}_k) \equiv \sum_{i=1}^{m} \lambda_i M_i(d(\underline{x}_k) - D_1 z) + \sum_{i=1}^{s} \lambda_i a_i,$$

where $\sum_{i=1}^{m} \lambda_i = 1$, $\lambda_i \geqq 0 \forall i$, $\gamma_i \geqq 0 \forall i$, and λ_i and γ_i are rational numbers. □

[6] Actually, the linear mappings D_1 and D_2 could be replaced by a Borel measurable mapping and a continuous mapping, respectively. See the proof of Proposition 4.6 in [4].

THEOREM 3.4 *Assume for each* $1 \leqq k \leqq K$:

(a) c_k *and* b_k *are Borel measurable*;

(b) $c_k(\cdot\,; \underline{x}_k)$ *is l.s.c. and convex for a.e.* $\underline{x}_k$;

(c) *There are functions* β_k *and* α_k *in* $L_1(R^{S_k})$ *with values in* $[-\infty, 0]$ *such that*

$$c_k(\underline{u}_k; \underline{x}_k) \geqq \beta_k(\underline{x}_k)\|\underline{u}_k\|_2 + \alpha_k(\underline{x}_k)$$

for every u_k *and every* x_k, *and*

$$E|\beta_k(\underline{X}_k)b_i(\underline{X}_i)| < +\infty \quad \text{if } 1 \leqq i \leqq k;^7$$

(d) $\{w: A_{kk}w = 0,\ w \geqq 0\} = \{0\}$.

Then $\bar{p}_1$ *is essentially defined, and for each* $1 \leqq k \leqq K$, *there is a Borel measurable function* $\bar{p}_k^0$ *such that* $\bar{p}_k(\cdot\,; \underline{x}_{k-1}) = \bar{p}_k^0(\cdot\,; \underline{x}_{k-1})$ *for a.e.* $\underline{x}_{k-1}$ *and* $\bar{p}_k^0(\cdot\,; \underline{x}_{k-1})$ *is l.s.c. and convex for a.e.* $\underline{x}_{k-1}$.

The theorem's proof uses the following lemma, which extends [6, Thm. 5].

LEMMA 3.5. *Let* $f: R^n \times R^m \to [-\infty, +\infty]$ *be Borel measurable. Let* μ *be a Borel probability measure on* R^m. *Assume that* $f(\cdot\,; x)$ *is l.s.c. and convex for a.e.* $x \in R^m$. *Then there is a countable collection* U *of Borel measurable functions* u: $R^m \to R^n$ *such that* $U(x) \cap \operatorname{dom} f(\cdot\,; x)$ *is dense in* $\operatorname{dom} f(\cdot\,; x)$ *for a.e.* x, *where*

$$U(x) = \{u(x): u \in U\}.$$

Proof. Redefine $f(\cdot\,; x)$ to be identically $+\infty$ on the set of measure 0 where it is not lower-semicontinuous and convex.

Define a multifunction $K_1 : R^m \to R^n$ by

$$K_1(x) \equiv \{\underline{u} \in R^n: f(\cdot\,; x) \in R\}.$$

Since f is Borel measurable, the graph of K_1 is a Borel set, and therefore by [6, Thm. 2], K_1 is measurable with respect to $\mathcal{B}(R^m)^*$, the completion of the σ-algebra of Borel sets under the measure μ. Consequently,

$$\begin{aligned} S_1 &\triangleq \{x: f(u; x) \in R \text{ for some } u\} \\ &= \{x: K_1(x) \neq \varnothing\} \\ &= K_1^{-1}(R^n) \in \mathcal{B}(R^m)^*. \end{aligned}$$

Define a multifunction K_2 by

$$K_2(x) \equiv \{u \in R^n: f(u; x) = -\infty\}.$$

An argument parallel to the preceding one reveals that

$$\begin{aligned} S_2 &\triangleq \{x: f(u; x) = -\infty \text{ for some } u\} \\ &= \{x: K_2(x) \neq \varnothing\} \in \mathcal{B}(R^m)^*. \end{aligned}$$

Let

$$g(u; x) = \begin{cases} 0, & x \in S_2 \text{ and } f(u; x) < +\infty, \\ f(u; x) & \text{otherwise.} \end{cases}$$

---

[7] The notation signifies that $\underline{X}_i$ is a subvector of $\underline{X}_k$.

Since $(R^n \times S_2) \cap \{(u, x): f(u; x) < +\infty\} \in \mathcal{B}(R^m)^*$, g is $\mathcal{B}(R^m)^*$-measurable. If $x \in S_2$, $f(\cdot\,; x)$ is identically $-\infty$ on its effective domain; hence, for each $x \in R^m$, $g(\cdot\,; x)$ is a lower-semicontinuous convex function with the same effective domain as $f(\cdot\,; x)$.

Let $T = S_1 \cup S_2$. By [6, Thm. 5], there is a countable collection U of $\mathcal{B}(R^m)^*$-measurable functions $u: T \to R^n$ such that $U(x) \cap \operatorname{dom} g(\cdot\,; x)$ is dense in $\operatorname{dom} g(\cdot\,; x)$ for every $x \in T$. Extend each function in U to a function on R^m by making it identically 0 outside T. If $u \in U$, there is a Borel measurable function $\hat{u}: R^m \to R^n$ such that $u = \hat{u}$ a.e. (μ) [8, p. 145]; let $\hat{U}$ be the collection consisting of one such $\hat{u}$ for each $u \in U$. Then $\hat{U}(x) = U(x)$ for a.e. $x \in R^m$ since U has countably many elements. Hence, there is a Borel subset T^0 of T such that $\mu(T^0) = \mu(T)$ and $\hat{U}(x) \cap \operatorname{dom} g(\cdot\,; x)$ is dense in $\operatorname{dom} g(\cdot\,; x)$ for each $x \in T^0$. Since $\operatorname{dom} f(\cdot\,; x) = \operatorname{dom} g(\cdot\,; x)$ for any $x \in R^m$ and $\operatorname{dom} f(\cdot\,; x) = \varnothing$ if $x \notin T$, the collection $\hat{U}$ has the desired property. □

Proof of Theorem 3.4. Let $1 \leqq k \leqq K$. Assume that $\bar{p}_{k+1}$ is essentially defined and there is a Borel measurable function $\bar{p}^0_{k+1}$ such that $\bar{p}_{k+1}(\cdot\,; \underline{x}_k) = \bar{p}^0_{k+1}(\cdot\,; \underline{x}_k)$ for a.e. $\underline{x}_k$ and $\bar{p}^0_{k+1}(\cdot\,; \underline{x}_k)$ is l.s.c. and convex for a.e. $\underline{x}_k$. Also assume that

$$\bar{p}^0_{k+1}(\underline{u}_k; \underline{x}_k) \geqq \delta(\underline{x}_k)\|\underline{u}_k\|_2 + \gamma(\underline{x}_k)$$

for every $\underline{u}_k$ and every $\underline{x}_k$, where δ and γ have the nonpositivity, mean, and covariance properties of β_k and α_k, respectively, in (c).

Let P^0_k be the problem formed from P_k by replacing $\bar{p}_{k+1}$ with $\bar{p}^0_{k+1}$. Proposition A.1 (Appendix) implies that $p^0_k(\cdot\,; \underline{x}_k) > -\infty$ for a.e. $\underline{x}_k$. Then by Proposition A.3, $p^0_k(\cdot\,; \underline{x}_k)$ is l.s.c. and convex for a.e. $\underline{x}_k$. Since $\bar{p}^0_{k+1}(\cdot\,; \underline{x}_k) > -\infty$ and $c_k(\cdot\,; \underline{x}_k) > -\infty$ for a.e. $\underline{x}_k$, $r^0_k(\cdot\,; \underline{x}_k)$ is l.s.c. and convex for a.e. $\underline{x}_k$ by [7, Thm. 9.3]. As a finite sum of Borel measurable functions, it is Borel measurable. Lemma 3.5 reveals the existence of a countable collection U of Borel measurable functions $u: R^{S_k} \to R^{N_k}$ such that $U(\underline{x}_k) \cap \operatorname{dom} r^0_k(\cdot\,; \underline{x}_k)$ is dense in $\operatorname{dom} r^0_k(\cdot\,; \underline{x}_k)$ for every $\underline{x}_k \in S$, where S is a Borel set of measure 1. Let T be a Borel subset of S of measure 1 such that $\underline{x}_k \in T$ implies $r^0_k(\cdot\,; \underline{x}_k)$ and $p^0_k(\cdot\,; \underline{x}_k)$ are l.s.c. and convex. Redefine r^0_k and p^0_k to be identically $+\infty$ outside T. Then for any $z \in R^{N_{k-1}}$ and $\underline{x}_k \in R^{S_k}$,

$$\begin{aligned} p^0_k(z; \underline{x}_k) &= \liminf_{z' \to z} \left(\inf_{u_k} r^0_k(z, u_k; \underline{x}_k)\right) \\ &= \liminf_{n \to \infty} \inf_{\underline{u}_k} \{r^0_k(\underline{u}_k; \underline{x}_k) : \|\underline{u}_{k-1} - z\|_2 < 1/n\} \\ &= \liminf_{n \to \infty} \{r^0_k(\underline{u}_k(\underline{x}_k); \underline{x}_k) : \|\underline{u}_{k-1}(\underline{x}_k) - z\|_2 < 1/n, \underline{u}_k \in U\}. \end{aligned}$$

It follows that p^0_k is Borel measurable. An argument identical to one in the proof of Proposition 3.2 shows that $\bar{p}_k$ is essentially defined and $\bar{p}_k(\cdot\,; \underline{x}_{k-1}) = \bar{p}^0_k(\cdot\,; \underline{x}_{k-1})$ for a.e. $\underline{x}_{k-1}$.

By Proposition A.1, there functions $\hat{\delta}$ and $\hat{\gamma}$ in $L_1(R^{S_{k-1}})$ with values in $[-\infty, 0]$ such that

$$\bar{p}_k^0(\underline{u}_{k-1}; \underline{x}_{k-1}) \geqq \hat{\delta}(\underline{x}_{k-1})\|\underline{u}_{k-1}\|_2 + \hat{\gamma}(\underline{x}_{k-1})$$

for every $\underline{u}_{k-1}$ and every $\underline{x}_{k-1}$; also, $E|\hat{\delta}((\underline{X}_{k-1})b_i(\underline{X}_i)| < +\infty$ if $1 \leqq i \leqq k$[8]. Then by Proposition A.4, $\bar{p}_k^0(\cdot\,; \underline{x}_{k-1})$ is l.s.c. and convex for a.e. $\underline{x}_{k-1}$.

That completes the induction step. The induction hypothesis is obviously true if $k = K$. □

Appendix. These propositions are restatements of some propositions in [5], whose numbers are displayed in parentheses.

PROPOSITION A.1 (2.2). *Assume for each* $0 \leqq k \leqq K$,

(a) *For some selected functions* β_k *and* α_k *in* $L_1(R^{S_k})$ *with values in* $[-\infty, 0]$,

$$c_k(\underline{u}_k; \underline{x}_k) \geqq \beta_k(\underline{x}_k)\|\underline{u}_k\|_2 + \alpha_k(\underline{x}_k)$$

for every $\underline{u}_k$ *and every* $\underline{x}_k$.

(b) $E|\beta_k(\underline{X}_k)b_i(\underline{X}_i)| < +\infty$ *if* $0 \leqq i \leqq k$. ($\underline{X}_i$ *is a subvector of* $\underline{X}_k$.)

(c) $\{w: A_{kk}w = 0,\ w \geqq 0\} = \{0\}$.

Then, for each $0 \leqq k \leqq K$, $E\bar{p}_k(\underline{u}_{k-1}; \underline{X}_k) < +\infty$, *and there are functions* δ_k *and* γ_k *in* $L_1(R^{S_k})$, *with values in* $[-\infty, 0]$, *such that*

$$\bar{p}_{k+1}(\underline{u}_k; \underline{x}_k) \geqq \delta_k(\underline{x}_k)\|\underline{u}_k\|_2 + \gamma_k(\underline{x}_k)$$

for every $\underline{u}_k$ *and every* $\underline{x}_k$.

PROPOSITION A.2 (3.2, 3.3). *Let* $1 \leqq k \leqq K$. *If* $c_k(\cdot\,; \underline{x}_k)$ *and* $\bar{p}_{k+1}(\cdot\,; \underline{x}_k)$ *are convex for every* $\underline{x}_k$, $\bar{p}_k(\cdot\,; \underline{x}_{k-1})$ *is convex for every* $\underline{x}_{k-1}$.

PROPOSITION A.3 (3.4). *Let* $1 \leqq k \leqq K$. *Assume that* $\{w : A_{kk}w = 0,\ w \geqq 0\} = \{0\}$. *Fix* $\underline{x}_k$. *If* $c_k(\cdot\,; \underline{x}_k)$ *and* $\bar{p}_{k+1}(\cdot\,\underline{x}_k)$ *are l.s.c. convex functions and* $p_k(\cdot\,; \underline{x}_k) > -\infty$, *then* $\bar{p}_k(\cdot\,; \underline{x}_k)$ *is l.s.c. and convex.*

PROPOSITION A.4 (3.5). *Let* $1 \leqq k \leqq K$. *Fix* $\underline{x}_{k-1}$. *If* $p_k(\cdot\,; \underline{x}_{k-1}, \hat{x}_k)$ *is l.s.c. and convex for every* $\hat{x}_k$ *and if* $\bar{p}_k(\cdot\,; \underline{x}_{k-1}) > -\infty$, *then* $\bar{p}_k(\cdot\,; \underline{x}_{k-1})$ *is l.s.c. and convex.*

REFERENCES

[1] R. B. ASH (1972), *Real Analysis and Probability*, Academic Press. New York.

[2] G. B. DANTZIG (1955), *Linear programming under uncertainty*, Management Sci., 1, pp. 197–206.

[3] ——— (1963), *Linear Programming and Extensions*, Princeton University Press, Princeton, N. J.

[4] P. OLSEN (1973), *Measurability in stochastic programming*, Tech. Rep. 196, Dept. of Operations Research, Cornell Univ., Ithaca, N. Y.

[5] ——— (1974), *Multistage stochastic programming with recourse: The equivalent deterministic problem*, this Journal, 14 (1976), pp. 495–517.

[6] R. T. ROCKAFELLAR (1969), *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28, pp. 4–25.

[7] ——— (1970), *Convex Analysis*, Princeton University Press, Princeton, N. J.

[8] W. RUDIN (1966), *Real and Complex Analysis*, McGraw-Hill, New York.

[9] D. W. WALKUP AND R. WETS (1967), *Stochastic programs with recourse*, SIAM J. Appl. Math., 15, pp. 1299–1314.

[10] R. WETS (1972), *Stochastic programs with recourse: A basic theorem for multistage problems*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 21, pp. 201–206.

[8] Strictly speaking, the argument is circular since Proposition A.1 makes no sense unless $\bar{p}_K, \cdots, \bar{p}_1$ are essentially defined. But the proposition's inductive proof can be woven into the theorem's (inductive) proof to avoid the circularity.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

MULTISTAGE STOCHASTIC PROGRAMMING WITH RECOURSE AS MATHEMATICAL PROGRAMMING IN AN L_p SPACE*

PAUL OLSEN†

Abstract. Multistage stochastic programming with recourse has been formulated in terms of a recursive sequence of parameterized, finite-dimensional mathematical programming problems. It has also been formulated as mathematical programming in an L_p-space. The two formulations are reconciled by showing that the L_p-space, or *static*, formulation is a restriction of the recursive, or *dynamic*, formulation and by deriving conditions under which any solution to the static formulation solves the dynamic formulation.

Introduction. Wets [15] identified two-stage *stochastic programming with recourse* with the problem

$$\text{SP2: } \underset{u}{\text{minimize}}\ c_0u_0 + E[\inf_{u_1}\{c_1u_1 : A_{11}u_1 = X_1 - A_{10}u_0,\ u_1 \geqq 0\}]$$

$$\text{subject to } A_{00}u_0 = b_0, \quad u_0 \geqq 0.$$

Dantzig [3] (also see [2, Chap. 25]) introduced essentially the same problem under the rubric *linear programming under uncertainty*. In stage 0 the decision-maker seeks $u_0 \in R^{n_0}$ to minimize the sum of current costs, c_0u_0, and expected future costs (anticipating an optimal decision in stage 1) while satisfying the stage 0 constraints

$$A_{00}u_0 = b_0, \qquad u_0 \geqq 0.$$

In stage 1 he observes a realization x_1 of the random vector X_1, which determines the stage 1 resources and requirements; he then seeks $u_1 \in R^{n_1}$ to minimize the current costs, c_1u_1, (the future costs are 0) while satisfying the stage 1 constraints

$$A_{11}u_1 = x_1 - A_{10}u_0, \qquad u_1 \geqq 0.$$

The two-stage problem's structure extends naturally to $K+1$ stages ($1 \leqq K < +\infty$). The following formulation of multistage stochastic programming (SP) with recourse is slightly more general than SP2, as it admits nonlinear, stochastic costs and lets the vector of resources and requirements be a function of a random vector instead of the random vector itself.

For $0 \leqq k \leqq K$, let X_k be a random vector describing the state of the world in stage k; to be consistent with SP2, let X_0 assume a single value, $\bar{x}_0$, with probability 1. Let s_k be the number of components of X_k. Let $\underline{X}_k$ be the random vector $(X_0, \cdots, X_k)$; let $S_k = \sum_{i=0}^{k} s_i$. Let "x_k" denote a realization of X_k, and "$\underline{x}_k$" a realization of $\underline{X}_k$. Let u_k be the vector of stage k activity levels; $u_k \in R^{n_k}$.

* Received by the editors July 19, 1974, and in revised form April 27, 1975.
† Institute for Defense Analyses, Arlington, Virginia 22202.

Let $\underline{u}_k$ be the vector $(u_0, \cdots, u_k)$; let $N_k = \sum_{i=0}^{k} n_i$. The random vectors $X_0, \cdots, X_K$ are assumed to be defined on the same sample space and to have a known joint distribution independent of the decisions $u_o, \cdots, u_K$.

Let $p_{K+1}(\underline{u}_k; \underline{x}_k) \equiv 0$. Now let $1 \leqq k \leqq K$, and suppose that $\bar{p}_{k+1}: R^{N_k} \times R^{S_k} \to [-\infty, +\infty]$ has been defined. Let

$$r_k(\underline{u}_k; \underline{x}_k) \equiv c_k(\underline{u}_k; \underline{x}_k) + \bar{p}_{k+1}(\underline{u}_k; \underline{x}_k) + \psi(\underline{u}_k; \underline{x}_k),$$

where

$$\psi(\underline{u}_k; \underline{x}_k) = \begin{cases} 0, & \sum_{j=0}^{k} A_{kj} u_j = b_k(\underline{x}_k), \quad \underline{u}_k \geqq 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

By convention, $+\infty+(-\infty)=-\infty+(+\infty)=+\infty$. The current cost function c_k: $R^{N_k} \times R^{S_k} \to [-\infty, +\infty]$, and b_k: $R^{S_k} \to R^{m_k}$; $A_{k0}, \ldots, A_{kk}$ are real matrices. At stage k the decision-maker seeks to solve the problem

$$P_k(\underline{u}_{k-1}; \underline{x}_k): \underset{v \in R^{n_k}}{\text{minimize}}\ r_k(\underline{u}_{k-1}, v; \underline{x}_k).$$

Let $p_k(\underline{u}_{k-1}; \underline{x}_k) = \inf(P_k(\underline{u}_{k-1}; \underline{x}_k))$, the problem's optimal value ($+\infty$ if the problem is inconsistent). Let $F_{X_k|\underline{X}_{k-1}}$ be a regular conditional distribution function for X_k given $\underline{X}_{k-1}$ (for the definition see [1, p. 263] or [8]); its existence follows from [1, pp. 263–66]. Define

$$\bar{p}_k(\underline{u}_{k-1}; \underline{x}_{k-1}) \equiv \int p_k(\underline{u}_{k-1}; \underline{x}_k)\, dF_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1}).$$

Given the convention $+\infty-(+\infty)=+\infty$, the integral is defined whenever the integrand is measurable; for any measure μ and measurable, extended real-valued function f,

$$\int f\, du = \int f^+\, du - \int f^-\, du,$$

where f^+ is the positive part of f, and f^- the negative part. The definition of $\bar{p}_k$ completes the recursion.

The stage 0 problem is

$$P_0(x_0): \text{minimize } c_0(u_0; x_0) + \bar{p}_1(u_0; x_0)$$
$$\text{subject to } A_{00}u_0 = b_0(x_0), \quad u_0 \geqq 0.$$

Since $X_0 = \bar{x}_0$ almost surely, it makes sense to identify $c_0(u_0; \cdot)$, $\bar{p}_1(u_0; \cdot)$ and b_0 with their values at $\bar{x}_0$ and to write the stage 0 problem as

$$P_0: \text{minimize } c_0(u_0) + \bar{p}_1(u_0)$$
$$\text{subject to } A_{00}u_0 = b_0, \quad u_0 \geqq 0.$$

P_0 is the SP problem's *equivalent deterministic problem.*

The problem sequence $P_0, \cdots, P_K$ constitutes the dynamic formulation of the SP problem. The formulation $P_0, \cdots, P_K$ is the natural extension of SP2 because

$$\bar{p}_{k+1}(\underline{u}_k; \underline{x}_k) = E[p_k(\underline{u}_k; \underline{X}_{k+1})|\underline{X}_k = \underline{x}_k]$$

for almost every (a.e.) $\underline{x}_k$ [8]. At stage k the decision-maker observes the current state of the world, x_k (a realization of X_k); he knows the past states of the world,

$\underline{x}_{k-1}$, and his past decisions, $\underline{u}_{k-1}$. He seeks a program u_k to minimize the sum of current costs, $c_k(\underline{u}_k; \underline{x}_k)$, and expected future costs, $\bar{p}_{k+1}(\underline{u}_k; \underline{x}_k)$, (anticipating optimal decisions in the future) while satisfying the stage k constraints

$$A_{kk}u_k = b_k(\underline{x}_k) - \sum_{j=0}^{k-1} A_{kj}u_j, \qquad u_k \geqq 0.$$

The dynamic formulation $P_0, \cdots, P_K$ is basically the same as the earlier formulations of Dantzig [3] and Wets [16], [17]. The principal difference is technical: Dantzig and Wets define $\bar{p}_{k+1}$ simply as a conditional expectation rather than an integral over a regular conditional distribution function, with the result that $\bar{p}_1$ may be ill-defined. This and related measure-theoretic issues are treated in [8] and will be skirted here.

In contrast to the dynamic formulation, which defines the SP problem recursively in finite-dimensional space, the static formulation defines it as mathematical programming in function space. Eisner [4] defined it as a mathematical programming problem in an L_p-space; the same approach was taken in [5]; Rockafellar and Wets [12] took a similar but more general approach. In this paper the static formulation is developed from the dynamic formulation. The resulting static formulation, which embraces Eisner's as a special case, is shown to be a restriction of the dynamic formulation. The two are shown to be equivalent under certain conditions on the problem data.

2. The existence of measurable optimal programs. For an arbitrary mathematical programming problem

$$\begin{aligned} &\text{MP: minimize } f(u) \\ &\text{subject to } u \in Q, \end{aligned}$$

one says that a program u *solves* MP if and only if: (i) $u \in Q$ and $f(u) = \inf$ (MP); or (ii) $\inf$ (MP) $= +\infty$. If MP has a solution, one may denote MP's optimal value by "min (MP)" instead of "inf (MP)". In view of the above, it is clear what is meant by the statement, "u_0 solves P_0."

DEFINITION 2.1. Let $1 \leqq k \leqq K$. Suppose $u_0, \cdots, u_{k-1}$ solve $P_0, \cdots, P_{k-1}$. Then $u_0, \cdots, u_k$ solve $P_0, \cdots, P_k$ if and only if: (i) $u_k: R^{S_k} \to R^{n_k}$ is Borel measurable, and $u_k(\underline{x}_k)$ solves $P_k(\underline{u}_{k-1}(\underline{x}_{k-1}); \underline{x}_k)$ for almost every (a.e.) $\underline{x}_k$; or (ii) $\inf(P_0) = +\infty$.

THEOREM 2.2. *Assume for each* $1 \leqq k \leqq K$:

(a) c_k *and* b_k *are Borel measurable.*

(b) $c_k(\cdot\,; \underline{x}_k)$ *is proper, lower-semicontinuous, and convex for a.e.* $\underline{x}_k$.[1]

(c) *There are functions* β_k *and* α_k *in* $L_1(R^{S_k})$ *with values in* $[-\infty, 0]$ *such that*

$$c_k(\underline{u}_k; \underline{x}_k) \geqq \beta_k(\underline{x}_k)|\underline{u}_k| + \alpha_k(\underline{x}_k)$$

for every $\underline{u}_k$ *and every* $\underline{x}_k$, *and*

$$E|\beta_k(\underline{X}_k)b_i(\underline{X}_i)| < +\infty$$

for each $1 \leqq i \leqq k$.[2]

(d) $\{w: A_{kk}w = 0, w \geqq 0\} = \{0\}$.

---

[1] An extended real-valued function is proper if it is nowhere $-\infty$ and not everywhere $+\infty$ [9].

[2] The notation implies that $\underline{X}_i$ is a subvector of $\underline{X}_k$.

Assume that c_0 *is lower-semicontinuous, proper and convex and that* $\{w: A_{00}w = 0,\ w \geqq 0\} = \{0\}$.

Then there is a sequence of programs $u_0, \cdots, u_K$ *solving* $P_0, \cdots, P_K$.

The conclusion is meaningless unless $P_0, \cdots, P_K$ are well-defined; Theorems 2.1 and 3.4 of [8] show that, given (a), (b), (c) and (d), $P_0, \cdots, P_K$ are well-defined.

Proof of Theorem 2.2. Assume that $\inf(P_0) < +\infty$; otherwise, the conclusion is trivially true.

For each $0 \leqq k \leqq K$, there is a Borel measurable function $\bar{p}^0_{k+1}$ such that: $\bar{p}^0_{k+1}(\cdot\,; \underline{x}_k) = \bar{p}_{k+1}(\cdot\,; \underline{x}_k)$ for a.e. $\underline{x}_k$; and $\bar{p}^0_{k+1}(\cdot\,; \underline{x}_k)$ is lower-semicontinuous (l.s.c.) and convex for a.e. $\underline{x}_k$ [8, Thm. 3.4]. Substitute $\bar{p}^0_{k+1}$ for $\bar{p}_{k+1}$ in the definition of P_k for each $0 \leqq k \leqq K$, and drop the superscript "0". This leaves $P_0, \cdots, P_K$ essentially unchanged, and therefore, if $u_0, \cdots, u_K$ solve the new sequence of problems, they also solve the original sequence.

By Proposition A.1 (Appendix), $\bar{p}_1 > -\infty$. Then since $c_0 > -\infty$, P_0's objective function is l.s.c. and convex [9, Thm. 9.3]. P_0's feasible region is bounded. It follows that P_0 is solvable. Choose u_0 to solve P_0; $\bar{p}_1(u_0) < +\infty$ because $\inf(P_0) < +\infty$.

Now let $1 \leqq k \leqq K$, and assume that $u_0, \cdots, u_{k-1}$ solve $P_0, \cdots, P_{k-1}$, with

$$\bar{p}_k(\underline{u}_{k-1}(\underline{x}_{k-1}); \underline{x}_{k-1}) < +\infty \quad \text{for a.e. } \underline{x}_{k-1}. \tag{1}$$

Define

$$\hat{r}_k(\cdot\,; \underline{x}_k) = r_k(\underline{u}_{k-1}(\underline{x}_{k-1}), \cdot\,; \underline{x}_k).$$

Since c_k, $\bar{p}_{k+1}$ and b_k are Borel measurable, so is r_k, and then since $\underline{u}_{k-1}$ is Borel measurable, so is $\hat{r}_k$. Since, by Proposition A.1, $p_k(\cdot\,; \underline{x}_k) > -\infty$ for a.e. $\underline{x}_k$, $\hat{r}_k(\cdot\,; \underline{x}_k) > -\infty$ for a.e. $\underline{x}_k$. Line (1) above implies that

$$p_k(\underline{u}_{k-1}(\underline{x}_{k-1}); \underline{x}_k) < +\infty \quad \text{for a.e. } \underline{x}_k,$$

and therefore, for a.e. $\underline{x}_k$, $\hat{r}_k(v; \underline{x}_k) < +\infty$ for some $v \in R^{n_k}$. Combining these results reveals the existence of a Borel set T in R^{S_k} such that: $P\{\underline{X}_k \in T\} = 1$; and for each $\underline{x}_k \in T$, $\hat{r}_k(\cdot\,; \underline{x}_k)$ is l.s.c., proper and convex, and

$$\inf_v \hat{r}_k(v; \underline{x}_k) = p_k(\underline{u}_{k-1}(x_{k-1}); \underline{x}_k)$$

is finite. Apply [11, Thm. 5] and then [11, Corollary 4.3] to verify the existence of a function $\hat{u}_k$: $T \to R^{n_k}$ such that: $\hat{u}_k$ is $\mathcal{T}$-measurable, $\mathcal{T}$ being the completion of the σ-algebra of Borel sets in T with respect to the measure induced by $\underline{X}_k$; and

$$r_k(\underline{u}_{k-1}(\underline{x}_{k-1}), \hat{u}_k(\underline{x}_k); \underline{x}_k) = p_k(\underline{u}_{k-1}(\underline{x}_{k-1}); \underline{x}_k) \quad \forall \underline{x}_k \in T.$$

Choose $u_k: R^{S_k} \to R^{n_k}$ to be any Borel measurable function equal to $\hat{u}_k$ a.e. on T. Of course,

$$\bar{p}_{k+1}(\underline{u}_k(\underline{x}_k); \underline{x}_k) < +\infty \quad \text{for a.e. } \underline{x}_k,$$

which completes the induction step. □

3. The static formulation of the SP problem. The preceding section having defined the concept of $u_0, \cdots, u_K$ solving $P_0, \cdots, P_K$, it is natural to inquire

whether there is a mathematical programming problem, P, to which the function $u=(u_0, \cdots, u_K)$ is a solution. Since u_k is a Borel measurable function on R^{S_k} (Definition 2.1), and therefore by extension a Borel measurable function on R^S, where S is the number of components of $(X_0, \cdots, X_K)$, the variables in P should be Borel measurable functions on R^S. In fact, the variables in P will be restricted to an L_p-space with $p \in [1, +\infty]$.

To introduce P some notation must be defined. Let $\mathcal{B}(R^S)$ be the σ-algebra of Borel sets in R^S. Let μ be the probability measure on $\mathcal{B}(R^S)$ determined by the random vector $X \triangleq (X_0, \cdots, X_K)$. If $u: (R^S, \mathcal{B}(R^S), \mu) \to R^n$, and $p \in (0, +\infty)$,

$$\|u\|_p \triangleq \left(\int |u(x)|^p \, d\mu \right)^{1/p},$$

where $|\cdot|$ is the Euclidean norm on R^n (any fixed norm on R^n would do), while $\|u\|_\infty$ is the essential supremum of the function $|u(x)|$. Let A be the matrix whose (k, j)-element is itself a matrix—namely, A_{kj} if $j \leqq k$ and 0 if not. (Recall the matrices $A_{k0}, \cdots, A_{kk}$ from the definitions of P_k.) Let M be the number of rows in A and N the number of columns $(N = N_K)$. Fix $p \in [1, +\infty]$. Define $E = L_p^N(R^S, \mathcal{B}(R^S), \mu)$ and $F = L_p^M(R^S, \mathcal{B}(R^S), \mu)$. E and F are considered seminormed vector spaces of functions; thus, if $u, v \in E$, $u = v$ if and only if $u(x) = v(x)$ for each $x \in R^S$.

E will be the optimization space for P. If $u_0, \cdots, u_K$ solve $P_0, \cdots, P_K$, the vector $u = (u_0, \cdots, u_K)$ has the property that

$$u_k(x') = u_k(x'') \quad \text{if } \underline{x}'_k = \underline{x}''_k \quad \forall\, 0 \leqq k \leqq K;$$

let U be the set of every $u \in E$ with that property. It is the subspace of E consisting of every function u whose stage k components, u_k, depend only on information known at stage k; in the terminology of Rockafellar and Wets [12], U is the set of *nonanticipative* functions in E.

Let $b = (b_0, \cdots, b_K)$. For any $u: (R^S, \mathcal{B}(R^S)) \to R^N$, let

$$c(u) = E\left[\sum_{k=0}^{K} c_k(u_k(X); \underline{X}_k) \right]$$

($\underline{X}_k$ is a subvector of X); the convention that $+\infty + (-\infty) = -\infty + (+\infty) = +\infty$ remains in force. Define $T: E \to F$ such that $(Tu)(x) = A(u(x))$ for each $x \in R^S$. T is continuous and linear.

The static formulation of the SP problem of which $P_0, \cdots, P_K$ constitute the dynamic formulation is

$$\begin{aligned} &P\text{: minimize } c(u), \quad u \in E, \\ &\text{subject to } Tu = b \quad \text{a.e.}(\mu), \\ &\qquad\qquad u \geqq 0 \quad \text{a.e.}(\mu), \quad u \in U. \end{aligned}$$

(Of course, P is inconsistent unless $b \in F$.) Eisner's formulation of multistage SP with recourse [4] is essentially P with c a continuous linear functional. P is a special case of the formulation in [12].

Although the dynamic formulation of multistage SP is the natural extension of the two-stage SP problem introduced by Dantzig [3] and investigated in depth by Walkup and Wets [13], [14], the static formulation has an independent rationale. (See [10].) The decision-maker assigns the cost

$$\sum_{k=0}^{K} c_k(v_0, \ldots, v_k; \underline{x}_k)$$

to the sequence of decisions $v_0, \cdots, v_K$ ($v_k \in R^{n_k}$) when the sequence of states of the world is $x_0, \cdots, x_K$. He seeks a program $u: R^S \to R^N$; $u_k(x)$ tells him what decision to make in stage k if state of the world is x_0 in stage $0, \cdots, x_k$ in stage $k, \cdots,$ and x_K in stage K. But in stage k he will not know $x_{k+1}, \cdots, x_K$; hence, u_k must depend only on $x_0, \cdots, x_k$—i.e., u must be chosen from U, the set of nonanticipative programs.

Both formulations of SP with recourse have appealing rationales: in the dynamic formulation, the decision-maker seeks a stage 0 program $u_0 \in R^{n_0}$ to minimize the sum of current cost and expected future cost, which is computed recursively; in the static formulation, he seeks a program $u \in U$, determining a response to each possible state of the world, so as to minimize the expected total cost from stages 0 through K. It would be distressing if the two formulations were not intimately related. A close relationship can be useful, as well as emotionally satisfying, because for some purposes one formulation is more convenient than the other. The static formulation is more computationally tractable than the dynamic formulation; it admits an elegant duality theory (see, for example, [4], [5] and [10]), and can be solved in some cases by solving a sequence of finite-dimensional "discretizations" [6]. The dynamic formulation, as a recursive sequence of (parameterized) finite-dimensional problems, is generally more fruitful than the static formulation for characterizing solutions. In the duality and discretization theories for P, it is important to know when P has a solution satisfying the constraints everywhere on some set of measure 1, not just almost everywhere; the dynamic formulation can furnish conditions. (See the use of results from [12] in [10] and results from [7] in [7a].

This paper's principal results are Theorems 3.1 and 3.2. They will be proved via two propositions interesting in their own right.

Define

$$\bar{c}(u) = E\left[\sum_{k=0}^{K} \min\{c_k(\underline{u}_k(X); \underline{X}_k), 0\}\right]$$

and

$$\underline{c}(u) = \sum_{k=0}^{K} E[c_k(\underline{u}_k(X); \underline{X}_k)]$$

for any $u: (R^S, \mathcal{B}(R^S)) \to R^N$. Of course, $\underline{c}(u) \leqq c(u) \leqq \bar{c}(u)$.

THEOREM 3.1. *Assume* $u_0, \cdots, u_K$ *solve* $P_0, \cdots, P_K$ *and* $u \in E$ $(u = (u_0, \cdots, u_K))$. *If* $-\infty < \underline{c}(u)$ *or* $\bar{c}(u) < +\infty$, *then* u *solves* P, *and* $\min(P) = \min(P_0)$.

THEOREM 3.2. *Assume* $\min(P) = \inf(P_0)$. *If* u *solves* P, *then* $u_0, \cdots, u_K$ *solve* $P_0, \cdots, P_K$.

Combining the two theorems produces

THEOREM 3.3. *Assume there is a sequence of programs* $\bar{u}_0, \cdots, \bar{u}_K$ *solving* $P_0, \cdots, P_K$ *such that* $\|\bar{u}\|_p < +\infty$ *and* $-\infty < \underline{c}(\bar{u})$ *or* $\bar{c}(\bar{u}) < +\infty$. *Then* P *is solvable, and if* u *solves* P, $u_0, \cdots, u_K$ *solve* $P_0, \cdots, P_K$.

Of course, Theorem 2.2 gives a sufficient condition for the existence of programs $u_0. \cdots, u_K$ solving $P_0, \cdots, P_K$. Moreover, its hypothesis leads to a simple condition assuring that $\|u\|_p < +\infty$. The hypothesis of Theorem 2.2 implies that

$$\{w: Aw = 0,\ w \geqq 0\} = \{0\}.$$

If $u_0, \cdots, u_K$ solve $P_0, \cdots, P_K$, $Au(x) = b(x)$ and $u(x) \geqq 0$ for a.e. x. It follows that if $\|b\|_p < +\infty$ and $u_0, \cdots, u_K$ solve $P_0, \cdots, P_K$, $\|u\|_p < +\infty$ (use [7, Lemma 2.1.]).

The assumption of Theorem 3.3. that $\underline{c}(u) > -\infty$ or $\bar{c}(u) < +\infty$ holds automatically if $c_0, \cdots, c_K$ have certain boundedness properties:

(i) Suppose that for each $1 \leqq k \leqq K$ for a.e. $\underline{x}_k$,

$$c_k(z; \underline{x}_k) \leqq \beta_k(\underline{x}_k)|z| + \alpha_k(\underline{x}_k)$$

for every $z \in R^{N_k}$, with $\beta_k \in L_q(R^{S_k})$ and $\alpha_k \in L_1(R^{S_k})$—$q$ being the conjugate exponent of p. Also suppose that $c_0 < +\infty$. Then $c(u) < +\infty$ for every $u \in E$ (by Hölder's inequality).

(ii) Suppose that for each $1 \leqq k \leqq K$ for a.e. $\underline{x}_k$,

$$c_k(z; \underline{x}_k) \geqq \beta_k(\underline{x}_k)|z| + \alpha_k(\underline{x}_k)$$

for every $z \in R^{N_k}$, with $\beta_k \in L_q(R^{S_k})$ and $\alpha_k \in L_1(R^{S_k})$. Also suppose that $c_0 > -\infty$. Then $\underline{c}(u) > -\infty$ for any $u \in E$.

Of course, if $c(u) \equiv \langle u, c^* \rangle$ for some $c^* \in L_q^N(R^S, \mathcal{B}(R^S), \mu)$—with $1/p + 1/q = 1$—the hypothesis of (i) and the hypothesis of (ii) both hold.

DEFINITION 3.4. u_0 satisfies the explicit constraints of P_0 if and only if $u_0 \in R^{n_0}$, $A_{00}u_0 = b_0$ and $u_0 \geqq 0$. The function $u_0, \cdots, u_k$ *satisfy the explicit constraints* of $P_0, \cdots, P_k$ if and only if: $u_0, \cdots, u_{k-1}$ satisfy the explicit constraints of $P_0, \cdots, P_{k-1}$; $u_k: (R^{S_k}, \mathcal{B}(R^{S_k})) \to R^{n_k}$; and for a.e. $\underline{x}_k$,

$$\sum_{j=0}^{k} A_{kj}u_j(\underline{x}_j) = b_k(\underline{x}_k), \qquad \underline{u}_k(\underline{x}_k) \geqq 0.$$

PROPOSITION 3.5. *If* $u_0, \cdots, u_K$ *satisfy the explicit constraints of* $P_0, \cdots, P_K$, *then*

$$\inf(P_0) \leqq E\left[\sum_{i=0}^{k} c_i(\underline{u}_i(\underline{X}_i); \underline{X}_i) + \bar{p}_{k+1}(\underline{u}_k(\underline{X}_k); \underline{X}_k)\right] \leqq c(u)$$

for each $0 \leqq k \leqq K$.

Proof. The first inequality clearly holds if $k = 0$. Let $0 \leqq k \leqq K$ and suppose that the first inequality holds for this value of k. Let $F_{\underline{X}_k}$ be the distribution function of $\underline{X}_k$.

$$\inf(P_0) \leqq \int \left[\sum_{i=0}^{k} c_i(\underline{u}_i(\underline{x}_i); \underline{x}_i) + \bar{p}_{k+1}(\underline{u}_k(\underline{x}_k); \underline{x}_k)\right] dF_{\underline{X}_k}(\underline{x}_k)$$

$$= \int \left[\sum_{i=0}^{k} c_i(\underline{u}_i(\underline{x}_i); \underline{x}_i)\right.$$

$$
\begin{aligned}
&\int p_{k+1}(\underline{u}_k(\underline{x}_k);\underline{x}_{k+1})\,dF_{\underline{X}_{k+1}|\underline{X}_k}(x_{k+1}|\underline{x}_k)\Big]\,dF_{\underline{X}_k}(\underline{x}_k)\\
&=\int\int\Big[\sum_{i=0}^{k} c_i(\underline{u}_i(\underline{x}_i);\underline{x})\\
&\qquad +p_{k+1}(\underline{u}_k(\underline{x}_k);\underline{x}_{k+1})\Big]\,dF_{\underline{X}_{k+1}|\underline{X}_k}(x_{k+1}|\underline{x}_k)\,dF_{\underline{X}_k}(\underline{x}_k) \qquad (2)\\
&\leqq\int\int\Big[\sum_{i=0}^{k+1} c_i(\underline{u}_i(\underline{x}_i);\underline{x}_i)\\
&\qquad +p_{k+2}(\underline{u}_{k+1}(\underline{x}_{k+1});\underline{x}_{k+1})\Big]\,dF_{\underline{X}_{k+1}|\underline{X}_k}(x_{k+1}|\underline{x}_k)\,dF_{\underline{X}_k}(\underline{x}_k)\\
&\leqq\int\Big[\sum_{i=0}^{k+1} c_i(\underline{u}_i(\underline{x}_i);\underline{x}_i)+\bar{p}_{k+2}(\underline{u}_{k+1}(\underline{x}_{k+1});\underline{x}_{k+1})\Big]\,dF_{\underline{X}_{k+1}}(\underline{x}_{k+1}).
\end{aligned}
$$

The last inequality follows from [8, Thm. 2.1]—a slight extension of Fubini's Theorem. □

COROLLARY 3.6. $\inf(P_0)\leqq\inf(P)$.

Proof. If $\inf(P)=+\infty$, the conclusion is trivially true. Suppose $\inf(P)<+\infty$. Let u be feasible in P. Then $u_0,\cdots,u_K$ satisfy the explicit constraints of $P_0,\cdots,P_K$, and the conclusion follows from Proposition 3.5. □

The corollary and its proof make precise and verify the first section's assertion that the static formulation is a restriction of the dynamic formulation.

PROPOSITION 3.7. *Assume* $u_0,\cdots,u_K$ *solve* $P_0,\cdots,P_K$. *If* $-\infty<\underline{c}(u)$ *or* $\bar{c}(u)<+\infty$, $\inf(P_0)=c(u)$.

Proof. If $\inf(P_0)=+\infty$, $\inf(P_0)=c(u)$ by Proposition 3.5. Suppose $\inf(P_0)<+\infty$. Suppose $\underline{c}(u)>-\infty$. Let $1\leqq k\leqq K$. Assume that

$$c(u)=\sum_{i=0}^{k} E[c_i(\underline{u}_i(\underline{X}_i);\underline{X}_i)]+E[\bar{p}_{k+1}(\underline{u}_k(\underline{X}_k);\underline{X}_k)]$$

and $E[\bar{p}_{k+1}^{-}(\underline{u}_k(\underline{X}_k);\underline{X}_k)]<+\infty$. ($\bar{p}_{k+1}^{-}$ is the negative part of $\bar{p}_{k+1}$.) Then since $E[\bar{c}_k^{-}(\underline{u}_k(\underline{X}_k);\underline{X}_k)]<+\infty$,

$$c(u)=\sum_{i=0}^{k-1} E[c_i(\underline{u}_i(\underline{X}_i);\underline{X}_i)]+E[g(\underline{X}_k)],$$

where

$$g(\underline{x}_k)\equiv c_k(\underline{u}_k(\underline{x}_k);\underline{x}_k)+\bar{p}_{k+1}(\underline{u}_k(\underline{x}_k);\underline{x}_k).$$

Since $u_k(\underline{x}_k)$ solves $P_k(\underline{u}_{k-1}(\underline{x}_{k-1});\underline{x}_k)$ for a.e. $\underline{x}_k$ and $\inf(P_0)<+\infty$,

$$E[g(\underline{X}_k)]=E[p_k(\underline{u}_{k-1}(\underline{X}_{k-1});\underline{X}_k)]$$

and $E[p_k^{-}(\underline{u}_{k-1}(\underline{X}_{k-1});\underline{X}_k)]<+\infty$. The latter fact implies that

$$E[p_k(\underline{u}_{k-1}(\underline{X}_{k-1});\underline{X}_k)]=E[\bar{p}_k(\underline{u}_{k-1}(\underline{X}_{k-1});\underline{X}_{k-1})]$$

[8, Thm. 2.1]. Combining these results yields

$$c(u)=\sum_{i=0}^{k-1} E[c_i(\underline{u}_i(\underline{X}_i);\underline{X}_i)]+E[\bar{p}_k(\underline{u}_{k-1}(\underline{X}_{k-1});\underline{X}_{k-1})]. \qquad (3)$$

To complete the induction step, observe that

$$
\begin{aligned}
&E[\bar{p}_k^-(\underline{u}_{k-1}(\underline{X}_{k-1}); \underline{X}_{k-1})] \\
&\quad = \int \left[\int p_k(\underline{u}_{k-1}(\underline{x}_{k-1}); \underline{x}_k)\, dF_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1}) \right] dF_{\underline{X}_{k-1}}(\underline{x}_{k-1}) \\
&\quad \leqq \int\int \bar{p}_k^-(\underline{u}_{k-1}(\underline{x}_{k-1}); \underline{x}_k)\, dF_{X_k|\underline{X}_{k-1}}(x_k|\underline{x}_{k-1})\, dF_{\underline{X}_{k-1}}(\underline{x}_{k-1}) \\
&\quad = \int \bar{p}_k^-(\underline{u}_{k-1}(\underline{x}_{k-1}); \underline{x}_k)\, dF_{\underline{X}_k}(\underline{x}_k) \\
&\quad < +\infty.
\end{aligned}
$$

The induction hypothesis holds if $k = K$ since $\underline{c}(u) > -\infty$.

The induction shows that

$$c(u) = c_0(u_0) + \bar{p}_1(u_0).$$

The right-hand side equals $\inf(P_0)$ (still under the assumption that $\inf(P_0) < +\infty$.

To demonstrate the conclusion under the assumption that $\bar{c}(u) < +\infty$, repeat the proof from the beginning through equation (3), but substitute the positive part for the negative part everywhere. The remainder of the proof need not be repeated since (3) and the fact that $c(u) = \bar{c}(u) < +\infty$ imply that

$$E[\bar{p}_k^+(\underline{u}_{k-1}(\underline{X}_{k-1}); \underline{X}_{k-1})] < +\infty. \qquad \square$$

Proof of Theorem 3.1. If $\inf(P_0) = +\infty$, $\inf(P) = \inf(P_0)$ by Corollary 3.6, and P is trivially solvable. Suppose $\inf(P_0) < +\infty$. Since $u_0, \cdots, u_K$ solve $P_0, \cdots, P_K$, they satisfy the explicit constraints of $P_0, \cdots, P_K$. Hence $Tu = b$ a.e. and $u \geqq 0$ a.e. Since $u \in E$, $u \in U$. By Proposition 3.7 and Corollary 3.6,

$$c(u) = \inf(P_0) \leqq \inf(P).$$

But u is feasible in P. $\square$

Proof of Theorem 3.2. Suppose $\inf(P_0) < +\infty$; otherwise, the conclusion is trivially true. Let u solve P. Since $\inf(P) < +\infty$, $u_0, \cdots, u_K$ satisfy the explicit constraints of $P_0, \cdots, P_K$. Since

$$\inf(P_0) = \inf(P) = c(u) < +\infty,$$

every inequality in Proposition 3.5 and its proof must hold as an equality (for this u). Take $k = 0$ in the proposition's conclusion (with the inequalities changed to equalities) to verify that

$$\inf(P_0) = c_0(u_0) + \bar{p}_1(u_0);$$

thus u_0 solves P_0. When changed to an equality, inequality (2) in the proof of Proposition 3.5 implies that for each $0 \leqq k < K$

$$
\begin{aligned}
&p_{k+1}(\underline{u}_k(\underline{x}_k); \underline{x}_{k+1}) \\
&\quad = c_{k+1}(\underline{u}_{k+1}(\underline{x}_{k+1}); \underline{x}_{k+1}) + \bar{p}_{k+2}(\underline{u}_{k+1}(\underline{x}_{k+1}); \underline{x}_{k+1})
\end{aligned} \tag{4}
$$

for a.e. $\underline{x}_{k+1}$. Thus $u_0, \cdots, u_K$ solve $P_0, \cdots, P_K$. $\square$

Appendix.

PROPOSITION A.1 [7, Proposition 2.2]. *Assume that for each* $1 \leqq k \leqq K$:

(a) *For some selected functions* β_k *and* α_k *in* $L_1(R^{S_k})$ *with values in* $[-\infty, 0]$,

$$c_k(\underline{u}_k; \underline{x}_k) \geqq \beta_k(\underline{x}_k)|\underline{u}_k| + \alpha_k(\underline{x}_k)$$

for every $\underline{u}_k$ *and every* $\underline{x}_k$.

(b) $E|\beta_k(\underline{X}_k)b_i(\underline{X}_i)| < +\infty$ *for each* $1 \leqq i \leqq k$ ($\underline{X}_i$ *is a subvector of* $\underline{X}_k$).

(c) $\{w: A_{kk}w = 0,\ w \geqq 0\} = \{0\}$.

Then for each $1 \leqq k \leqq K$ *there are functions* δ_k *and* γ_k *in* $L_1(R^{S_k})$ *such that*

$$p_k(\underline{u}_{k-1}; \underline{x}_k) \geqq \delta_k(\underline{x}_k)|\underline{u}_{k-1}| + \gamma_k(\underline{x}_k)$$

for every $\underline{u}_{k-1}$ *and every* $\underline{x}_k$ *with* $\delta_k(\underline{x}_k)$ *and* $\gamma_k(\underline{x}_k) \in [-\infty, 0]$ *for every* $\underline{x}_k$. *For each* $0 \leqq k \leqq K$ *there are functions* $\bar{\delta}_{k+1}$ *and* $\bar{\gamma}_{k+1}$ *in* $L_1(R^{S_k})$ *such that*

$$\bar{p}_{k+1}(\underline{u}_k; \underline{x}_k) \geqq \bar{\delta}_{k+1}(\underline{x}_k)|\underline{u}_k| + \bar{\gamma}_{k+1}(\underline{x}_k)$$

for every $\underline{u}_k$ *and every* $\underline{x}_k$ *with* $\bar{\delta}_{k+1}(\underline{x}_k)$ *and* $\bar{\gamma}_{k+1}(\underline{x}_k) \in [-\infty, 0]$ *for every* $\underline{x}_k$.

REFERENCES

[1] R. B. ASH, *Real Analysis and Probability*, Academic Press, New York, 1972.

[2] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, N.J., 1963.

[3] ———, *Linear programming under uncertainty*, Management Sci., 1 (1955), pp. 197–206; reprinted in Mathematical Studies in Management Science, A. F. Veinott, Jr., ed., Macmillan, New York, 1965.

[4] M. J. EISNER, *On duality in infinite-player games and chance-constrained programming*, Ph.D. thesis, Cornell Univ., Ithaca, N.Y., 1970.

[5] M. J. EISNER AND P. OLSEN, *Duality for stochastic programming interpreted as LP in* L_p *space*, SIAM J. Appl. Math., 28 (1975), pp. 779–792.

[6] P. OLSEN, *Discretizations of multistage stochastic programming problems*, Stochastic Systems: Modeling, Identification, and Optimization, Mathematical Programming Studies, to appear.

[7] ———, *Multistage stochastic programming with recourse: The equivalent deterministic problem*, this Journal, 14 (1976), pp. 495–517.

[7a] ———, *Polyhedral convex feasible regions in stochastic programming with recourse*, Proc. 1975 IEEE Conference on Decision and Control, to appear.

[8] ———, *When is a multistage stochastic programming problem well defined?*, this Journal, 14 (1976), pp. 518–527.

[9] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[10] ———, *Lagrange multipliers for an N-stage model in stochastic convex programming*, Colloque d'Analyse Convexe, J. P. Aubin, ed., Springer-Verlag, New York, 1975.

[11] ———, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

[12] R. T. ROCKAFELLAR AND R. WETS, *Continuous versus measurable recourse in N-stage stochastic programming*, Ibid., 48 (1974), pp. 836–859.

[13] D. W. WALKUP AND R. WETS, *Stochastic programs with recourse*, SIAM J. Appl. Math., 15 (1967), pp. 1299–1314.

[14] ———, *Stochastic programs with recourse. II: On the continuity of the objective*, Ibid., 17 (1969), pp. 98–103.

[15] R. WETS, *Programming under uncertainty: The equivalent convex program*, Ibid., 14 (1966), pp. 89–105.

[16] ———, *Programming under uncertainty: The solution set*, Ibid, 14 (1966), pp. 1143–1151.

[17] ———, *Stochastic programs with recourse: A basic theorem for multistage problems*, Z. Wahrscheinlichkeitstheorie und Verw. Gebeite, 21 (1972), pp. 201–206.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

N-PLAYER STOCHASTIC DIFFERENTIAL GAMES*

PRAVIN VARAIYA†

Abstract. The paper presents conditions which guarantee that the control strategies adopted by N players constitute an efficient solution, an equilibrium, or a core solution. The system dynamics are described by an Ito equation, and all players have perfect information. When the set of instantaneous joint costs and velocity vectors is convex, the conditions are necessary.

1. Introduction. N players are simultaneously controlling the evolution of a system described by the Ito equation

$$dz_t = f(t, z, u_t^1, \cdots, u_t^N)\,dt + dB_t, \qquad t \in [0, 1], \tag{1}$$

where (z_t) is the state process, (B_t) is Brownian motion and (u_t^i) is the control of the ith player. Player i chooses this control so as to minimize the cost

$$J^i(\alpha) = E\left[\int_0^1 c^i(t, z, u_t)\,dt + \gamma^i(z)\right], \tag{2}$$

where $\alpha = (u_t) = (u_t^1, \cdots, u_t^N)$.

Different solution concepts of the resulting game are studied. Sufficient conditions are given which guarantee that $\alpha^* = (u_t^{1*}, \cdots, u_t^{N*})$ is a (Nash) equilibrium, a (Pareto) efficient solution, or a member of the core. When the set of admissible cost-drift vectors $(c^1, \cdots, c^N, f)$ possesses a certain convexity property, these sufficient conditions become necessary.

The next section gives a precise model of the game. The convexity property is stated, and its main implications are drawn out in § 3. The main results are given in § 4. A priori conditions on the c^i and f which imply the convexity property are examined in § 5.

2. The model.

2.1. Admissible controls. The sample paths of the state process (z_t) are evidently continuous, hence members of the Banach space C of all continuous functions $\omega: [0, 1] \to R^n$. Let ξ_t be the evaluation functional on C, that is, $\xi_t(\omega) = \omega(t)$. Let $\mathcal{F}_t$ be the σ-field of subsets of C generated by $\{\xi_s | 0 \leqq s \leqq t\}$. Let $\mathcal{F} = \mathcal{F}_1$.

For each i, U_i is a compact metric space, the set of *actions* available to i. A function $\alpha^i : [0, 1] \times C \to U_i$ is an (admissible) *control* for i if

(i) α^i is jointly measurable,

(ii) $u_t^i = \alpha^i(t, \cdot)$ is $\mathcal{F}_t$-measurable for all t.

$\mathcal{U}_i$ denotes the set of controls for i.

---

* Received by the editors December 18, 1974.

† Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This research was initiated at the University of California, Berkeley, under the Joint Services Electronic Program, Contract F44260-71-C-0087, and continued at the Decision and Control Sciences Group of the M.I.T. Electronic Systems Laboratory with partial support provided by AFOSR under Grant 72-2273 and by NASA Ames Research Center under Grant NGL-22-009-124.

Denote $U = U_1 \times \cdots \times U_N$ with elements $u = (u_1, \cdots, u_N)$ and $\mathcal{U} = \mathcal{U}^1 \times \cdots \times \mathcal{U}^N$ with elements $\alpha = (\alpha^1, \cdots, \alpha^N)$. For $u \in U$, $v \in U$ and $i \in \{1, \cdots, N\}$, denote $(u_{\bar{i}}, v_i) = (u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_N)$. More generally, for $S \subset \{1, \cdots, N\}$, denote $(u_{\bar{S}}, v_S)$ to be the N-tuple obtained from u upon replacing u_i by v_i for each $i \in S$. In exactly the same way one defines $(\alpha^{\bar{i}}, v^i)$ and $(\alpha^{\bar{S}}, v^S)$ when α and v are in $\mathcal{U}$.

2.2. Dynamics. The function $f : [0, 1] \times C \times U \to R^n$ in (1) satisfies the following conditions:

(i) f is jointly measurable,

(ii) $f(t, \cdot, u)$ is $\mathcal{F}_t$-measurable for all t, u and $f(t, \omega, \cdot)$ is continuous for all t, ω,

(iii) There is a constant k such that $|f(t, \omega, u)| \leqq k(1 + \|\omega\|)$ for all t, ω, u.

Let P denote Wiener measure on $(C, \mathcal{F})$. Let (z_t) be the family of evaluation functionals on C so that $(z_t, \mathcal{F}_t, P)$ is an n-dimensional, standard, Brownian motion. For $\alpha \in \mathcal{U}$, define the *drift* $(\phi_t^\alpha, \mathcal{F}_t, P)$ by

$$\phi_t^\alpha = f(t, z, \alpha(t, z))$$

and the *density* $(\rho_t^\alpha, \mathcal{F}_t, P)$ by

$$\rho_t^\alpha = \exp\left[\int_0^t \phi_s^\alpha \, dz_s - \frac{1}{2}\int_0^t |\phi_s^\alpha|^2 \, ds\right].$$

Denote $\rho^\alpha = \rho_1^\alpha$. The next result is well known [1], [2].

THEOREM 1 (Beneš). $E(\rho_t^\alpha) \equiv 1$. *Hence P^α is a probability measure on $(C, \mathcal{F})$, where*

$$P^\alpha(F) = \int_F \rho^\alpha(z) P(dz), \qquad F \in \mathcal{F}.$$

Furthermore, the process $(w_t^\alpha, \mathcal{F}_t, P^\alpha)$ defined by

$$w_t^\alpha = z_t - \int_0^t \phi_s^\alpha \, ds$$

is a Brownian motion.

This theorem justifies the following definition. The *solution* of (1) corresponding to $\alpha \in \mathcal{U}$ is the process $(z_t, \mathcal{F}_t, P^\alpha)$.

2.3. Solutions of the game. Conditions analogous to those imposed on f are also imposed on the functions c^i in (2). The functions $\gamma^i : C \to R$ are $\mathcal{F}$-measurable and integrable with respect to P^α for all α. In addition, the c^i and γ^i are nonnegative.

The *cost* to player i of $\alpha \in \mathcal{U}$ is defined to be

$$J^i(\alpha) = E^\alpha\left[\int_0^1 c^i(t, \cdot, u_t) \, dt + \gamma^i(\cdot)\right],$$

where E^α denotes expectation with respect to P^α.

Recall the following definitions. $u^* = (u^{1*}, \cdots, u^{N*})$ is

(a) an *equilibrium* if there is no i and no u such that

$$J^i(u^{*\bar{i}}, u^i) < J^i(u^*),$$

(b) *efficient* if there is no u such that

$$J^i(u) < J^i(u^*) \quad \text{for all } i,$$

(c) in the *core* if there is no S and no u such that

$$J^i(u^{*\bar{S}}, u^S) < J^i(u^*) \quad \text{for all } i \in S.$$

To avoid confusion it should be pointed out that the definitions (b) and (c) are not the standard ones. Usually, u^* is said to be efficient if there is no u such that $J^i(u) \leqq J^i(u^*)$ for all i with the strict inequality holding for at least one i. If one adopts this definition, then the observation at the beginning of § 4.3 below needs to be modified and so do the subsequent results; these modifications are slight but clumsy, and the definition given here avoids the clumsiness. In any case, the difference is slight. The core is usually defined only for games where comparison of inter-personal utilities is permitted and where side payments are allowed. For games where such comparison is not permitted, as is the normal posture in mathematical economics, one is naturally led to the definition given here.

3. The convexity property. Let $g(t, z, u) = (c^1(t, z, u), \cdots, c^N(t, z, u), f(t, z, u))$.

g is an $(N+n)$-dimensional vector.

The game is said to have the *convexity property* if for all t, z,

$$\{g(t, z, u) | u \in U\}$$

is a convex set. It is said to have the *strong convexity property* if for all t, z, u and for all S,

$$\{g(t, z, (u_{\bar{S}}, v_S)) | v \in U\}$$

is a convex set.

In [1] and [2] it is shown that the convexity property implies that the set of densities obtained by using all possible admissible controls is convex. The two lemmas below follow readily from these results.

LEMMA 1. *Suppose the game has the convexity property. Then*

$$\mathcal{J} = \{(J^1(u), \cdots, J^N(u)) | u \in \mathcal{U}\}$$

is a convex subset of R^N.

LEMMA 2. *Suppose the game has the strong convexity property. Let* $u \in \mathcal{U}$ *and* $S \subset \{1, \cdots, N\}$. *Then*

$$\mathcal{J}(u^{\bar{S}}) = \{(J^1(u^{\bar{S}}, v^S), \cdots, J^N(u^{\bar{S}}, v^S)) | v \in \mathcal{U}\}$$

is a convex subset of R^N.

4. The main results.

4.1. A result from control theory. Suppose $N = 1$ so that the game is simply an optimal control problem. Dropping superscripts and subscripts, the control

problem is to find $u^* \in \mathcal{U}$ so as to minimize

$$J(u) = E^u\left[\int_0^1 c(t, \cdot, u_t)\, dt + \gamma(\cdot)\right].$$

A minimizing control is said to be *optimal.*

The result below has been proved in [3] in a slightly more restrictive form than necessary.

THEOREM 2. *u^* is an optimal control if and only if there exist a constant J^*, and processes (ΛV_t), (∇V_t) with values in R, R^n, respectively, such that*

(i) $$J^* + \int_0^1 \Lambda V_t\, dt + \int_0^1 \nabla V_t\, dz_t = \gamma \quad a.s.,$$

(ii) $$\Lambda V_t + \min_{u \in U} \{\nabla V_t f(t, z, u) + c(t, z, u)\} = 0,$$

and the minimum is achieved at $u^*(t, z)$ for almost all t, z. Furthermore, $J^* = J(u^*)$ is the minimum cost; in fact,

$$J^* + \int_0^t \Lambda V_s\, ds + \int_0^t \nabla V_s\, dz_s = \min_{u \in \mathcal{U}} E^u\left\{\int_t^1 c^i(s, z, u_s)\, ds + \gamma^i | \mathcal{F}_t\right\}.$$

4.2. Conditions for equilibrium. The controls $u^* = (u^{*1}, \cdots, u^{*N})$ constitute an equilibrium if and only if for each i, u^{*i} minimizes $J^i(u^{*\bar{i}}, u^i)$ over the set $\mathcal{U}^i$. Theorem 2, therefore, immediately yields the next result.

THEOREM 3. *$u^* = (u^{*1}, \cdots, u^{*N})$ is an equilibrium if and only if for each i there exist a constant J^{*i}, and processes (ΛV_t^i), (∇V_t^i) such that*

(i) $$J^{*i} + \int_0^1 \Lambda V_t^i\, dt + \int_0^1 \nabla V_t^i\, dz_t = \gamma^i \quad a.s.,$$

(ii) $$\Lambda V_t^i + \min_{u_i \in U_i} \{\nabla V_t^i f(t, z, (u^{*\bar{i}}(t, z), u_i)) + c^i(t, z, (u^{*\bar{i}}(t, z), u_i))\} = 0,$$

and the minimum is achieved at $u^{*i}(t, z)$ a.s. Furthermore, $J^{*i} = J^i(u^*)$.

4.3. Conditions for efficiency. Consider the set $\mathcal{J} = \{J(u) = (J^1(u), \cdots, J^N(u)) | u \in \mathcal{U}\}$, the set of attainable cost vectors. Suppose there exists a *nonnegative* vector $\lambda = (\lambda_1, \cdots, \lambda_N) \neq 0$ and u^* such that

$$\lambda J(u^*) \leqq \lambda J \quad \text{for all } J \in \mathcal{J}. \tag{3}$$

It is then immediate that u^* is an efficient solution. It is also well known that (3) is a *necessary* condition in the event that $\mathcal{J}$ is a convex set. This observation, in conjunction with Theorem 2 and Lemma 1, imply the next result.

THEOREM 4. (a) *u^* is an efficient solution if there exist $\lambda \geqq 0$, $\lambda \neq 0$, and for each i a constant J^{*i}, and processes (ΛV_t^i), (∇V_t^i) such that*

(i) $$\sum \lambda_i \left[J^{*i} + \int_0^1 \Lambda V_t^i\, dt + \int_0^1 \nabla V_t^i\, dz_t\right] = \sum \lambda_i \gamma^i \quad a.s.,$$

(ii) $$\sum \lambda_i \Lambda V_t^i + \min_{u \in U} \sum \lambda_i \{\nabla V_t^i f(t, z, u) + c^i(t, z, u)\} = 0,$$

and the minimum is achieved at $u^*(t, z)$ a.s.

(b) *If the game has the convexity property, then the conditions above are necessary for efficiency.*

From a game-theoretic viewpoint, an efficient solution is of interest only insofar as it is also an equilibrium. The combination of the results above gives the first intriguing result. Its proof is given in the Appendix.

THEOREM 5. (a) u^* *is an efficient equilibrium if there exist for each* i *a constant* J^{*i}, *and processes* (ΛV_t^i), (∇V_t^i) *such that*

$$\text{(i)} \qquad J^{*i} + \int_0^1 \Lambda V_t^i \, dt + \int_0^1 \nabla V_t^i \, dz_t = \gamma^i \text{ a.s.},$$

$$\text{(ii)} \quad \Lambda V_t^i + \min_{u_i \in U_i} \{\nabla V_t^i f(t, z, (u^{*\bar{i}}(t, z), u_i)) + c^i(t, z, (u^{*\bar{i}}(t, z), u_i))\} = 0,$$

and the minimum is achieved at $u^{*i}(t, z)$ *a.s.*,

(iii) *there exist* $\lambda \geqq 0$, $\lambda \neq 0$ *such that*

$$\sum \lambda_i \{\nabla V_t^i f(t, z, u^*(t, z)) + c^i(t, z, u^*(t, z))\} = \min_{u \in U} \sum \lambda_i \{\nabla V_t^i f(t, z, u) + c^i(t, z, u)\} \text{ a.s.}$$

(b) *If the game has the convexity property, then the conditions above are also necessary.*

Remark. Define the Hamiltonian $H^i(t, z, u) = \nabla V_t^i f(t, z, u) + c^i(t, z, u)$. Condition (ii) above says that the ith Hamiltonian must be minimized along the ith "coordinate" u_i. Condition (iii) says that in order that the "private" minimization (implied in the equilibrium concept) also be "socially" efficient, this private minimization should lead to the "global" minimization of the social cost obtained as a weighted combination of the private costs. *The intriguing part of the result is that these weights, the* λ_i, *are constant, that is, they do not depend on time* t *or the random state* z.

4.4. Conditions for the core. The result for the core follows in the same way as Theorem 5.

THEOREM 6. (a) u^* *is in the core if there exist for each* i *a constant* J^{*i}, *and processes* (ΛV_t^i), (∇V_t^i) *such that*

$$\text{(i)} \qquad J^{*i} + \int_0^1 \Lambda V_t^i \, dt + \int_0^1 \nabla V_t^i \, dz_t = \gamma^i \quad \text{a.s.},$$

$$\text{(ii)} \quad \Lambda V_t^i + \min_{u_i \in U_i} \{\nabla V_t^i f(t, z, (u^{*\bar{i}}(t, z), u_i)) + c^i(t, z, (u^{*\bar{i}}(t, z), u_i))\} = 0,$$

and the minimum is achieved at $u^{*i}(t, z)$ *a.s.*;

(iii) *for each* S *there exist constants* $\lambda_i^S \geqq 0$, $i \in S$, *not all zero, such that*

$$\sum_{i \in S} \lambda_i^S \{\nabla V_t^i f(t, z, u^*(t, z)) + c^i(t, z, u^*(t, z))\} = \min_{u \in U} \sum_{i \in S} \lambda_i^S \{\nabla V_t^i f(t, z, (u^{*\bar{S}}(t, z), u_S)) + c^i(t, z, (u^{*\bar{S}}(t, z), u_S))\} \quad \text{a.s.}$$

(b) *If the game has the strong convexity property, then the conditions above are also necessary.*

Remark. It may appear reasonable, at first sight, to conjecture that the weights, λ_i^S, should not depend upon S. However, upon further reflection, the reader should become convinced that this is unlikely. Thus the weights associated with different players will vary with the coalition S in which they are being considered as members.

5. Randomized strategies. The convexity property is evidently quite restrictive. However, if one permits randomized controls, then convexity is guaranteed. To see this, define M_i as the set of all probability measures on U_i. U_i can then be regarded as a subset of M_i and the function f can be extended to the domain $[0, 1]\times C\times M_1\times\cdots\times M_N$ by setting

$$(4)\qquad f(t, z, m_1, \cdots, m_N)=\int_{U_1}\cdots\int_{U_N} f(t, z, u_1, \cdots, u_N)m_1(du_1)\cdots m_N(du_N).$$

The cost functions c^i can be extended analogously. The spaces M_i can be made compact and metrizable in a standard manner and $f(t, z, \cdot)$ remains continuous on $M=M_1\times\cdots\times M_N$. The controls for i are now randomized controls, that is, functions $m^i: [0, 1]\times C\to M_i$. The previous results continue to hold for this "extended" game. But notice from (4) that this extended game enjoys the convexity property, and if joint randomization is allowed, it also enjoys the strong convexity property.

6. Conclusions. These remarks are mainly suggestions for further research.

It is known that for deterministic differential games the condition that the weights λ_i are constant is sufficient but not necessary even when the game has the convexity property. The results presented here therefore convey surprise. However, it is not evident that these results should be regarded as curiosities or as significant. To decide this, it is necessary to clarify the precise role played by the Brownian motion in (1). Such clarification should also aid in restoring a measure of unity to the currently disparate traditions in the literature on deterministic and stochastic differential games. In the cases of control problems and two-player zero-sum games, this has been achieved by the important work of Fleming [4], [5] and subsequent work of Danskin [6] and Friedman [7], but it is not clear that these directions will prove useful for the many-player games.

This paper is not addressed to the important question of existence of solutions. For efficient controls, this question is immediately settled by known results on existence of optimal controls. A recent study [8] has nicely resolved the problem of existence of saddle points and value for two-player, zero-sum, stochastic differential games. It seems likely that the methods used in that study combined with the usual fixed-point arguments will help in proving existence of equilibrium solutions and the core.

Finally, the condition of complete information is a serious a priori restriction on the family of games considered in this paper. It is likely that results similar to those obtained here hold when all players have the *same* information even if it is incomplete [9]. The game is enormously more complicated when different players

have different information. In the context of static games, many important insights are provided by the results reported in [10], [11].

Note. Reference [9] contains several incorrect statements.

Appendix. *Proof of Theorem* 5. Part (a) of the theorem follows immediately from Theorem 3 and part (a) of Theorem 4. Hence it only remains to prove part (b).

By Theorem 3, there exist for each i, J^{*i}, and processes (ΛV_t^i), (∇V_t^i) such that

$$\text{(A.1)}\qquad \begin{aligned} &J^{*i}+\int_0^1 \Lambda V_t^i\,dt+\int_0^1 \nabla V_t^i\,dz_t=\gamma^i \quad \text{a.s.},\\ &\Lambda V_t^i+\min_{u_i\in U_i}\{\nabla V_t^i f(t,z,(u^{*\bar{i}}(t,z),u_i))+c^i(t,z,(u^{*\bar{i}}(t,z),u_i))\}=0 \end{aligned}$$

and the minimum is achieved at $\alpha^{*i}(t,z)$. On the other hand, by part (b) of Theorem 4, there exist $\lambda\geqq 0, \lambda\neq 0$ and for each i, K^{*i}, and processes (ΛW_t^i), (∇W_t^i) such that

$$\text{(A.2)}\qquad \begin{aligned} &\sum\lambda_i\left[K^{*i}+\int_0^1 \Lambda W_t^i\,dt+\int_0^1 \nabla W_t^i\,dz_t\right]=\sum\lambda_i\gamma^i \quad \text{a.s.},\\ &\sum\lambda_i\Lambda W_t^i+\min_{u\in U}\sum\lambda_i\{\nabla W_t^i f(t,z,u)+c^i(t,z,u)\}=0, \end{aligned}$$

and the minimum is achieved at $\alpha^*(t,z)$ a.s.

Comparison of these two sets of conditions reveals that it is enough to show that whenever (A.1) and (A.2) are both satisfied, then (A.2) is also satisfied by choosing

$$K^{*i}=J^{*i},\quad \Lambda W_t^i=\Lambda V_t^i,\quad \text{and}\quad \nabla W_t^i=\nabla V_t^i.$$

Now, by the last part of Theorem 2,

$$\begin{aligned} &J^{*i}+\int_0^t \Lambda V_s^i\,ds+\int_0^t \nabla V_s^i\,dz_s\\ &\quad=\lim_{\alpha\in\mathcal{U}} E^{\alpha}\left\{\int_t^1 c^i(s,z_s,(\alpha_s^{*i},\alpha_s^i))\,ds+\gamma^i|\mathcal{F}_t\right\}\\ &\quad=E^{\alpha^*}\left\{\int_t^1 c^i(s,z_s,\alpha_s^*)+\gamma^i|\mathcal{F}_t\right\}, \end{aligned}$$

and similarly,

$$\sum\lambda_i\left[K^{*i}+\int_0^t \Lambda W_s^i\,ds+\int_0^t \nabla W_s^i\,dz_s\right]=E^{\alpha^*}\left\{\sum\lambda_i\left[\int_t^1 c^i(s,z_s,\alpha_s^*)+\gamma^i\right]|\mathcal{F}_t\right\}.$$

Hence

$$\sum\lambda_i\left[J^{*i}+\int_0^t \Lambda V_s^i\,ds+\int_0^t \nabla V_s^i\,dz_s\right]=\sum\lambda_i\left[K^{*i}+\int_0^t \Lambda W_s^i\,ds+\int_0^t \nabla W_s^i\,dz_s\right].$$

Setting $t = 0$ yields $\sum \lambda_i J^{*i} = \sum \lambda_i K^{*i}$ and so

$$\int_0^t \left(\sum \lambda_i V_s^i - \sum \lambda_i W_s^i\right) ds = \int_0^t \left(\sum \lambda_i \nabla W_s^i - \sum \lambda_i \nabla V_s^i\right) dz_s.$$

But, under the measure P, (z_t) is a Brownian motion so that the term on the right is a continuous martingale, whereas the term on the left is a process with integrable variation. It follows that both terms must vanish so that $\sum \lambda_i V_s^i = \sum \lambda_i W_s^i$ and $\sum \lambda_i \nabla V_s^i = \sum \lambda_i \nabla W_s^i$ and the result follows.

REFERENCES

[1] V. E. BENEŠ, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.

[2] T. E. DUNCAN AND P. VARAIYA, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.

[3] M. H. A. DAVIS AND P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.

[4] W. H. FLEMING, *The convergence problem for differential games, II*, Advances in Game Theory, Ann. Math. Studies, no. 52, 1964, pp. 195–210.

[5] ——— *The Cauchy problem for degenerate parabolic equations*, J. Math. Mech., 13 (1964), pp. 987–1008.

[6] J. DANSKIN, *Stochastic differential games*, to appear.

[7] A. FRIEDMAN, *Stochastic differential games*, Differential Equations, 11 (1972), pp. 79–108.

[8] R. ELLIOTT, *The existence of value in stochastic differential games*, to appear.

[9] P. VARAIYA, *N-person stochastic differential games*, The Theory and Application of Differential Games, J. D. Grote, ed., D. Reidel, Boston, 1975, pp. 97–106.

[10] T. BASAR AND Y-C. HO, *Informational properties of the Nash solutions of two stochastic nonzero-sum games*, J. Economic Theory, 7 (1974), pp. 370–387.

[11] Y-C. HO, I. BLAU AND T. BASAR, *A tale of four information structures*, Control Theory, Numerical Methods and Computer Systems Modelling, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Economics and Mathematical Systems, no. 107, Springer-Verlag, New York, 1975, pp. 85–96.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

CONTROLLABILITY AND NECESSARY CONDITIONS IN UNILATERAL PROBLEMS WITHOUT DIFFERENTIABILITY ASSUMPTIONS*

J. WARGA†

Abstract. We study the attainable set and derive necessary conditions for relaxed, original and strictly original minimum in control problems defined by ordinary differential equations with unilateral restrictions. The functions defining the problem are assumed to be Lipschitz-continuous in their dependence on the state variables except for the unilateral restriction where continuous differentiability is also required. We define an extremal control as one satisfying a generalized Pontryagin maximum principle, with set-valued "derivate containers" replacing nonexistent derivatives. We prove that a nonextremal control (either original or relaxed) yields an interior point of the attainable set generated by original controls, and that, in normal problems, a minimizing original solution must also be a minimizing relaxed solution. The proofs are carried out with the help of an inverse function theorem for Lipschitz-continuous functions that is formulated in terms of derivate containers.

1. Introduction. We shall study control problems that involve the relations

$$\dot{y}(t) = f(t, y(t), u(t)) \qquad \text{a.e. in } T = [t_0, t_1], \tag{1}$$

$$u(t) \in R^{\#}(t) \subset R \quad \text{a.e. in } T, \tag{2}$$

$$y(t_0) \in A_0 \subset \mathbb{R}^n, \tag{3}$$

$$h^2(t, y(t)) \in (-\infty, 0]^{m_2} \qquad (t \in T^h \subset T), \tag{4}$$

with a particular emphasis on two closely related subjects:

I. Properties of the attainable set of a function $h^1(y(t_1))$, that is, of the set of points $h^1(y(t_1)) \in \mathbb{R}^m$ corresponding to choices of (y, u) that satisfy relations (1)–(4);

II. Necessary conditions for a couple (y, u) to yield a minimum of $h^0(y(t_1))$ subject to relations (1)–(4) and the end condition $h^1(y(t_1)) \in A_1$.

We shall assume that the functions $f(t, v, r)$, $h^0(v)$, $h^1(v)$ and $h^2(t, v)$ are measurable in t, continuous in (v, r), and Lipschitz-continuous in v over bounded sets, and that h^2 is continuous and has a continuous partial derivative with respect to v. We shall consider the *original* problem, in which the *original control function* u is chosen from the set $\mathscr{R}^{\#}$ of measurable selections of $\mathscr{R}^{\#}$, as well as its relaxed version in the formulation of Gamkrelidze [3]. In that formulation, the original control functions u are embedded in the set $\mathscr{S}_G^{\#}$ of *Gamkrelidze controls* σ such that $\sigma(t)$ is a probability measure concentrated at $n+1$ or fewer points of $R^{\#}(t)$.

For $(\sigma, a) \in \mathscr{S}_G^{\#} \times A_0$, let $y(f, \sigma, a)$ denote the unique absolutely continuous solution y of the equation

$$y(t) = a + \int_{t_0}^{t} d\tau \int f(\tau, y(\tau), r)\sigma(\tau)(dr) \qquad (t \in T),$$

* Received by the editors April 9, 1975, and in revised form May 18, 1975.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115. This work was supported in part by the National Science Foundation under Grant GP-37507X.

if such a solution exists. Our present investigations have three primary objectives:

(a) To prove that for every "nonextremal" control $(\sigma_0, a_0) \in \mathcal{S}_G^{\#} \times A_0$, the point $h^1(y(f, \sigma_0, a_0)(t_1))$ is in the interior of the *original attainable set*

$$\{h^1(y(f, u, a)(t_1)) | u \in \mathcal{R}^{\#}, a \in A_0, h^2(t, y(f, u, a)(t)) \in (-\infty, 0]^{m_2} (t \in T^h)\};$$

(b) to derive necessary conditions for an original minimum; and

(c) to show that a strict original solution (that is, a minimizing original solution that is not at the same time a minimizing relaxed solution) cannot be present in "normal" problems.

The result (a) generalizes Yorke's [10] and overlaps with Schwarzkopf's [6] results for problems in which all the defining functions are continuously differentiable with respect to v. The result (c) generalizes [7, VI.2.3, p. 357]. The necessary conditions (b) extend those of Clarke [2] that apply to "free" problems in which restriction (4) and one of the boundary conditions at t_0 or t_1 are absent. On the other hand, Clarke's results apply to problems involving differential inclusions that appear more general than relations (1)–(2) and do not require representation in terms of controls. Furthermore, because we require the assumption that $h^2(t, \cdot)$ is differentiable, we are unable to completely extend to original unilateral problems the necessary conditions derived in [9] for relaxed unilateral problems.

As in every investigation related to necessary conditions under constraints, we are faced with the need to apply an "implicit point" theorem. However, no theorem of this kind employed until now in the study of necessary conditions (such as the finite-dimensional implicit function theorem, or its refinements by Halkin [4], or Brouwer's fixed point theorem, etc.) appeared helpful when we attempted to derive results of the type described in (a). This was the case even if the functions f, h^0 and h^1 were assumed continuously differentiable with respect to v and restriction (4) was dropped. Ultimately, we were able to complete our arguments with the help of a rather simple but apparently new "convex mapping" theorem (Lemma 3.3) that generalizes the finite-dimensional inverse function theorem to Lipschitz-continuous (but not necessarily differentiable) functions defined on convex sets. The basic tool in formulating and deriving this theorem, as well as in the study of the control problem proper, was the use of "derivate containers", first introduced in [8], that contain the derivatives of particular C^1 approximations to a given Lipschitz-continuous function. These derivate containers represent a kind of set-valued Fréchet derivatives of Lipschitz-continuous functions between finite-dimensional spaces and thus seem conceptually related to the subdifferentials of convex functions [5] and Clarke's [1] "generalized gradients". The concept of an "extremal" control (that satisfies a Pontryagin maximum principle) is defined in terms of these derivate containers instead of ordinary derivatives that may not exist.

Our basic results are stated in § 2. Some auxiliary lemmas, including the "convex mapping" theorem, are derived in § 3, while §§ 4 and 5 contain the proofs of the results stated in § 2.

2. Definitions and basic results. We denote the Lebesgue measure on $T = [t_0, t_1]$ by μ, and use the terms "a.e.", "a.a." (almost all) and "measurable" in the sense of Lebesgue. We assume we are given an open set $V \subset \mathbb{R}^n$, closed convex

sets $A_0 \subset V$ and $A_1 \subset \mathbb{R}^m$, a compact metric space R, a compact set $T^h \subset T$, and functions

$$f: T \times V \times R \to \mathbb{R}^n, \quad h^0: V \to \mathbb{R}, \quad h^1: V \to \mathbb{R}^m, \quad h^2: T^h \times V \to \mathbb{R}^{m_2}$$

that are measurable in t and continuous in (v, r). We also assume that, for every compact set $V^* \subset V$, the functions

$$f|T \times V^* \times R, \quad h^0|V^*, \quad h^1|V^*, \quad h^2|T^h \times V^*$$

are bounded and have a common Lipschitz-constant with respect to v, independent[1] of the arguments t and r, and that h^2 is continuous and has a continuous partial derivative with respect to v. We shall write h^2_v or generally g_v to indicate such a (partial) derivative.

Our problem involves a measurable mapping $R^{\#}: T \to \mathcal{K}(R)$, where $\mathcal{K}(R)$ is the collection of nonempty closed subsets of R with the Hausdorff metric. We denote by $\mathcal{R}^{\#}$ the set of all measurable selections of $R^{\#}$, and refer to them as *original control functions*. If X is a compact metric space, we denote by $\mathrm{rpm}(X)$ the set of all Radon probability measures on X with the relative weak $*$ topology of $C(X)^*$. We denote by $\mathcal{S}^{\#}$ the collection of all measurable functions $\sigma: T \to \mathrm{rpm}(R)$ such that $\sigma(t)(R^{\#}(t)) = 1$ a.e. in T. We embed the set $\mathcal{S}^{\#}$ of *relaxed control functions* in $L^1(\mu, C(R))^*$ with its weak $*$ topology by identifying $\sigma \in \mathcal{S}^{\#}$ with the continuous linear functional

$$\phi \to \int_{t_0}^{t_1} dt \int \phi(t, r)\sigma(t)(dr),$$

and recall [7, IV.3.11, p. 287] that $\mathcal{S}^{\#}$ is convex and compact and its topology is derived from a "weak" norm $|\cdot|_w$ on $L^1(\mu, C(R))^*$ [7, IV.1.9, p. 272]. We embed $\mathcal{R}^{\#}$ in $\mathcal{S}^{\#}$ by identifying each $r \in R$ with the Dirac measure δ_r at r.

We write

$$f(t, v, \sigma(t)) = \int f(t, v, r)\sigma(t)(dr),$$

and similarly for other functions.

Let $\mathcal{S}^{\#}_G$ denote the set of all $\sigma \in \mathcal{S}^{\#}$ such that $\sigma(t)$ is, for all $t \in T$, concentrated at $n+1$ or fewer points of $R^{\#}(t)$. (This type of control was apparently first used by Gamkrelidze [3]). It is well known [7, VI.3.2, p. 370] that, for every $(\sigma, a_0) \in \mathcal{S}^{\#} \times A_0$ for which the equation

$$y(t) = a_0 + \int_{t_0}^{t} f(\tau, y(\tau), \sigma(\tau))\, d\tau \qquad (t \in T),$$

[1] The assumption about f can be replaced by one stating that, for each compact $V^* \subset V$, there exists an integrable function $\psi: T \to \mathbb{R}$ that is both a bound and a Lipschitz-constant for $f(t, \cdot, r)$. Since our future arguments will apply to a fixed set V^*, we can replace ψ by a constant by choosing the indefinite integral of ψ as a new independent variable instead of t.

has a solution $y(f, \sigma, a_0)$ (which must be unique), there exists $\sigma_G \in \mathcal{S}^{\#}_G$ such that $y(f, \sigma, a_0) = y(f, \sigma_G, a_0)$. This observation justifies[2] our formulating our results in terms of the elements of $\mathcal{S}^{\#}_G$ which we shall treat as a topological subspace of $\mathcal{S}^{\#}$. However, $\mathcal{S}^{\#}$ will remain a basic tool in the derivation of these results.

It is clear that for every $\sigma_G \in \mathcal{S}^{\#}_G$ there exist $\rho_j \in \mathcal{R}^{\#}$ and measurable $\alpha^j : T \to [0, 1]$ $(j = 0, \cdots, n)$ such that

$$\sigma_G(t)(\{\rho_j(t)\}) = \alpha^j(t), \qquad \sum_{j=0}^{n} \alpha^j(t) = 1 \qquad (t \in T).$$

We may, and henceforth always shall, assume that the points $\rho_j(t)$ $(\alpha^j(t) \neq 0)$ are distinct for each $t \in T$. We shall write $\sigma_G = [\alpha^j, \rho_j]$ to describe the above relations and assumption.

We let e_k denote the kth column of the unit matrix I of appropriate dimension; $d(x, y)$, $d[x, A]$, $d[B, A]$ the distance between two points, a point and a set and two sets, respectively; let

$$S^F(x, \alpha) = \{y | d(y, x) \leqq \alpha\}, \qquad S^F(A, \alpha) = \{y | d[y, A] \leqq \alpha\};$$

let $\phi'(v)$ or $\phi_v(v)$ denote the Fréchet derivative; $\mathcal{L}(\mathbb{R}^a, \mathbb{R}^b)$ the space of real $b \times a$ matrices; co and $\overline{co}$, the convex hull and the convex closure; and A°, $\bar{A}$, ∂A the interior, the closure and the boundary of A. For

$$x = (x^1, \cdots, x^a) \in \mathbb{R}^a, \qquad M = (M_{ij}) \in \mathcal{L}(\mathbb{R}^a, \mathbb{R}^b),$$

we set

$$|x| = \max_i |x^i| \quad \text{and} \quad |M| = \max_i \sum_j |M_{ij}|.$$

Unless otherwise specified, we define distances in $\mathbb{R}^a$ and $\mathcal{L}(\mathbb{R}^a, \mathbb{R}^b)$ accordingly. We write M^T for a transposed matrix, v^T for a row vector and $v^T w$ or $v \cdot w$ for the scalar product. Thus, when convenient, we shall write $e_j \cdot v$ or $e_j^T v$ for the jth component of v.

DEFINITION 2.1. Let A be an open subset of $\mathbb{R}^a$ and $\phi : A \to \mathbb{R}^b$ have a Lipschitz-constant c_ϕ. A bounded indexed family $\{\Lambda^\varepsilon \phi(v) | \varepsilon > 0, v \in A\}$ of closed subsets of $\mathcal{L}(\mathbb{R}^a, \mathbb{R}^b)$, also referred to as $\Lambda^\varepsilon \phi$, is a *derivate container* for ϕ if

$$\Lambda^\varepsilon \phi(v) \subset \Lambda^{\varepsilon'} \phi(v) \qquad (\varepsilon < \varepsilon', v \in A)$$

and for every compact subset A^* of A there exists a sequence (ϕ_i) of C^1 functions defined in a neighborhood of A^* and such that $\lim_i \phi_i = \phi$ uniformly on A^* and for every $\varepsilon > 0$ there exist $i(\varepsilon, A^*)$ and $\delta(\varepsilon, A^*) > 0$ such that

$$\phi_i'(v) \in \Lambda^\varepsilon \phi(w) \qquad (i \geqq i(\varepsilon, A^*), \quad w \in A^*, |v - w| \leqq \delta(\varepsilon, A^*)).$$

[2] While elements of $\mathcal{S}^{\#}_G$ yield the same set of solutions of controlled ordinary differential equations as $\mathcal{S}^{\#}$, this is no longer the case for various integral and functional-integral equations. However, it follows from Theorem IV.3.14 and the arguments of VII.1.4 and VIII.1.3 of [7] that, for functional-integral equations, $\mathcal{S}^{\#}$ can be replaced by $\mathcal{S}^{\#}_F$, the set of all $\sigma \in \mathcal{S}^{\#}$ for which there exists $n'(\sigma) \in \{1, 2, \cdots\}$ such that $\sigma(t)$ is, for all $t \in T$, concentrated at $n'(\sigma) + 1$ or fewer points of $R^{\#}(t)$. Our present results and arguments remain valid, without any change whatsoever, when $\mathcal{S}^{\#}_G$ is replaced by $\mathcal{S}^{\#}_F$. Thus we are hopeful that our present methods may be useful in the study of control problems defined by functional-integral equations.

We shall say that a bounded indexed family

$$\{\Lambda^\varepsilon f(t, v, r) | \varepsilon > 0, (t, v, r) \in T \times V \times R\}$$

of closed subsets of $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^n)$, also referred to as $\Lambda^\varepsilon f$, is a *derivate container for* f (*with respect to* v) if

$$\Lambda^\varepsilon f(t, v, r) \subset \Lambda^{\varepsilon'} f(t, v, r) \quad \text{for all } \varepsilon' > \varepsilon \text{ and all } t, v, r,$$

and for every compact subset V^* of V there exist a neighborhood $\tilde{V}$ of V^* in V and a sequence of functions $f^i : T \times \tilde{V} \times R \to \mathbb{R}^n$ such that each f^i has a partial derivative f^i_v, both f^i and f^i_v are measurable in t and continuous in (v, r), $\lim_i f^i = f$ uniformly on $T \times V^* \times R$, and for every $\varepsilon > 0$ there exist $i(\varepsilon, V^*)$ and $\delta(\varepsilon, V^*) > 0$ such that

$$f^i_v(t, v, r) \in \Lambda^\varepsilon f(t, w, r) \qquad (i \geqq i(\varepsilon, V^*), (t, w, r) \in T \times V^* \times R, |v - w| \leqq \delta(\varepsilon, V^*)).$$

This definition generalizes a concept introduced in [8]. The argument used there shows that a particular derivate container for $\phi = (\phi^1, \ldots, \phi^b)$ can be constructed in one of the following ways: we set

$$B^\varepsilon_{ik}(v) = \{(2\alpha)^{-1}[\phi^i(x + \alpha e_k) - \phi^i(x - \alpha e_k)] \,|\, |x - v| \leqq \varepsilon, 0 < \alpha \leqq \varepsilon, x, x \pm \alpha e_k \in V\},$$

and define $\Delta^\varepsilon \phi(v)$ as the collection of all $M = (M_{ik}) \in \mathscr{L}(\mathbb{R}^a, \mathbb{R}^b)$ with $M_{ik} \in \overline{\text{co}} B^\varepsilon_{ik}(v)$. Then $\Delta^\varepsilon \phi$ is a derivate container for ϕ. Furthermore, if $\phi = \tilde{\phi}_1 \circ \cdots \circ \tilde{\phi}_l$, with each $\tilde{\phi}_j$ defined on an open subset of some $\mathbb{R}^{b_j}$, and $\Lambda^\varepsilon \tilde{\phi}_j$ is a derivate container for $\tilde{\phi}_j$, then

$$\Lambda^\varepsilon \phi(v) = \{M_1 \cdot M_2 \cdot \cdots \cdot M_l | M_j \in \Lambda^\varepsilon \tilde{\phi}_j(\tilde{\phi}_{j+1} \circ \cdots \circ \tilde{\phi}_l(v))\}$$

also defines a derivate container for ϕ. An analogous procedure permits one to define a derivate container for f with respect to v. It also follows from the definition that if ϕ' exists and is continuous, then we may define $\Lambda^\varepsilon \phi(v)$ as $\{\phi'(w) | |w - v| \leqq \varepsilon/2\}$ for all ε and v.

We now generalize the concept of an extremal control usually defined under assumptions of differentiability.

DEFINITION 2.2 (extremal control). Let $(\bar{\sigma}, \bar{a}) \in \mathscr{S}^\#_G \times A_0$ be such that $\bar{y} = y(f, \bar{\sigma}, \bar{a})$ exists, $\bar{\sigma} = [\alpha^j, \rho_j]$, $\Lambda^\varepsilon f$ and $\Lambda^\varepsilon h^1$ be derivate containers for f and h^1 with respect to v, and

$$\Omega = (f, \Lambda^\varepsilon f, h^1, \Lambda^\varepsilon h^1, h^2, A_0).$$

We say that $(\bar{\sigma}, \bar{a})$ is *extremal relative to* Ω if

$$h^2(t, \bar{y}(t)) \in (-\infty, 0]^{m_2} \qquad (t \in T^h)$$

and there exist $l_1 \in \mathbb{R}^m$, nonnegative Radon measures $\omega_1, \cdots, \omega_{m_2}$ on T^h, a measurable $F : T \to \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n)$, and $H \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ such that

$$|l_1| + \sum_{j=1}^{m_2} \omega_j(T^h) > 0, \tag{1}$$

$$\omega_j(E_j) = 0, \text{ where } E_j = \{t \in T^h | e_j \cdot h^2(t, \bar{y}(t)) < 0\}, \tag{2}$$

$$H \in \bigcap_{\varepsilon > 0} \Lambda^\varepsilon h^1(\bar{y}(t_1)), \tag{3}$$

$$F(t) \in \bigcap_{\varepsilon>0} co \bigcup_{j=0}^{n} \Lambda^{\varepsilon} f(t, \bar{y}(t), \rho_j(t)) \quad \text{a.e. in } T, \tag{4}$$

$$k(t)^T f(t, \bar{y}(t), \bar{\sigma}(t)) = \min_{r \in R^{\#}(t)} k(t)^T f(t, \bar{y}(t), r) \quad \text{a.e. in } T, \tag{5}$$

$$k(t_0)^T \bar{a} = \min_{a_0 \in A_0} k(t_0)^T a_0, \tag{6}$$

where

$$k(t)^T = \left[l_1^T H + \sum_{j=1}^{m_2} \int_{[t,t_1] \cap T^h} e_j^T h_v^2(\tau, \bar{y}(\tau)) Z(\tau)^{-1} \omega_j(d\tau) \right] Z(t)$$

and

$$Z(t) = I + \int_t^{t_1} Z(\tau) F(\tau)\, d\tau \qquad (t \in T).$$

If this is the case, we also say that $(\bar{\sigma}, \bar{a}, l_1, \omega_j, F, H)$ is *extremal relative to* Ω.

If $\bar{\sigma} \in \mathcal{S}_{G'}^{\#}$, $\bar{y} = y(f, \bar{\sigma}, \bar{a})$ exists, and

$$h^2(t, \bar{y}(t)) \in (-\infty, 0]^{m_2} \qquad (t \in T^h)$$

but $(\bar{\sigma}, \bar{a})$ is not extremal relative to Ω, then we call $(\bar{\sigma}, \bar{a})$ *nonextremal relative to* Ω.

We can now state our basic results.

THEOREM 2.3. *Let* $(\sigma_0, \bar{a}_0)$ *be nonextremal relative to* $\Omega = (f, \Lambda^{\varepsilon} f, h^1, \Lambda^{\varepsilon} h^1, h^2, A_0)$. *Then there exist a finite collection* $\{u_1, \cdots, u_N\} \subset \mathcal{R}^{\#}$ *and* $\kappa > 0$ *such that*

$$\begin{aligned} S^F(h^1(y(f, \sigma_0, \bar{a}_0)(t_1)), \kappa) \subset \{ & h^1(y(f, u, a_0)(t_1)) | u \in \mathcal{R}^{\#}, a_0 \in A_0, \\ & h^2(t, y(f, u, a_0)(t)) \in (-\infty, -\kappa]^{m_2} (t \in T^h), \\ & u(t) \in \{u_1(t), \cdots, u_N(t)\} (t \in T)\}. \end{aligned}$$

Furthermore, there exists a sequence $((\hat{u}^i, a_0^i))$ *in* $\mathcal{R}^{\#} \times A_0$ *such that*

$$\hat{u}^i(t) \in \{u_1(t), \cdots, u_N(t)\} \qquad (i = 1, 2, \cdots, t \in T),$$

$$\lim_i \hat{u}^i = \sigma_0, \qquad \lim_i a_0^i = \bar{a}_0,$$

$$h^1(y(f, \hat{u}^i, a_0^i)(t_1)) = h^1(y(f, \sigma_0, \bar{a}_0)(t_1)) \qquad (i = 1, 2, \cdots),$$

$$h^2(t, y(f, \hat{u}^i, a_0^i)(t)) \in (-\infty, 0)^{m_2} \qquad (t \in T^h, \quad i = 1, 2, \cdots).$$

We shall refer to $(\bar{\sigma}, \bar{a}) \in \mathcal{S}_G^{\#} \times A_0$ as a *minimizing relaxed solution* if $(\bar{\sigma}, \bar{a})$ minimizes $h^0(y(f, \sigma, a_0)(t_1))$ on the set

$$\begin{aligned} \mathcal{A}(\mathcal{S}_G^{\#}) = \{ & (\sigma, a_0) \in \mathcal{S}_G^{\#} \times A_0 | h^1(y(f, \sigma, a_0)(t_1)) \in A_1, \\ & h^2(t, y(f, \sigma, a_0)(t)) \in (-\infty, 0]^{m_2} (t \in T^h)\}. \end{aligned}$$

We similarly define a *minimizing original solution* $(\bar{\sigma}, \bar{a}) \in \mathcal{R}^{\#} \times A_0$, with $\mathcal{R}^{\#}$ replacing $\mathcal{S}^{\#}_G$. We refer to a minimizing original solution as a *strict original solution* if it is not a minimizing relaxed solution.

THEOREM 2.4. *Let $(\sigma_0, \bar{a}_0)$ be either a minimizing relaxed solution or a minimizing original solution, and let $\Lambda^{\varepsilon} f$ and $\Lambda^{\varepsilon}(h^0, h^1)$ be derivate containers with respect to v for f and (h^0, h^1). Then there exist (l_0, l_1), ω_j, F and H such that $(\sigma_0, \bar{a}_0, (l_0, l_1), \omega_j, F, H)$ is extremal relative to*

$$\Omega^{0,1} = (f, \Lambda^{\varepsilon} f, (h^0, h^1), \Lambda^{\varepsilon}(h^0, h^1), h^2, A_0)$$

and

$$l_0 \geqq 0, \qquad l_1^T h^1(y(f, \sigma_0, \bar{a}_0)(t_1)) = \max_{a_1 \in A_1} l_1^T a_1.$$

THEOREM 2.5. *Let $\Omega^{0,1}$ be defined as in Theorem 2.4 and $(\bar{u}, \bar{a}_0)$ be a strict original solution. Then the set*

$$\mathcal{M}^- = \{(\sigma, a_0) \in \mathcal{S}^{\#}_G \times A_0 | h^0(y(f, \sigma, a_0)(t_1)) < h^0(y(f, \bar{u}, \bar{a}_0)(t_1)),$$
$$h^1(y(f, \sigma, a_0)(t_1)) \in A_1, h^2(t, y(f, \sigma, a_0)(t)) \in (-\infty, 0]^{m_2} (t \in T^h)\}$$

is nonempty and for every $(\tilde{\sigma}, \tilde{a}_0) \in \mathcal{M}^-$ there exist (l_0, l_1), ω_j, F and H such that $(\tilde{\sigma}, \tilde{a}_0, (l_0, l_1), \omega_j, F, H)$ is extremal relative to $\Omega^{0,1}$ and

$$l_0 = 0, \qquad l_1^T h^1(y(f, \tilde{\sigma}, \tilde{a}_0)(t_1)) = \max_{a_1 \in A_1} l_1^T a_1.$$

COROLLARY. *Let $(\tilde{\sigma}, \tilde{a}_0)$ be a minimizing relaxed solution. If $(\tilde{\sigma}, \tilde{a}_0, (l_0, l_1), \omega_j, F, H)$ can be extremal with respect to $\Omega^{0,1}$ only for $l_0 \neq 0$ (in which case we refer to the problem as "normal"), then there exists no strict original solution.*

3. Some auxiliary lemmas.

LEMMA 3.1. *Let A be a convex subset of $\mathbb{R}^n$, $\tilde{A}$ an open neighborhood of A and $\phi : \tilde{A} \to \mathbb{R}^m$ continuously differentiable. Then*

$$\phi(v) - \phi(w) \in \operatorname{co}\{\phi'(w + t(v - w))(v - w) | 0 \leqq t \leqq 1\} \qquad (v, w \in A).$$

Furthermore, there exist points $t_1, \cdots, t_m \in [0, 1]$ and numbers $\alpha_1, \cdots, \alpha_m$ such that $\alpha_j \geqq 0$, $\sum_{j=1}^m \alpha_j = 1$ and

$$\phi(v) - \phi(w) = \sum_{j=1}^{m} \alpha_j \phi'(w + t_j(v - w))(v - w).$$

Proof. Let $v, w \in A$ and

$$\psi(t) = \phi(w + t(v - w)) \qquad (0 \leqq t \leqq 1).$$

Then ψ is continuously differentiable and

$$\psi'(t) = \phi'(w + t(v - w))(v - w).$$

It follows that

$$\psi(1) - \psi(0) = \int_0^1 \phi'(w + t(v - w))(v - w)\, dt;$$

hence

$$\phi(v)-\phi(w)\in co\,\{\phi'(w+t(v-w))(v-w)|0\leqq t\leqq 1\}.$$

The second statement now follows by applying Caratheodory's theorem about the representation of the convex hull of a connected set. Q.E.D.

LEMMA 3.2. *Let A be a closed convex subset of $\mathbb{R}^n$, $\tilde{A}$ an open neighborhood of A, $0\in A$, $\phi:\tilde{A}\to\mathbb{R}^n$ have a continuous derivative, and $|\phi'(v)^{-1}|\leqq c$ $(v\in A)$. Let*

$$\hat{A}=\{\gamma v|v\in A,\ \gamma\geqq 0\},$$

$$0\leqq\alpha\leqq\sup\{\gamma\geqq 0|\gamma v\in A(v\in\hat{A},|v|=1)\}.$$

If $a\in\phi'(v)\hat{A}$ for every $v\in A$ and $|a|=1$, then

$$\phi(0)+[0,\alpha/c]a\subset\phi(A\cap S^F(0,\alpha)).$$

If furthermore, v, $w\in A$ and $|M^{-1}|\leqq c_1$ for every

$$M\in F(v,w)=\text{co}\,\{\phi'(\theta v+(1-\theta)w)|0\leqq\theta\leqq 1\},$$

then

$$|\phi(v)-\phi(w)|\geqq c_1^{-1}|v-w|.$$

Proof. If $\alpha=0$, then our first statement is trivially satisfied. We assume therefore that $\alpha>0$.

For each $x\in\mathbb{R}^n$, let $s(x)$ denote the unique point in A that minimizes the Euclidean distance $|x-s(x)|_2$ to x. Then $|s(x)-s(y)|_2\leqq|x-y|_2$ for all x, $y\in\mathbb{R}^n$, and the function $\psi:\mathbb{R}^n\to\mathscr{L}(\mathbb{R}^n,\mathbb{R}^n)$, defined by

$$\psi(x)=\phi'(s(x)),$$

is continuous. We now consider the differential equation

$$\dot{u}(t)=\psi(u(t))^{-1}a\qquad(t\geqq 0),\quad u(0)=0.$$

Since $|\psi(x)^{-1}|\leqq c$ for all $x\in\mathbb{R}^n$ and ψ is continuous, this equation has a continuously differentiable solution u for all $t\geqq 0$. Since, for each $v\in A$, $\phi'(v)^{-1}a\in\hat{A}$, we have

$$\dot{u}(t)\in\hat{A}\quad\text{and}\quad|\dot{u}(t)|\leqq c\qquad(t\geqq 0).$$

It follows that

$$u(t)=\int_0^t\dot{u}(\tau)\,d\tau\in\hat{A}$$

and

$$|u(t)|\leqq\int_0^t|\dot{u}(\tau)|\,d\tau\leqq ct.$$

Thus, for $t\in[0,\alpha/c]$, we have $u(t)\in A$ and

$$\phi(u(t))-\phi(0)=\int_0^t\phi'(u(\tau))\dot{u}(\tau)\,d\tau$$

$$=\int_0^t\psi(u(\tau))\dot{u}(\tau)\,d\tau=ta;$$

hence

$$\phi(0)+[0, \alpha/c]a \subset \phi(A \cap S^F(0, \alpha)).$$

Now assume that $|M^{-1}| \leqq c_1$ for every $M \in F(v, w)$. By Lemma 3.1,

$$\phi(v)-\phi(w) = \bar{M}(v-w)$$

for some $\bar{M} \in F(v, w)$, and therefore

$$|v-w| = |\bar{M}^{-1}[\phi(v)-\phi(w)]| \leqq c_1|\phi(v)-\phi(w)|. \qquad \text{Q.E.D.}$$

LEMMA 3.3. *Let A be a convex body in $\mathbb{R}^n$ (that is, a closed convex set in $\mathbb{R}^n$ with $A^\circ \neq \varnothing$), $0 \in A$,*

$$\hat{A} = \{\gamma v | v \in A, \gamma \geqq 0\},$$

$$0 \leqq \alpha \leqq \sup\{\gamma \geqq 0 | \gamma v \in A (v \in \hat{A}, |v| = 1)\},$$

and $\phi : A \to \mathbb{R}^n$ Lipschitz-continuous. If $\Lambda^\varepsilon \phi$ is a derivative container for $\phi|A^\circ$, $\varepsilon_0 > 0$, $a \in \mathbb{R}^n$, $|a| = 1$ and if, for every $v \in A^\circ$ and every $M \in \Lambda^{\varepsilon_0}\phi(v)$, we have

$$|M^{-1}| \leqq c \quad \text{and} \quad a \in M\hat{A},$$

then

$$\phi(0)+[0, \alpha/c]a \subset \phi(A \cap S^F(0, \alpha)).$$

If $v, w \in A^\circ$ and $|M^{-1}| \leqq c_1$ for every

$$M \in F(v, w) = \operatorname{co} \bigcup_{0 \leqq \theta \leqq 1} \Lambda^{\varepsilon_0}\phi(\theta v + (1-\theta)w),$$

then

$$\phi(v)-\phi(w) \geqq c_1^{-1}|v-w|.$$

Proof. We shall assume that $\alpha > 0$, the first statement being trivially true if $\alpha = 0$. Let p be an interior point of $\hat{A}$ with $|p| = 1$, $0 < \eta < \frac{1}{2}\alpha$, and

$$A_\eta = \eta p + A \cap S^F(0, \alpha - 2\eta).$$

Then A_η is a convex body and $A_\eta \subset A^\circ$.

Now let (ϕ_j) be a sequence of C^1 uniform approximations to $\phi|A_\eta$ associated with $\Lambda^\varepsilon\phi$. We can determine a positive integer j_0 such that

$$\phi_j'(v) \in \Lambda^{\varepsilon_0}\phi(v) \qquad (v \in A_\eta, \quad j \geqq j_0).$$

We set

$$\psi_j(v) = \phi_j(\eta p + v) \qquad (v \in A_\eta - \eta p, \quad j \geqq j_0).$$

Then the conditions of Lemma 3.2 are satisfied, with ϕ, A, α replaced by ψ_j, $A_\eta - \eta p$, $\alpha - 2\eta$, respectively. It follows that

$$\psi_j(0)+[0, (\alpha-2\eta)/c]a \subset \psi_j(A_\eta - \eta p);$$

hence

$$\phi_j(\eta p)+[0, (\alpha-2\eta)/c]a \subset \phi_j(A_\eta).$$

Since $\lim_j \phi_j = \phi$ uniformly on A_η, A_η is compact, and ϕ and ϕ_j continuous, we conclude that

$$\phi(\eta p)+[0,(\alpha-2\eta)/c]a \subset \phi(A_\eta) \subset \phi(A \cap S^F(0,\alpha))$$

for all sufficiently small positive η. This implies that

$$\phi(0)+[0,\alpha/c]a \subset \phi(A \cap S^F(0,\alpha)).$$

Now assume that $v, w \in A^\circ$ and $|M^{-1}| \leqq c_1$ for all $M \in F(v,w)$. We denote by A^* the closed segment joining v and w and determine a sequence (ϕ_j) of C^1 uniform approximations to $\phi|A^*$ associated with $\Lambda^\varepsilon\phi$. Then $\phi_j'(\theta v+(1-\theta)w) \in F(v,w)$ for large j and, by Lemma 3.2,

$$\phi_j(v)-\phi_j(w) \geqq c_1^{-1}|v-w|.$$

We obtain our final conclusion by letting $j \to \infty$. Q.E.D.

We shall henceforth denote by χ_A the characteristic function of A.

LEMMA 3.4. *Let $\phi \in L^\infty(\mu, \mathbb{R}^m)$ and $\Phi(s)=\int_{t_0}^s \phi(\tau)\,d\tau$ $(s \in T)$. Then, for every $\varepsilon>0$ and every subset $\mathbf{N}$ of T with $\mu(\mathbf{N})=0$, there exist $l \in \{1,2,\cdots\}$, $t^1,\cdots,t^l \in T\sim\mathbf{N}$ and $a^1,\cdots,a^l \in (0,\varepsilon]$ such that the points $t^1,\cdots,t^l$ are distinct, and*

$$\sum_{k=1}^{l} a^k = t_1-t_0, \qquad \left|\Phi(s)-\sum_{k=1}^{l} a^k\phi(t^k)\chi_{[t_0,s]}(t^k)\right| \leqq \varepsilon \qquad (s \in T).$$

Proof. We set $\varepsilon' = \min(\varepsilon, \frac{1}{2}(t_1-t_0))$, choose a closed subset T_ε of $T\sim\mathbf{N}$ such that $\mu(T\sim T_\varepsilon) \leqq \varepsilon'[4|\phi|_\infty]^{-1}$ and $\phi|T_\varepsilon$ is continuous, and let $\delta>0$ be such that $\delta \leqq \min([8|\phi|_\infty]^{-1}\varepsilon', \frac{1}{2}\varepsilon')$ and

$$|\phi(t')-\phi(t'')|<\varepsilon'[4(t_1-t_0)]^{-1}$$

if $|t'-t''| \leqq \delta$, $t', t'' \in T_\varepsilon$. We partition T into consecutive subintervals $\tilde{I}_1, \tilde{I}_2, \cdots, \tilde{I}_l$ of lengths not exceeding δ, set $I_k = \tilde{I}_k \cap T_\varepsilon$, $\beta^k = \mu(I_k)$, choose in each nonempty I_k a point t^k, and set $t^k = t_0$ if $I_k = \varnothing$. We denote by $O(a)$ an element x of $\mathbb{R}^m$ with $|x| \leqq a$. For each $s \in \tilde{I}_j$, we have

$$\Phi(s) = \sum_{k=1}^{j-1} \beta^k\phi(t^k) + O\left(\frac{\varepsilon}{4}\right) + O(\delta|\phi|_\infty) + O\left(\frac{\varepsilon}{4}\right) \tag{1}$$

$$= \sum_{k=1}^{j-1} \beta^k\phi(t^k) + O\left(\frac{5}{8}\varepsilon\right).$$

Now

$$\sum_{k=1}^{j-1} \beta^k\phi(t^k) = \sum_{k=1}^{l} \beta^k\phi(t^k)\chi_{[t_0,s]}(t^k) + O(\delta|\phi|_\infty). \tag{2}$$

We set $a^k = (t_1-t_0)\beta^k/\sum_{i=1}^l \beta^i$ and observe that

$$(t_1-t_0) \geqq \sum_{i=1}^{l} \beta^i \geqq (t_1-t_0)-[4|\phi|_\infty]^{-1}\varepsilon';$$

hence

$$a^k - \varepsilon'[4|\phi|_\infty(t_1-t_0)]^{-1}a^k \leqq \beta^k \leqq a^k.$$

Then, by (1) and (2), we have

$$\Phi(s)=\sum_{k=1}^{l} a^k\phi(t^k)\chi_{[t_0,s]}(t^k)+O\left(\frac{5\varepsilon}{8}\right)+O\left(\frac{\varepsilon}{4}\right)+O(\delta|\phi|_\infty).$$

Since

$$\beta^k\leqq\mu(\tilde{I}_k)\leqq\delta\leqq\tfrac{1}{2}\varepsilon$$

and

$$\sum_{i=1}^{l}\beta^i=\mu(T_\varepsilon)\geqq(t_1-t_0)-\varepsilon'\geqq\tfrac{1}{2}(t_1-t_0),$$

we conclude that $a^k\in[0,\varepsilon]$. The points t^k corresponding to $a^k\neq 0$ are distinct. Q.E.D.

LEMMA 3.5. *Let* $\sigma_0=[\alpha^j,\rho_j]\in\mathscr{S}^{\#}_G$. *Then there exist sequences* $(\{A^i_0,\cdots,A^i_n\})_{i=1}^{\infty}$ *of measurable partitions of* T *and* (u_i) *in* $\mathscr{R}^{\#}$ *such that*

(1) $\lim_i\mu(A^i_j\cap B)=\int_B\alpha^j(t)\,dt$ $(j=0,\cdots,n,\ B$ *measurable*),

(2) $u_i(t)=\rho_j(t)\qquad(t\in A^i_j,\quad j=0,\cdots,n)$

and

(3) $$\lim_i u_i=\sigma_0.$$

Proof. Let

$$d(t)=\tfrac{1}{4}\sup\{\alpha\leqq\text{diameter }(R)\mid d(\rho_j(t),\rho_k(t))\geqq\alpha(k\neq j,\ \alpha^j(t)\neq 0,\ \alpha^k(t)\neq 0)\},$$

$$R^{\#}_1(t)=\{\rho_j(t)|\alpha^j(t)\neq 0,\ j=0,\cdots,n\}\qquad(t\in T),$$

$\mathscr{R}^{\#}_1$ be the collection of measurable selections of $R^{\#}_1$ and $\mathscr{S}^{\#}_1$ the collection of measurable functions $\sigma: T\to\text{rpm}(R)$ with $\sigma(t)(R^{\#}_1(t))=1$ $(t\in T)$. Then $\sigma_0\in\mathscr{S}^{\#}_1$ and, by [7, IV.3.10, p. 287], there exists a sequence (u_i) in $\mathscr{R}^{\#}_1$ such that $\lim_i u_i=\sigma_0$, that is,

(4) $$\lim_i\int_{t_0}^{t_1}\phi(t,u_i(t))\,dt=\int_{t_0}^{t_1}dt\int\phi(t,r)\sigma_0(t)(dr)\qquad[\phi\in L^1(\mu,C(R))].$$

Since $u_i\in\mathscr{R}^{\#}_1$, there exist measurable partitions $\{A^i_0,\cdots,A^i_n\}$ of T such that

$$u_i(t)=\rho_j(t)\qquad(t\in A^i_j,\ j=0,\cdots,n)$$

and

$$A^i_j\subset\{t\in T|\alpha^j(t)\neq 0\}.$$

For each $j=0,1,\cdots,n$ and $t\in T$ such that $\alpha^j(t)\neq 0$, let

$$F_j(t)=S^F(\rho_j(t),d(t)),$$

$$H_j(t)=\{r\in R|d(r,\rho_j(t))\geqq 2d(t)\},$$

$$\psi_j(t,r)=(d[r,H_j(t)]+d[r,F_j(t)])^{-1}\,d[r,H_j(t)].$$

If $\alpha^j(t)=0$, we set $\psi_j(t,r)=0$ $(r\in R)$.

If B is a measurable subset of T and $\phi_j(t, r) = \psi_j(t, r)\chi_B(t)$, then, for every $i = 1, 2, \cdots$,

$$\int_{t_0}^{t_1} \phi_j(t, u_i(t))\, dt = \mu(A_j^i \cap B)$$

and

$$\int_{t_0}^{t_1} dt \int \phi_j(t, r)\sigma_0(t)(dr) = \int_B \alpha^j(t)\, dt.$$

Thus relation (1) follows from (4). Q.E.D.

4. Proof of Theorem 2.3. We shall first show that it suffices to prove the theorem for the special case where $A_0 = \{\bar{a}_0\}$. Indeed, let A_0^* be a compact and convex neighborhood of $\bar{a}_0$ in A_0. We now consider a related problem in which relations (1)–(3) of § 1 are replaced by

$$\dot{y}(t) = a_0(t) - \bar{a}_0 \quad \text{a.e. in } [t_0 - 1, t_0),$$

$$\dot{y}(t) = f(t, y(t), u(t)) \quad \text{a.e. in } [t_0, t_1], \tag{1'}$$

$$(u(t), a_0(t)) \in R \times A_0^* \quad \text{a.e. in } [t_0 - 1, t_0),$$

$$(u(t), a_0(t)) \in R^{\#}(t) \times A_0^* \quad \text{a.e. in } [t_0, t_1], \tag{2'}$$

$$y(t_0 - 1) = \bar{a}_0. \tag{3'}$$

Then $(\sigma_0, \bar{a}_0)$ is extremal for the old problem if and only if $((\sigma_0, \bar{a}_0), \bar{a}_0)$ is extremal for the new problem, where $\sigma_0(t) = \sigma_0(t_0)$ for $t < t_0$ and each derivate container is defined as before for $t \in T$ and as containing only an appropriate zero matrix for $t < t_0$. Then l_1, ω_j, $\bar{y}(t)$, $Z(t)$ and $k(t)$ remain the same for the new problem when $t \in T$, while $Z(t) = Z(t_0)$ and $k(t) = k(t_0)$ when $t < t_0$. Thus Definition 2.2(5) for $t \leqq t_0$ and the new problem is equivalent to Definition 2.2(6).

Thus Theorem 2.3 remains valid for arbitrary closed convex sets $A_0 \subset \mathbb{R}^n$ if it is valid for sets A_0 consisting of one element only. For this reason, we shall assume, in the remainder of this section, that $A_0 = \{\bar{a}_0\}$, and shall write $y(g, \sigma)$ for $y(g, \sigma, \bar{a}_0)$. We shall use the terms "extremal respectively nonextremal" to mean "extremal respectively nonextremal relative to Ω."

Let $\sigma_0 = [\alpha^j, \rho_j]$, V^* be a compact subset of V containing $y(f, \sigma_0)(T)$ in its interior, and let $(f^p)_{p=1}^{\infty}$, $(h^{1,p})_{p=1}^{\infty}$ represent the uniform approximations to $f | T \times V^* \times R$ respectively $h^1 | V^*$ associated with the definition of $\Lambda^e f$ and $\Lambda^e h^1$. It follows easily from Gronwall's inequality and [7, Chap. VI, pp. 346 ff.] that

$$\lim_p y(f^p, \sigma)(t) = y(f, \sigma)(t)$$

uniformly for σ near σ_0 and $t \in T$, and that the function $\sigma \to y(g, \sigma)$ is continuous near σ_0 for large p and $g = f, f^p$. We may therefore assume that $y(f^p, \sigma)(T) \subset V^*$ for $p = 1, 2, \cdots$ provided $|\sigma - \sigma_0|_w$ (the distance in $\mathcal{S}^{\#}$) is sufficiently small.

We set

$$\mathscr{H}^\varepsilon = \Lambda^\varepsilon h^1(y(f, \sigma_0)(t_1))$$

and denote by $\mathscr{F}^\varepsilon$ the collection of all measurable $F : T \to \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n)$ such that

$$F(t) \in \operatorname{co} \bigcup_{j=0}^{n} \Lambda^\varepsilon f(t, y(f, \sigma_0)(t), \rho_j(t)) \quad \text{a.e. in } T.$$

For every $F \in L^\infty(\mu, \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n))$, the matrix-differential equation

$$Z(t) = I + \int_t^{t_1} Z(\tau)F(\tau)\, d\tau \qquad (t \in T)$$

has a unique solution $Z(F)$, and the matrices $Z(F)(t)$ and $Z(F)(t)^{-1}$ are uniformly bounded for $t \in T$ and $|F|_\infty$ uniformly bounded. We shall denote by c a common bound of

$$1, f(t, v, r), f^p(t, v, r), h^1(v), h^2(t, v), h_v^2(t, v), Z(F)(t), Z(F)(t)^{-1}$$

and all elements of $\Lambda^\varepsilon f(t, v, r)$ and $\Lambda^\varepsilon h^1(v)$ for

$$(t, v, r) \in T \times V^* \times R,$$

$$\varepsilon > 0 \quad \text{and} \quad |F|_\infty \leqq \sup\{|M| \,|M \in \Lambda^\varepsilon f(t, v, r)\},$$

and a common Lipschitz-constant for $f(t, \cdot, r)$, $f^p(t, \cdot, r)$, h^1 and $h^2(t, \cdot)$ over $T \times V^* \times R$.

We shall write δ_r for the Dirac measure at $r \in R$,

$$r' \sim r'' = \delta_{r'} - \delta_{r''}, \qquad r' \sim \sigma(t) = \delta_{r'} - \sigma(t),$$

and $|\cdot|_{\sup}$ for the sup norm. By [7, IV.3.10, p. 287], $\mathscr{R}^\#$ contains a denumerable subset $\mathscr{R}_\infty^\#$ that is dense in the compact metric space $\mathscr{S}^\#$.

LEMMA 4.1. *Let $\mathscr{R}_\infty^\#$ be as defined above. Then there exists $\gamma \in (0, 1]$ such that, for every choice of a set $\mathbf{N} \subset T$ with $\mu(\mathbf{N}) = 0$ and of*

$$e > 0, \quad F \in \mathscr{F}^\gamma, \quad H \in \mathscr{H}^\gamma \quad \text{and} \quad \xi^i \in S^F(0, \gamma) \subset \mathbb{R}^m \qquad (i = 1, \cdots, m)$$

we can determine

$$l \in \{1, 2, \cdots\}, \quad t^{ik} \in T \sim \mathbf{N}, \quad a^{ik} \in [0, 1], \quad \rho^i \in \mathscr{R}_\infty^\#$$
$$(i = 1, \cdots, m, \quad k = 1, \cdots, l)$$

and corresponding

$$\delta^i(t) = Z(F)(t)^{-1} \sum_{k=1}^{l} a^{ik} Z(F)(t^{ik}) \cdot f(t^{ik}, y(f, \sigma_0)(t^{ik}), \rho^i(t^{ik}) \sim \sigma_0(t^{ik})) \chi_{[t_0, t]}(t^{ik}),$$

such that the points t^{ik} are all distinct,

$$a^{ik} \leqq (32c^2 \exp[c(t_1 - t_0)])^{-1}\gamma, \qquad \sum_{k=1}^{l} a^{ik} = t_1 - t_0,$$

$$|\xi^i - H\delta^i(t_1)| \leqq e$$

and

$$h^2(t, y(f, \sigma_0)(t)) + h_v^2(t, y(f, \sigma_0)(t))\delta^i(t) \in (-\infty, -\gamma]^{m_2} \qquad (t \in T^h, \quad i = 1, \cdots, m).$$

Proof. Step 1. For $F \in \bigcup_{\varepsilon>0} \mathscr{F}^\varepsilon$, $H \in \bigcup_{\varepsilon>0} \mathscr{H}^\varepsilon$, $\sigma \in \mathscr{S}^\#$ and $t \in T^h$, let

$$x_1(F, H, \sigma) = H \int_{t_0}^{t_1} Z(F)(\tau) f(\tau, y(f, \sigma_0)(\tau), \sigma(\tau) - \sigma_0(\tau))\, d\tau,$$

$$x_2(F, \sigma)(t) = h_v^2(t, y(f, \sigma_0)(t)) Z(F)(t)^{-1} \cdot \int_{t_0}^{t} Z(F)(\tau) f(\tau, y(f, \sigma_0)(\tau), \sigma(\tau) - \sigma_0(\tau))\, d\tau,$$

$$W(F, H) = \{(x_1(F, H, \sigma), x_2(F, \sigma)) | \sigma \in \mathscr{S}^\#\} \subset \mathbb{R}^m \times C(T^h, \mathbb{R}^{m_2}),$$

$$\hat{h}(t) = h^2(t, y(f, \sigma_0)(t)), \qquad \hat{h}_v(t) = h_v^2(t, y(f, \sigma_0)(t)).$$

For each F and H, the function

$$\sigma \to (x_1(F, H, \sigma), x_2(F, \sigma)) : \mathscr{S}^\# \to \mathbb{R}^m \times C(T^h, \mathbb{R}^{m_2})$$

is continuous. Since $\mathscr{S}^\#$ is convex and compact, it follows that $W(F, H)$ is also convex and compact, and we have

$$0 = (x_1(F, H, \sigma_0), x_2(F, \sigma_0)) \in W(F, H).$$

We shall show that there exists $\beta \in (0, 1]$ such that, for every $F \in \mathscr{F}^\beta$ and $H \in \mathscr{H}^\beta$, $W(F, H)$ *contains a point* (w_1, w_2) satisfying the relations

$$(1) \qquad w_1 = 0, \qquad \hat{h}(t) + w_2(t) \in (-\infty, -\beta]^{m_2} \qquad (t \in T^h)$$

Indeed, assume the contrary. Then there exists a sequence $((\beta_i, F_i, H_i))$, with β_i decreasing to 0, $F_i \in \mathscr{F}^{\beta_i}$ and $H_i \in \mathscr{H}^{\beta_i}$ such that, for each i, the closed convex set

$$\Phi_i = \{0\} \times \{\phi_2 \in C(T^h, \mathbb{R}^{m_2}) | \phi_2(t) + \hat{h}(t) \in (-\infty, -\beta_i]^{m_2} (t \in T^h)\}$$

has no points in common with the compact convex set $W(F_i, H_i)$. It follows that there exist $l^i = (l_1^i, l_2^i) \in \mathbb{R}^m \times C(T^h, \mathbb{R}^{m_2})^*$ such that

$$(2) \qquad |l^i| = 1 \quad \text{and} \quad l^i w \geqq l^i \phi \qquad [w \in W(F_i, H_i), \phi \in \Phi_i].$$

We can represent $l_2^i \in C(T^h, \mathbb{R}^{m_2})^*$ by Radon measures $\omega_1^i, \cdots, \omega_{m_2}^i$ on T^h. Since $0 \in W(F_i, H_i)$, (2) yields

$$\sum_{j=1}^{m_2} \int e_j \cdot \phi_2(t) w_j^i(dt) \leqq 0 \qquad [(0, \phi_2) \in \Phi_i].$$

This implies that each ω_j^i is nonnegative,

$$|l_2^i| = \sum_{j=1}^{m_2} \omega_j^i(T^h) \leqq 1$$

and

$$(3) \qquad 0 \leqq \omega_j^i(A_i) \leqq \sqrt{\beta_i}, \quad \text{where} \quad A_i = \{t \in T^h | e_j \cdot \hat{h}(t) < -\beta_i - \sqrt{\beta_i}\}.$$

Furthermore, setting $\phi_2 = -\beta_i(1, \cdots, 1)$ in (2), we obtain

$$l^i w \geqq -\beta_i \qquad [w \in W(F_i, H_i)]. \tag{4}$$

Since the families $\{Z(F) | F \in \mathscr{F}^{\beta_1}\}$ and $\mathscr{H}^{\beta_1}$ are conditionally compact in $C(T, \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n))$ and $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, respectively, we can find a sequence $J \subset (1, 2, \cdots)$ and $\bar{Z}$, $\bar{H}$, l_1 and ω_j such that

$$\lim_{i \in J} Z(F_i) = \bar{Z} \text{ uniformly}, \qquad \lim_{i \in J} H_i = \bar{H},$$

$$\lim_{i \in J} l_1^i = l_1, \qquad \lim_{i \in J} \omega_j^i = \omega_j \text{ weakly}.$$

Then, by (2),

$$|l_1| + \sum_{j=1}^{m_2} \omega_j(T^h) = 1.$$

We set

$$\hat{f}(\sigma)(\tau) = f(\tau, y(f, \sigma_0)(\tau), \sigma(\tau) - \sigma_0(\tau)).$$

Then, by (4),

$$\begin{aligned} l_1^{iT} H_i \int_{t_0}^{t_1} Z(F_i)(\tau) \hat{f}(\sigma)(\tau)\, d\tau + \sum_{j=1}^{m_2} \int e_j^T \hat{h}_v(t) Z(F_i)(t)^{-1} \omega_j^i(dt) \\ \cdot \int_{t_0}^{t} Z(F_i)(\tau) \hat{f}(\sigma)(\tau)\, d\tau \geqq -\beta_i \qquad (\sigma \in \mathscr{S}^{\#}). \end{aligned} \tag{5}$$

Since $\mathscr{F}^{\beta_1}$ is bounded, the sets

$$\operatorname{co} \bigcup_{j=0}^{n} \Lambda^{\beta_i} f(t, y(f, \sigma_0)(t), \rho_j(t))$$

are convex and compact, and

$$\frac{d}{dt} Z(F_i)(t) = -Z(F_i)(t) F_i(t) \quad \text{a.e. in } T,$$

it follows from standard arguments of optimal control, with F_i playing the role of control functions (or from a special case of [7, IV.3.14, p. 291]), that there exists $\bar{F} \in \bigcap_{\varepsilon > 0} \mathscr{F}^{\varepsilon}$ such that $\bar{Z} = Z(\bar{F})$. Thus relation (5) implies, letting $i \to \infty$, $i \in J$, that

$$\int_{t_0}^{t_1} k(\tau)^T f(\tau, y(f, \sigma_0)(\tau), \sigma(\tau) - \sigma_0(\tau))\, d\tau \geqq 0 \qquad (\sigma \in \mathscr{S}^{\#}),$$

where $k(\tau)$ is defined as in Definition 2.2 for $F = \bar{f}$, $H = \bar{H}$. We deduce from the above relation (as in [7, VI.2.3, step 2, pp. 360–361]) that

$$\begin{aligned} k(t)^T f(t, y(f, \sigma_0)(t), \sigma(t)) \\ = \min_{r \in R^{\#}(t)} k(t)^T f(t, y(f, \sigma_0)(t), r) \quad \text{a.e. in } T. \end{aligned} \tag{6}$$

Finally, since the ω_j^i are nonnegative, relation (3) implies that

$$\omega_j(\{t \in T^h | e_j \cdot \hat{h}(t) < 0\}) = 0 \qquad (j = 1, \cdots, m_2).$$

This shows that σ_0 is extremal, contrary to assumption.

Step 2. Let β be as defined in Step 1. We next show that there exists $\alpha > 0$ such that, for all $F \in \mathscr{F}^\beta$ and $H \in \mathscr{H}^\beta$, we have

$$S^F(0, \alpha) \subset W_1(F, H) = \{x_1(F, H, \sigma) | \sigma \in \mathscr{S}^\#\}.$$

Indeed, otherwise there exist sequences (F_i) in $\mathscr{F}^\beta$ and (H_i) in $\mathscr{H}^\beta$ such that each convex and compact set $W_1(F_i, H_i) \subset \mathbb{R}^m$ contains a boundary point w_i, with $\lim_i w_i = 0$. For each such boundary point w_i we can determine an inward normal l_1^i such that $|l_1^i| = 1$ and

$$l_1^i w \geqq l_1^i w_i \qquad [w \in W_1(F_i, H_i), \quad i = 1, 2, \cdots]. \tag{7}$$

As in Step 1, we select $J \subset (1, 2, \cdots)$ such that

$$\lim_{i \in J} l_1^i = l_1, \quad \lim_{i \in J} Z(F_i) = Z(\bar{F}), \quad \lim_{i \in J} H_i = \bar{H},$$

and it follows from (7) that relation (6) is satisfied with $k(t)^T = l_1^T \bar{H} Z(\bar{F})(t)$ (corresponding to $\omega_1 = \cdots = \omega_{m_2} = 0$). Again, this contradicts the assumption that σ_0 is nonextremal.

Step 3. Let α and β be defined as in Steps 1 and 2, $F \in \mathscr{F}^\beta$ and $H \in \mathscr{H}^\beta$. Then $S^F(0, \alpha) \subset W_1(F, H)$ and there exists a point $\bar{w} = (0, \bar{w}_2) \in W(F, H)$ such that

$$\bar{w}_2(t) + \hat{h}(t) \in (-\infty, -\beta]^{m_2} \qquad (t \in T^h).$$

The number $c' = c^4(t_1 - t_0) + c$ is an upper bound of c and all $|x_1(F, H, \sigma)|$ and $|x_2(F, \sigma)(t)|$. We set $\beta' = \frac{1}{2}\beta/(c' + \beta)$ and, for each $w = (w_1, w_2) \in W(F, H)$,

$$\tilde{w} = (\tilde{w}_1, \tilde{w}_2) = \beta' w + (1 - \beta')\bar{w} \in W(F, H).$$

Since $|w_2|_{\sup} \leqq c'$ and $|\hat{h}|_{\sup} \leqq c'$, we have

$$e_j \cdot \tilde{w}_2(t) \leqq \beta' c' + (1 - \beta')(-e_j \cdot \hat{h}(t) - \beta) \qquad (j = 1, \cdots, m_2, \quad t \in T^h);$$

hence

$$e_j \cdot [\tilde{w}_2(t) + \hat{h}(t)] \leqq -\tfrac{1}{2}\beta \qquad (j = 1, 2, \cdots, m_2, \quad t \in T^h).$$

Since

$$S^F(0, \beta'\alpha) \subset \beta' W_1(F, H) = \{\tilde{w}_1 | w \in W(F, H)\},$$

it follows that

$$S^F(0, \beta'\alpha) \subset \{w_1 | (w_1, w_2) \in W(F, H), \hat{h}(t) + w_2(t) \in (-\infty, -\tfrac{1}{2}\beta]^{m_2} (t \in T^h)\}.$$

We now set $\gamma = \min(1, \beta'\alpha, \frac{1}{8}\beta)$. Then the above relation implies that for every choice of $F \in \mathscr{F}^\beta$, $H \in \mathscr{H}^\beta$ and $\xi^1, \cdots, \xi^m \in S^F(0, \gamma)$, there exist $\sigma^1, \cdots, \sigma^m \in \mathscr{S}^\#$ such that

$$\xi^i = x_1(F, H, \sigma^i), \qquad \hat{h}(t) + x_2(F, \sigma^i)(t) \in (-\infty, -4\gamma]^{m_2} \qquad (t \in T^h).$$

Since $\mathscr{R}_\infty^\#$ is dense in $\mathscr{S}^\#$, for every $e>0$ we can determine $\rho^i \in \mathscr{R}_\infty^\# (i=1,\cdots,m)$ sufficiently close to σ^i so that

$$|\xi^i - x_1(F,H,\rho^i)| < e/2, \qquad \hat{h}(t)+x_2(F,\rho^i)(t) \in (-\infty,-2\gamma]^{m_2} \qquad (t\in T^h).$$

We then apply Lemma 3.4 to approximate $x_1(F,H,\rho^i)$ and $x_2(F,\rho^i)(t)$ by $H\delta^i(t_1)$ and $\hat{h}_v(t)\delta^i(t)$, respectively, choosing for each i points t^{ik} distinct from all t^{pq} for $p<i$. We may clearly assume that the number of points t^{ik} and a^{ik} is the same for $i=1,\cdots,m$, choosing for l the largest of these numbers, setting $a^{ik}=0$ for the "missing" values of k, and selecting corresponding t^{ik} arbitrarily but distinct from each other and the previously selected points. Q.E.D.

4.2. Auxiliary definitions. We shall continue to use the notation introduced in this section, and shall denote by $[a_1,\cdots,a_m]$ the matrix with columns $a_1,\cdots,a_m$.

I. It is easily seen that we can determine a finite collection $\mathscr{G}$ of m-simplices in $S^i(0,\gamma)\subset\mathbb{R}^m$ and a number

$$\varepsilon_1 \in (0,[80c^3(t_1-t_0)+1]^{-1}\gamma)$$

with the property that for every $w\in\mathbb{R}^m$ with $|w|=\gamma/(2m)$ there exists a corresponding $G=\operatorname{co}\{0,\xi^1,\cdots,\xi^m\}\in\mathscr{G}$ such that the ξ^i are mutually orthogonal and of norm γ, and

$$w\in\operatorname{co}\{0,\tilde{\xi}^1,\cdots,\tilde{\xi}^m\} \quad\text{and}\quad |[\tilde{\xi}^1,\cdots,\tilde{\xi}^m]^{-1}| \leqq 2m/\gamma$$

provided

$$|\tilde{\xi}^i-\xi^i| \leqq [18c^3(t_1-t_0)+1]\varepsilon_1 \qquad (i=1,\cdots,m).$$

II. We can determine a number $\varepsilon_2\in(0,\frac{1}{6}\varepsilon_1]$ such that, for all $t,\tau\in T$, $F\in\mathscr{F}^\gamma$, $v\in\mathbb{R}^n$ and K in either $\mathscr{L}(\mathbb{R}^n,\mathbb{R}^m)$ or $\mathscr{L}(\mathbb{R}^n,\mathbb{R}^{m_2})$, with $|v|,|K|\leqq c$, the expressions

$$Z(F)(t)^{-1}Z(F)(\tau)v \quad\text{and}\quad KZ(F)(t)^{-1}Z(F)(\tau)v$$

change by terms of norms not exceeding $(t_1-t_0+1)^{-1}\varepsilon_1$ whenever $(Z(F),K,v,\tau)$ is replaced by (Z^*,K^*,v^*,τ^*) such that

$$|Z(F)-Z^*|_{\sup}\leqq 2\varepsilon_2, \quad |K-K^*|\leqq 2\varepsilon_2, \quad |v-v^*|\leqq\varepsilon_2, \quad |\tau-\tau^*|\leqq\varepsilon_2.$$

III. Since the family $\{Z(F)|F\in\mathscr{F}^\gamma\}$ is bounded and equicontinuous and the set $\mathscr{H}^\gamma$ bounded, we can determine finite collections $\mathscr{F}\subset\mathscr{F}^\gamma$ and $\mathscr{H}\subset\mathscr{H}^\gamma$ with the property that for each $F\in\mathscr{F}^\gamma$ and $H\in\mathscr{H}^\gamma$ there exist $F^*\in\mathscr{F}$ and $H^*\in\mathscr{H}$ such that

$$|Z(F)-Z(F^*)|_{\sup}\leqq\varepsilon_2, \qquad |H-H^*|\leqq\varepsilon_2.$$

IV. Let the finite sets $\mathscr{G}$, $\mathscr{F}$ and $\mathscr{H}$ be defined as above. We set $\mathscr{L}=\mathscr{G}\times\mathscr{F}\times\mathscr{H}$ and, for each $L=(G^*,F^*,H^*)\in\mathscr{L}$, we denote by ξ^{Li} $(i=1,\cdots,m)$ the points ξ^i corresponding to G^* as defined in I.

We shall say that a function $\phi: T\to\mathbb{R}^b$ is *approximately continuous* at $\bar{t}$ if, for each $\varepsilon>0$,

$$\lim_{\alpha\to+0}(2\alpha)^{-1}\mu(\{t\in[\bar{t}-\alpha,\bar{t}+\alpha]\cap T \mid |\phi(t)-\phi(\bar{t})|>\varepsilon\})=0.$$

We shall henceforth denote by T^* the set of all points where

$$t \to f(t, y(f, \sigma_0)(t), \rho(t)) \quad \text{and} \quad t \to \alpha^j(t)$$

are approximately continuous for $j = 0, \cdots, n$, $\rho = \rho_j$ and $\rho \in \mathscr{R}_\infty^{\#}$. As is well known, $\mu(T^*) = \mu(T)$ (because $\mathscr{R}_\infty^{\#}$ is denumerable).

By Lemma 4.1, for each $L = (G^*, F^*, H^*) \in \mathscr{L}$ we can determine

$$l^L, t^{Lik} \in T^*, \quad a^{Lik} \in [0, 1], \quad \rho^{Li} \in \mathscr{R}_\infty^{\#}, \quad \delta^{Li}(t)$$
$$(i = 1, \cdots, m, \quad k = 1, \cdots, l^L, \quad t \in T),$$

satisfying the conditions of that lemma with

$$\mathbf{N} = T \sim T^*, \quad e = \varepsilon_1, \quad F = F^*, \quad H = H^*, \quad \xi^i = \xi^{Li}, \quad \delta^i = \delta^{Li}.$$

We may assume that the points t^{Lik} differ from each other and from t_1 because we may determine these points for any $L \in \mathscr{L}$ by defining the set $\mathbf{N}$ of Lemma 4.1 as the union of $[t_0, t_1) \sim T^*$ and the collection of all previously determined t^{L_1, i_1, k_1} Furthermore, since $\mathscr{L}$ is finite, we may replace l^L by

$$l = \sup \{l^{L_1} | L_1 \in \mathscr{L}\}$$

by defining the "missing" A^{Lik} as 0 and the "missing" t^{Lik} as arbitrary points in T^* distinct from each other, from t_1 and from those already determined.

V. Let d be the smallest of the distances between the various points t^{Lik} or t_1 and $c_1 = 2c \exp[c(t_1 - t_0)]$. We shall write $\mu(\sigma \neq \sigma')$ for $\mu(\{t \in T | \sigma(t) \neq \sigma'(t)\})$, and $|\sigma_1, \sigma_2|_{w,\mu} \leqq \alpha$ if there exists $\sigma_3 \in \mathscr{S}^{\#}$ such that $|\sigma_1 - \sigma_3|_w \leqq \alpha$ and $\mu(\sigma_3 \neq \sigma_2) \leqq \alpha$.

We can determine a positive integer p_0 and a number

$$\varepsilon_3 \in (0, \min[(cc_1)^{-1}\varepsilon_1, \varepsilon_2, d])$$

such that, for every choice of

$$L \in \mathscr{L}, \quad i \in \{1, \cdots, m\}, \quad F \in \mathscr{F}^\gamma, \quad p \geqq p_0, \quad r \in R, \quad \tau \in T, \quad t \in T^h$$

and $\sigma \in \mathscr{S}^{\#}$, with $|\sigma_0, \sigma|_{w,\mu} \leqq \varepsilon_3$, we have

$$y(f^p, \sigma)(T) \subset V^*, \qquad |f - f^p|_{\sup} \leqq \varepsilon_1,$$
$$f_v^p(\tau, y(f^p, \sigma)(\tau), \rho_j(\tau)) \in \Lambda^\gamma f(\tau, y(f, \sigma_0)(\tau), \rho_j(\tau)) \qquad (j = 0, \cdots, n),$$
$$h_v^{1,p}(y(f^p, \sigma)(t_1)) \in \mathscr{H}^\gamma,$$

and the values of

$$Z(F)(\tau), \quad f^p(\tau, y(f^p, \sigma)(\tau), r), \quad h^2(t, y(f^p, \sigma)(t)), \quad h_v^2(t, y(f^p, \sigma)(t))$$

and

$$h^2(t, y(f^p, \sigma)(t)) + h_v^2(t, y(f^p, \sigma(t))\delta^{Li}(t)$$

change by at most $\min(\varepsilon_2, \gamma/8)$ whenever any combination of the following changes is made: f^p is replaced anywhere by f, or σ is replaced anywhere by σ' with $|\sigma_0, \sigma'|_{w,\mu} \leqq \varepsilon_3$, or F is replaced by some measurable $\tilde{F}$ such that $|\tilde{F}|_{\sup} \leqq c$ and $\mu(F \neq \tilde{F}) \leqq \varepsilon_3$.

VI. Since $t^{Lik} \in T^*$ for each $L \in \mathscr{L}$, $i = 1, \cdots, m$, $k = 1, \cdots, l$, we can determine sets $T^{Lik} \subset [t^{Lik}, t_1)$ of positive measures and such that, for $t \in T^{Lik}$, $j = 0, \cdots, n$ and $\rho = \rho_j$, ρ^{Li} we have

$$|f(t, y(f, \sigma_0)(t), \rho(t)) - f(t^{Lik}, y(f, \sigma_0)(t^{Lik}), \rho(t^{Lik}))| \leqq \varepsilon_2,$$

$$\alpha^j(t) \geqq \tfrac{3}{4}\alpha^j(t^{Lik}), \qquad \text{diameter } (T^{Lik}) \leqq \varepsilon_3,$$

and the union of all T^{Lik} has a measure not exceeding ε_3. We may choose these sets so that they are disjoint and observe that, for each L, i and k, $\overline{\text{co}}\, T^{Lik}$ contains no point t^{L_1, i_1, k_1} with $(L_1, i_1, k_1) \neq (L, i, k)$. We set

$$\bar{\beta} = \tfrac{1}{2} \min \{\mu(T^{Lik}) | L \in \mathscr{L}, i = 1, \cdots, m, k = 1, \cdots, l\}.$$

VII. Let c_1 be as defined in V and $c_2 = 8m^2 c c_1 (t_1 - t_0)$. By VI and Lemma 3.5, we can determine a measurable partition $\{A_0, \cdots, A_n\}$ of T and a corresponding function $u_0 \in \mathscr{R}^{\#}$ such that

$$|\sigma_0 - u_0|_w \leqq \varepsilon_3,$$

$$u_0(t) = \rho_j(t) \qquad (t \in A_j, \quad j = 0, \cdots, n),$$

$$|h^{1,p}(y(f^p, u_0)(t_1)) - h^1(y(f, \sigma_0)(t_1))| \leqq \gamma\bar{\beta}/(4m),$$

$$|h^2(t, y(f^p, u_0)(t)) - h^2(t, y(f, \sigma_0)(t))| \leqq \gamma^2\bar{\beta}/(2c_2) \qquad (t \in T^h),$$

for all sufficiently large p, and

$$\mu(T^{Lik} \cap A_j) \geqq \tfrac{1}{2}\mu(T^{Lik})\alpha(t^{Lik}) \quad \text{for all } L, i, k, j. \tag{1}$$

VIII. For each L, i, k and j, we set

$$T^{Likj} = T^{Lik} \cap A_j$$

and determine a family $T^{Likj}(\alpha)$ $(\alpha \in [0, \bar{\beta}])$ of subsets of T^{Likj} such that

$$T^{Likj}(\alpha) \subset T^{Likj}(\alpha') \quad \text{if } \alpha < \alpha',$$

$$T^{Likj}(0) = \varnothing,$$

$$\mu(T^{Likj}(\alpha)) = \min(\alpha, \mu(T^{Likj})).$$

It follows from VII(1) and the definition of $\bar{\beta}$ in VI that $\mu(T^{Likj}(\alpha)) = \alpha$ whenever $\alpha \leqq \bar{\beta}\alpha^j(t^{Lik})$.

For each choice of $\omega = (\omega^{Likj})$, with $\omega^{Likj} \in [0, \bar{\beta}]$, we set

$$u(\omega)(t) = \begin{cases} \rho^{Li}(t) & [t \in T^{Likj}(\omega^{Likj})], \\ \\ u_0(t) & \text{elsewhere in } T. \end{cases}$$

We also denote by $|\mathscr{L}|$ the number of elements of $\mathscr{L}$, and set

$$\mathscr{T} = \{\theta = (\theta^{Li})_{L \in \mathscr{L}, i = 1, \cdots, m} | \theta^{Li} \in [0, \bar{\beta}]\} = [0, \bar{\beta}]^{m|\mathscr{L}|},$$

$$\omega^{Likj}(\theta) = a^{Lik}\alpha^j(t^{Lik})\theta^{Li} \qquad (\theta = (\theta^{Li}) \in \mathscr{T}),$$

$$\omega(\theta) = (\omega^{Likj}(\theta)).$$

We observe that $\mu(T^{Likj}(\omega^{Likj}(\theta))) = \omega^{Likj}(\theta)$ for every $\theta \in \mathscr{T}$.

In the remainder of this section we shall refer to the items I–VIII above and to the various objects defined there. We shall use letters L, i, j, k, with or without additional symbols, to represent elements of $\mathscr{L}$, $\{1, \cdots, m\}$, $\{0, \cdots, n\}$, $\{1, \cdots, l\}$, respectively. We shall use the symbol $O(a)$ (in a more restricted sense than usual) to represent an element x of a normed space such that $|x| \leqq a$.

LEMMA 4.3. *Let c_1 and c_2 be as defined in* V *and* VII, $p \geqq p_0$, $\mathbf{n} \in \mathbb{R}^m$, $|\mathbf{n}| = 1$, $\tilde{\theta} \in \mathscr{T}$, $\tilde{\beta} = \bar{\beta} - |\tilde{\theta}|$, $\tilde{u} = u(\omega(\tilde{\theta}))$ *and* $\mathbf{h}^1 = h^{1,p}(y(f^p, \tilde{u})(t_1))$. *Then for every* $b \in [0, \tilde{\beta}]$ *there exists* $\theta \in \mathscr{T}$ *such that, for all* $t \in T^h$ *and* $j = 1, \cdots, m_2$,

$$0 \leqq \theta^{Li} - \tilde{\theta}^{Li} \leqq b \quad (L \in \mathscr{L},\ i = 1, \cdots, m), \tag{1}$$

$$h^{1,p}(y(f^p, u(\omega(\theta)))(t_1)) = \mathbf{h}^1 + [\gamma b/(2m)]\mathbf{n}, \tag{2}$$

$$e_j \cdot h^2(t, y(f^p, u(\omega(\theta)))(t)) \leqq \max(e_j \cdot h^2(t, y(f^p, \tilde{u})(t)) - b\gamma^2/c_2, -\gamma/4). \tag{3}$$

Proof. Step 1. Let $\mathscr{B}$ denote the collection of all numbers $b' \in [0, \tilde{\beta}]$ such that for every $b \in [0, b']$ there exists $\theta \in \mathscr{T}$ satisfying relations (1)–(3). The point $b = 0$ belongs to $\mathscr{B}$ and corresponds to $\theta = \tilde{\theta}$. Furthermore, $\mathscr{B}$ is closed because the function $\theta \to y(f^p, u(\omega(\theta))) : \mathscr{T} \to C(T, \mathbb{R}^n)$ is continuous and $\mathscr{T}$ is compact. Now let $0 \leqq \bar{b} < \bar{\beta}$ and $\bar{b} \in \mathscr{B}$. We shall show that there exists $\beta_0 > 0$ such that $\bar{b} + \beta_0 \in \mathscr{B}$. Since $\mathscr{B}$ is closed, this will imply that $\mathscr{B} = [0, \tilde{\beta}]$.

Step 2. Let $\bar{\theta}$ be the value of θ corresponding to $\bar{b}$. We can determine a simplex $G^* \in \mathscr{G}$ corresponding to the point $[\gamma/(2m)]\mathbf{n}$ as in I. We let

$$\alpha_0 = \min\{\tfrac{1}{2}(\tilde{\beta} - \bar{b})a^{Lik}\alpha^j(t^{Lik}) \mid a^{Lik}\alpha^j(t^{Lik}) > 0, \text{ all } L, i, k, j\},$$

$$\bar{v}(t) = y(f^p, u(\omega(\bar{\theta})))(t),$$

$$\bar{F}(t) = f^p_v(t, \bar{v}(t), u(\omega(\bar{\theta}))(t)).$$

Since $\mu(u(\omega(\bar{\theta})) \neq u_0) \leqq \varepsilon_3$ and $|\sigma_0 - u_0|_w \leqq \varepsilon_3$, we may, by III and V, select elements $F^* \in \mathscr{F}$ and $H^* \in \mathscr{H}$ such that

$$|H^* - H^{1,p}_v(\bar{v}(t_1))| \leqq \varepsilon_2, \qquad |Z(F^*) - Z(\bar{F})|_{\sup} \leqq 2\varepsilon_2, \tag{4}$$

and we set

$$L^* = (G^*, F^*, H^*).$$

Since the function

$$\omega \to y(f^p, u(\omega)) : [0, \bar{\beta}]^{|\mathscr{L}|ml(n+1)} \to C(T, \mathbb{R}^n)$$

is continuous, we can determine $\beta_0 \in (0, \alpha_0]$ such that

$$\int_{t_0}^{t_1} \sup_{r \in R, 0 \leqq a \leqq 1} |f^p_v(t, ay(f^p, u(\omega))(t) + (1-a)\bar{v}(t), r) - f^p_v(t, \bar{v}(t), r)|\, dt \leqq \varepsilon_1/c_1,$$

$$\sup_{0 \leqq a \leqq 1} |h^{1,p}_v(ay(f^p, u(\omega))(t_1) + (1-a)\bar{v}(t_1)) - h^{1,p}_v(\bar{v}(t_1))| \leqq \varepsilon_2, \tag{5}$$

$$\sup_{t \in T^h, 0 \leqq a \leqq 1} |h^2_v(t, ay(f^p, u(\omega))(t) + (1-a)\bar{v}(t)) - h^2_v(t, \bar{v}(t))| \leqq [16c^3(t_1 - t_0) + 1]^{-1}\gamma,$$

provided $|\omega^{Likj} - \omega^{Likj}(\bar{\theta})| \leqq 2\beta_0$ for all L, i, k and j.

For each choice of $\zeta=(\zeta^{ikj})$ with elements in $[0, 2\beta_0]$, we consider the array $\tilde{\omega}(\zeta)=(\tilde{\omega}^{Likj}(\zeta))$, defined by

$$\tilde{\omega}^{Likj}(\zeta)=\omega^{Likj}(\bar{\theta}) \qquad (L\neq L^*),$$
$$\tilde{\omega}^{L^*ikj}(\zeta)=\omega^{L^*ikj}(\bar{\theta})+\zeta^{ikj}.$$

We choose a fixed (i^*, k^*, j^*) with $a^{L^*i^*k^*}\alpha^{j^*}(t^{L^*i^*k^*})>0$, denote $\tilde{\omega}(\zeta)$ by $\hat{\omega}(\alpha)$ when $\zeta^{i^*k^*j^*}=\alpha$ while all other ζ^{ikj} are kept fixed, and set

$$\hat{u}(\alpha)=u(\hat{\omega}(\alpha)), \qquad \hat{y}(\alpha)=y(f^p, \hat{u}(\alpha)),$$
$$M^*=(L^*, i^*, k^*, j^*),$$
$$\hat{T}(\alpha, \beta)=T^{M^*}(\hat{\omega}^{M^*}(\alpha+\beta))\sim T^{M^*}(\hat{\omega}^{M^*}(\alpha)) \qquad (\alpha, \beta\in[0, \beta_0]),$$
$$\Delta(\alpha, \beta)(t)=\Delta(t)=\beta^{-1}[\hat{y}(\alpha+\beta)(t)-\hat{y}(\alpha)(t)] \qquad (\alpha, \beta\in[0, \beta_0], \beta>0),$$
$$\tau^*=t^{L^*i^*k^*}, \qquad f^*=f(\tau^*, y(f, \sigma_0)(\tau^*)), \rho^{L^*i^*}(\tau^*)\sim\rho_{j^*}(\tau^*)).$$

We observe that for all $\alpha, \beta\in[0, \beta_0]$, we have

$$\omega^{M^*}(\bar{\theta})+\alpha+\beta\leqq\bar{\beta}a^{L^*i^*k^*}\alpha^{j^*}(\tau^*)\leqq\bar{\beta}\alpha^{j^*}(\tau^*),$$

and therefore $\mu(\hat{T}(\alpha, \beta))=\beta$.

Step 3. We shall next show that, for every choice of

$$i^*, k^*, j^*, \alpha\in[0, \beta_0], \quad \beta\in(0, \beta_0], \quad t\in T,$$

we have

$$\Delta(t)=0 \qquad (t\leqq\tau^*), \qquad |\Delta(t)|\leqq c_1 \qquad (\tau^*\leqq t\leqq\tau^*+\varepsilon_3),$$
$$|\Delta(t)-Z(F^*)(t)^{-1}Z(F^*)(\tau^*)f^*|\leqq 5c^2\varepsilon_1 \qquad (t\geqq\tau^*+\varepsilon_3). \tag{6}$$

For all $\alpha, \beta\in[0, \beta_0]$, we have

$$\hat{u}(\alpha+\beta)(\tau)=\hat{u}(\alpha)(\tau) \qquad (\tau\in T\sim\hat{T}(\alpha, \beta)),$$
$$\hat{u}(\alpha+\beta)(\tau)=\rho^{L^*i^*}(\tau), \quad \hat{u}(\alpha)(\tau)=u_0(\tau)=\rho_{j^*}(\tau) \qquad (\tau\in\hat{T}(\alpha, \beta)),$$
$$\hat{T}(\alpha, \beta)\in[\tau^*, t_1], \qquad \mu(\hat{T}(\alpha, \beta))=\beta.$$

Thus

$$\hat{y}(\alpha, \beta)(t)=\bar{a}_0+\int_{t_0}^{t} f^p(\tau, \hat{y}(\alpha+\beta)(\tau), \hat{u}(\alpha+\beta)(\tau))\,d\tau$$
$$=\bar{a}_0+\int_{t_0}^{t} f^p(\tau, \hat{y}(\alpha+\beta)(\tau), \hat{u}(\alpha)(\tau))\,d\tau$$
$$+\int_{\hat{T}(\alpha,\beta)\cap[t_0,t]} f^p(\tau, \hat{y}(\alpha+\beta)(\tau), \hat{u}(\alpha+\beta)(\tau)\sim\tilde{u}(\alpha)(\tau))\,d\tau \qquad (t\in T),$$

and, for $\beta>0$,

$\Delta(t)=0$ if $t\leqq\tau^*$, and otherwise

$$\Delta(t)=\beta^{-1}\int_{\tau^*}^{t}[f^p(\tau, \hat{y}(\alpha+\beta)(\tau), \hat{u}(\alpha)(\tau))-f^p(\tau, \hat{y}(\alpha)(\tau), \hat{u}(\alpha)(\tau))]\,d\tau$$
$$+\beta^{-1}\int_{\hat{T}(\alpha,\beta)\cap[t_0,t]} f^p(\tau, \hat{y}(\alpha+\beta)(\tau), \rho^{L^*i^*}(\tau)\sim\rho_{j^*}(\tau))\,d\tau. \tag{7}$$

By V, $\hat{y}(\alpha+\beta)(T)\subset V^*$. Since $\mu(\hat{T}(\alpha,\beta))=\beta$, (7) implies that

$$|\Delta(t)|\leqq c\int_{t^*}^{t}|\Delta(\tau)|\,d\tau+2c \qquad (t\geqq\tau^*),$$

and we deduce by Gronwall's inequality that

$$|\Delta(t)|\leqq c_1 \qquad (t\geqq\tau^*). \tag{8}$$

Furthermore, it follows from V and VI, that, for $\rho=\rho^{L^*i^*}$, ρ_{j^*} and all $\tau\in\hat{T}(\alpha,\beta)$,

$$|\,f^p(\tau,\hat{y}(\alpha+\beta)(\tau),\rho(\tau))-f(\tau^*,y(f,\sigma_0)(\tau^*),\rho(\tau^*))|\leqq 3\varepsilon_2\leqq\varepsilon_1/2.$$

Thus (5), (7) and (8) imply that, for $\alpha,\beta\in[0,\beta_0]$, $\beta>0$,

$$\Delta(t)=\begin{cases}0 & \text{if } t\leqq\tau^*,\\ \displaystyle\int_{\tau^*}^{t}f_v^p(\tau,y(f^p,u(\omega(\bar{\theta})))(\tau),\hat{u}(\alpha)(\tau))\Delta(\tau)\,d\tau & \\ \quad+\beta^{-1}\mu(\hat{T}(\alpha,\beta)\cap[t_0,t])f^*+O(2\varepsilon_1) & \text{otherwise.}\end{cases} \tag{9}$$

Since $\hat{u}(\alpha)$ coincides with $u(\omega(\bar{\theta}))$ except on a set of measure at most ε_3 and $cc_1\varepsilon_3\leqq\varepsilon_1$, it follows from (9) that

$$\Delta(t)=\int_{\tau^*+\varepsilon_3}^{t}\bar{F}(\tau)\Delta(\tau)\,d\tau+f^*+O(4\varepsilon_1) \qquad (t\geqq\tau^*+\varepsilon_3),$$

which, together with (4), (8), (9) and II, implies the validity of relation (6).

Step 4. Now let

$$\psi(\zeta)(t)=y(f^p,u(\tilde{\omega}(\zeta)))(t) \qquad (t\in T,\quad \zeta^{ikj}\in[0,2\beta_0]),$$
$$\zeta^{ikj}(\eta)=a^{L^*ik}\alpha^j(t^{L^*ik})\eta^i \qquad (\eta=(\eta^1,\cdots,\eta^m)\in[0,2\beta_0]^m),$$
$$\zeta(\eta)=(\zeta^{ikj}(\eta)),\qquad \phi(\eta)=\psi(\zeta(\eta)),$$

and let $\Delta^{i^*k^*j^*}(t)=\Delta^{i^*k^*j^*}(\alpha,\beta)(t)$ denote $\Delta(t)$ of Steps 2 and 3. Then $\Delta^{i^*k^*j^*}(\alpha,\beta)(t)$ represents a partial difference quotient of $\psi(\zeta)(t)$ with respect to $\zeta^{i^*k^*j^*}$ between the values α and $\alpha+\beta$. It follows that, for each choice of i, of $\eta\in[0,\beta_0]^m$, and of $\beta\in(0,\beta_0]$,

$$\beta^{-1}[\phi(\eta+\beta e_i)(t)-\phi(\eta)(t)]=\sum_{j,k}a^{L^*ik}\alpha^j(t^{L^*ik})\Delta^{ikj}(t), \tag{10}$$

where each Δ^{ikj} is evaluated for appropriate choices of ζ with elements in $[0,2\beta_0]$. By IV and Lemma 4.1,

$$a^{Lik}\in[0,1],\quad a^{Lik}\leqq(16cc_1)^{-1}\gamma,\quad \sum_k a^{Lik}=t_1-t_0;$$

and by V, any subinterval of T of length ε_3 contains at most one of the points t^{Lik} or t_1. Since

$$f(t,v,\sigma_0(t))=\sum_j\alpha^j(t)f(t,v,\rho_j(t)),$$

we deduce from (6) and (10) that

$$\begin{aligned}&\beta^{-1}[\phi(\eta+\beta e_i)(t)-\phi(\eta)(t)]\\ &=\sum_k a^{L^*ik}Z(F^*)(t)^{-1}Z(F^*)(t^{L^*ik})\hat{f}^{ik}\chi_{[t_0,t]}(t^{L^*ik})\\ &\quad+O(5c^2(t_1-t_0)\varepsilon_1+(16c)^{-1}\gamma\chi_{[t_0,t_1)}(t)),\end{aligned}\tag{11}$$

where

$$\hat{f}^{ik}=f(t^{L^*ik},\, y(f,\sigma_0)(t^{L^*ik}),\, \rho^{L^*i}(t^{L^*ik})\sim\sigma_0(t^{L^*ik})).$$

If we denote the first (main) term on the right of (11) by $\delta^{L^*i}(t)$ (as in IV and Lemma 4.1), then we deduce from (11) that

$$\phi(\eta)(t)-\phi(0)(t)=\sum_{i=1}^m \eta^i[\delta^{L^*i}(t)+O(5c^2(t_1-t_0)\varepsilon_1+(16c)^{-1}\gamma)]$$

and, by the definition of ε_1 in I,

$$\phi(\eta)(t)-\phi(0)(t)=\sum_{i=1}^m \eta^i[\delta^{L^*i}(t)+O((8c)^{-1}\gamma)].\tag{12}$$

It is easily verified that $|\delta^{L^*i}(t)|\leqq 2c^3(t_1-t_0)$ for all i and t. We can now deduce from (4), (5), (11), (12), II and IV that

$$\begin{aligned}&\beta^{-1}[h^{1,p}(\phi(\eta+\beta e_i)(t_1))-h^{1,p}(\phi(\eta)(t_1))]\\ &=(H^*+O(2\varepsilon_2))\delta^{L^*i}(t_1)+O(18c^3(t_1-t_0)\varepsilon_1)\\ &=\xi^{L^*i}+O([18c^3(t_1-t_0)+1]\varepsilon_1)\end{aligned}\tag{13}$$

and, setting $c'=[16c^3(t_1-t_0)+1]^{-1}$, that

$$\begin{aligned}h^2(t,\phi(\eta)(t))&=h^2(t,\phi(0)(t))+[h_v^2(t,\phi(0)(t))+O(c'\gamma)][\phi(\eta)(t)-\phi(0)(t)]\\ &=h^2(t,\bar{v}(t))+\sum_{i=1}^m \eta^i[h_v^2(t,\bar{v}(t))\delta^{L^*i}(t)+O(\gamma/4)].\end{aligned}\tag{14}$$

Now, by V and Lemma 4.1,

$$e_j\cdot[h^2(t,\bar{v}(t))+h_v^2(t,\bar{v}(t))\delta^{L^*i}(t)]\leqq-\tfrac{7}{8}\gamma\tag{15}$$

for all i, j and $t\in T^h$. We first consider values of j and $t\in T^h$ for which

$$e_j\cdot h^2(t,\bar{v}(t))\leqq-\tfrac{3}{8}\gamma.$$

Then, by V,

$$e_j\cdot h^2(t,\phi(\eta)(t))\leqq-\gamma/4.$$

For other j and t, it follows from (15) that

$$e_j\cdot h_v^2(t,\bar{v}(t))\delta^{L^*i}(t)\leqq-\gamma/2\quad\text{for all } i,$$

and therefore, by (14),

$$e_j\cdot h^2(t,\phi(\eta)(t))\leqq e_j\cdot h^2(t,\bar{v}(t))-\frac{\gamma}{4}\sum_{i=1}^m \eta^i.$$

Thus

$$e_j \cdot h^2(t, \phi(\eta)(t)) \leqq \max\left(e_j \cdot h^2(t, \bar{v}(t)) - \frac{\gamma}{4}\sum_{i=1}^{m}\eta^i, -\frac{\gamma}{4}\right) \tag{16}$$

$$(j = 1, \cdots, m_2, \quad t \in T^h, \quad \eta \in [0, \beta_0]^m).$$

Step 5. Let Δ^ε be the derivate container operator described in the remarks following Definition 2.1, and let

$$\hat{\phi}(\eta) = h^{1,p}(\phi(\eta)(t_1)) \quad \text{and} \quad \beta' \in (0, \beta_0].$$

By (13), for every $\eta \in (0, \beta')^m$ and every

$$\Phi = [\Phi^1, \cdots, \Phi^m] \in \Delta^{\beta_0}\hat{\phi}(\eta),$$

we have

$$|\Phi^i - \xi^{L^*i}| \leqq [18c_*^3(t_1 - t_0) + 1]\varepsilon_1 \quad \text{for all } i;$$

hence by I,

$$|\Phi^{-1}| \leqq 2m/\gamma \quad \text{and} \quad \mathbf{n} \in \Phi[0, \infty)^m.$$

It follows, by Lemma 3.3, that

$$\hat{\phi}(0) + [0, \beta'/(2m/\gamma)]\mathbf{n} \subset \hat{\phi}([0, \beta']^m).$$

Thus there exist $\eta \in [0, \beta']^m$ and a corresponding $\theta \in \mathscr{T}$ such that

$$\theta^{Li} = \bar{\theta}^{Li} \leqq b \qquad (L \neq L^*), \qquad \theta^{L^*i} = \bar{\theta}^{L^*i} + \eta^i \leqq \bar{b} + \beta' \tag{17}$$

and

$$\begin{aligned}\hat{\phi}(\eta) &= h^{1,p}(y(f^p, u(\omega)(\theta)))(t_1)) \\ &= \mathbf{h}^1 + [(\bar{b} + \beta')\gamma/(2m)]\mathbf{n} = \hat{\phi}(0) + [\beta'\gamma/(2m)]\mathbf{n}.\end{aligned} \tag{18}$$

Finally, we deduce from (8)–(10) that ϕ has a Lipschitz constant $mc_1(t_1 - t_0)$ and therefore $\hat{\phi}$ has a Lipschitz constant $mcc_1(t_1 - t_0)$. It follows therefore from (18) that

$$\beta'\gamma/(2m) = |\hat{\phi}(\eta) - \hat{\phi}(0)| \leqq mcc_1(t_1 - t_0)|\eta| \leqq mcc_1(t_1 - t_0)\sum_{i=1}^{m}\eta^i;$$

hence, by (16),

$$\begin{aligned}e_j \cdot h^2(t, \phi(\eta)(t)) &\leqq \max\,(e_j \cdot h^2(t, \bar{v}(t)) - \beta'\gamma^2/c_2, -\gamma/4) \\ &= \max\,(e_j \cdot h^2(t, y(f^p, \tilde{u})(t)) - (\bar{b} + \beta')\gamma^2/c_2, -\gamma/4)\end{aligned} \tag{19}$$

for all $t \in T^h$ and $j = 1, \cdots, m_2$. Relations (17)–(19) show that $\bar{b} + \beta_0 \in \mathscr{B}$, and therefore $\mathscr{B} = [0, \tilde{\beta}]$. Q.E.D.

4.4. Proof of Theorem 2.3. Let $p \geqq p_0$,

$$w(\theta)(t) = y(f^p, u(\omega(\theta)))(t) \qquad (\theta \in \mathscr{T}, \quad t \in T),$$

$\tilde{\theta} \in \mathscr{T}$ and $0 \leqq s \leqq (\bar{\beta} - |\tilde{\theta}|)\gamma/(2m)$. Then it follows from Lemma 4.3 that for every

point $v \in \mathbb{R}^m$ at a distance s from $h^{1,p}(w(\tilde{\theta})(t_1))$ there exists $\theta \in \tilde{\theta} + [0, 2ms/\gamma]^{m|\mathscr{L}|} \in \mathscr{T}$ such that $v = h^{1,p}(w(\theta)(t_1))$ and

$$e_j \cdot h^2(t, w(\theta)(t)) \leqq \max\,(e_j \cdot h^2(t, w(\tilde{\theta})(t)) - (2m\gamma s/c_2), -\gamma/4)$$
$$(t \in T^h, \quad j = 1, \cdots, m_2).$$

This implies that if v is the endpoint of a polygonal line in $\mathbb{R}^m$ of length $\bar{\beta}\gamma/(2m)$ and originating at $h^{1,p}(y(f^p, u_0)(t_1))$, then there exists $\theta_p \in \mathscr{T}$ such that $v = h^{1,p}(w(\theta_p)(t_1))$ and

$$e_j \cdot h^2(t, w(\theta_p)(t)) \leqq \max\,(e_j \cdot h^2(t, y(f^p, u_0)(t)) - (\bar{\beta}\gamma^2/c_2), -\gamma/4)$$
$$(t \in T^h, \quad j = 1, \cdots, m_2).$$

It is clear that every point in $S^F(v_0, \varepsilon)$ is the endpoint of a polygonal line originating at v_0 and of length ε. Therefore, the above relation together with the relations in VII shows that for all sufficiently large p and all

$$(1) \qquad v \in S^F(h^1(y(f, \sigma_0)(t_1)), \gamma\bar{\beta}/(4m)) \subset S^F(h^{1,p}(y(f^p, u_0)(t_1)), \gamma\bar{\beta}/(2m))$$

there exists $\theta_p \in \mathscr{T}$ such that

$$(2) \qquad v = h^{1,p}(y(f^p, u(\omega(\theta_p)))(t_1))$$

and

$$(3) \qquad e_j \cdot h^2(t, y(f^p, u(\omega(\theta_p)))(t)) \leqq \max\,(-\tfrac{1}{2}\bar{\beta}\gamma^2/c_2, -\gamma/4)$$
$$(t \in T^h, \quad j = 1, \cdots, m_2).$$

Because $\mathscr{T}$ is compact, we may choose a sequence $P \subset (1, 2, \cdots)$ such that $(\theta_p)_{p \in P}$ converges to some $\bar{\theta} \in \mathscr{T}$. We can enumerate the finite collection of the control functions

$$\rho_j, \rho^{Li} \qquad (j = 0, \cdots, n, \quad i = 1, \cdots, m, \quad L \in \mathscr{L}),$$

as $u_1, \cdots, u_N$. Then

$$u(\omega(\theta))(t) \in \{u_1(t), \cdots, u_N(t)\} \qquad (\theta \in \mathscr{T}, \quad t \in T),$$

and the first part of Theorem 2.3 follows therefore from relations (1)–(3) by letting $p \to \infty$, $p \in P$, and setting

$$\kappa = \min\,(\gamma\bar{\beta}/(4m), \tfrac{1}{2}\bar{\beta}\gamma^2/c_2, \gamma/4).$$

In particular, there exists $\theta_0 \in \mathscr{T}$ such that

$$h^1(y(f, u(\omega(\theta_0)))(t_1)) = h^1(y(f, \sigma_0)(t_1))$$

and

$$h^2(t, y(f, u(\omega(\theta_0)))(t)) \in (-\infty, 0)^{m_2} \qquad (t \in T^h).$$

Finally, our previous arguments remain valid if we replace the number ε_3, chosen in V, by any smaller positive number. In particular, if (ε_3^i) is a sequence in $(0, \varepsilon_3]$ decreasing to 0, then for every i we can replace ε_3 by ε_3^i. We then denote by

u_0^i and $\hat{u}^i$ the control functions corresponding to u_0 and $u(\omega(\theta_0))$. By VII and VIII,

$$|u_0^i - \sigma_0|_w \leqq \varepsilon_3^i \quad \text{and} \quad \mu(\hat{\mu}^i \neq u_0^i) \leqq \varepsilon_3^i.$$

It follows that $\lim_i \hat{u}^i = \sigma_0$, while

$$h^1(y(f, \hat{u}^i)(t_1)) = h^1(y(f, \sigma_0)(t_1))$$

and

$$h^2(t, y(f, \hat{u}^i)(t)) \in (-\infty, 0)^{m_2} \qquad (t \in T^h),$$

for all $i = 1, 2, \cdots$. Q.E.D.

5. Proofs of Theorems 2.4 and 2.5.

5.1. A related problem. We shall apply Theorem 2.3 to a related problem for which an interval $[\alpha, \beta]$ is given and relations (1) and (3) of § 1 are replaced, respectively, by

$$\dot{y}(t) = f(t, y(t), u(t)), \quad \dot{\xi}_0(t) = 0, \quad \dot{\xi}(t) = 0 \quad \text{a.e. in } T \tag{1'}$$

and

$$(y(t_0), \xi_0(t_0), \xi(t_0)) \in A_0 \times [\alpha, \beta] \times A_1 \quad \text{a.e. in } T, \tag{3'}$$

while $h^1(y(t_1))$ is replaced by

$$\tilde{h}^1((y, \xi_0, \xi)(t_1)) = (h^0(y(t_1)) + \xi_0(t_1), h^1(y(t_1)) - \xi(t_1)).$$

We deduce from the remarks that follow Definition 2.1 (of a derivate container) that, if a component ϕ^j of a Lipschitz-continuous function $\phi = (\phi^1, \cdots, \phi^b) : A \to \mathbb{R}^b$ has a continuous derivative, then we may choose the jth row of elements of $\Lambda^\varepsilon \phi(a)$ as $\{\phi_v^j(v) \mid |v - a| \leqq \varepsilon/2\}$. Similarly, if $\phi = \phi^* + \phi^{**}$ and ϕ_v^{**} exists and is continuous then we may set

$$\Lambda^\varepsilon(\phi^* + \phi^{**})(a) = \Lambda^\varepsilon \phi^*(a) + \{\phi_v^{**}(v) \mid |v - a| \leqq \varepsilon/2\} \qquad (a \in A).$$

Thus, if $\Lambda^\varepsilon f$ and $\Lambda^\varepsilon(h^0, h^1)$ are given, then we can easily determine corresponding derivate containers for $(f, 0, 0)$ and $\tilde{h}$. A simple computation then shows that if $(\bar{\sigma}, (\bar{a}_0, 0, \bar{a}_1))$ is extremal relative to

$$\Omega_{\alpha,\beta} = ((f, 0, 0), \Lambda^\varepsilon(f, 0, 0), \tilde{h}^1, \Lambda^\varepsilon \tilde{h}^1, h^2, A_0 \times [\alpha, \beta] \times A_1),$$

then there exist $(l_0, l_1) \in \mathbb{R}^{m+1}$, nonnegative Radon measures $\omega_1, \cdots, \omega_{m_2}$ on T^h, a measurable $F : T \to \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n)$ and $H \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^{m+1})$ that satisfy relations (1)–(5) of Definition 2.2 with l_1 replaced by (l_0, l_1) and h^1 by (h^0, h^1), and that, furthermore,

$$0 = \min_{\xi_0 \in [\alpha, \beta]} l_0 \xi_0 \tag{1}$$

and

$$l_1^T \bar{a}_1 = \max_{a_1 \in A_1} l_1^T a_1. \tag{2}$$

5.2. Proof of Theorem 2.4. Let $(\sigma_0, \bar{a}_0)$ be either a minimizing relaxed solution or a minimizing original solution, $\bar{y} = y(f, \sigma_0, \bar{a}_0)$, $\bar{a}_1 = h^1(\bar{y}(t_1))$, $\alpha = 0$ and $\beta = 1$. If $(\sigma_0, (\bar{a}_0, 0, \bar{a}_1))$ is nonextremal relative to $\Omega_{\alpha,\beta} = \Omega_{0,1}$ (as defined in

5.1) then, by Theorem 2.3, there exist points $a_0 \in A_0$, $\xi_0 \in [0, 1]$, $\xi \in A_1$ and $u \in \mathcal{R}^{\#}$ such that

$$h^0(y(f, u, a_0)(t_1)) + \xi_0 < h^0(\bar{y}(t_1)),$$
$$h^1(y(f, u, a_0)(t_1)) - \xi = 0,$$
$$h^2(t, y(f, u, a_0)(t)) \in (-\infty, 0]^{m_2} \qquad (t \in T^h).$$

Since $\xi_0 \geqq 0$ and $\xi \in A_1$, this contradicts the assumption that $(\sigma_0, \bar{a}_0)$ is either a minimizing relaxed solution or a minimizing original solution. Thus $(\sigma_0, (\bar{a}_0, 0, \bar{a}_1))$ is extremal relative to $\Omega_{0,1}$ and it follows from 5.1 that $(\sigma_0, \bar{a}_0)$ is extremal relative to $\Omega^{0,1}$ as defined in the statement of Theorem 2.4. Relation (1) now shows that $l_0 \geqq 0$ and relation (2) of 5.1 shows that $l_1^T \bar{a}_1 = \max_{a_1 \in A_1} l_1^T a_1$. Q.E.D.

5.3. Proof of Theorem 2.5. Now assume that $(\bar{u}, \bar{a}_0)$ is a strict original solution. Then the set $\mathcal{M}^-$ is nonempty by the very definition of a strict original solution. Now let $(\tilde{\sigma}, \tilde{a}_0) \in \mathcal{M}^-$ and set

$$\beta = -\alpha = \tfrac{1}{2}[h^0(y(f, \bar{u}, \bar{a}_0)(t_1)) - h^0(y(f, \tilde{\sigma}, \tilde{a}_0)(t_1))],$$
$$\tilde{a}_1 = h^1(y(f, \tilde{\sigma}, \tilde{a}_0)(t_1)).$$

If $(\tilde{\sigma}, (\tilde{a}_0, 0, \tilde{a}_1))$ is nonextremal relative to $\Omega_{\alpha,\beta}$ (as defined in 5.1), then the same argument as in the proof of Theorem 2.4 shows that there exist $a_0 \in A_0$, $\xi_0 \in [\alpha, \beta]$, $\xi \in A_1$ and $u \in \mathcal{R}^{\#}$ such that

$$h^0(y(f, u, a_0)(t_1)) + \xi_0 < h^0(y(f, \tilde{\sigma}, \tilde{a}_0)(t_1)) < h^0(y(f, \bar{u}, \bar{a}_0)(t_1)),$$
$$h^1(y(f, u, a_0)(t_1)) \in A_1,$$
$$h^2(t, y(f, u, a_0)(t_1)) \in (-\infty, 0]^{m_2} \qquad (t \in T^h).$$

The first of these relations implies that

$$h^0(y(f, u, a_0)(t_1)) < h^0(y(f, \bar{u}, \bar{a}_0)(t_1)),$$

contradicting the assumption that $(\bar{u}, \bar{a}_0)$ is a minimizing original solution. Thus $\tilde{\sigma}, (\tilde{a}_0, 0, \tilde{a}_1))$ is extremal relative to $\Omega_{\alpha,\beta}$ and it follows from 5.1 that there exist (l_0, l_1), ω_j, F and H such that $(\tilde{\sigma}, \tilde{a}_0, (l_0, l_1), \omega_j, F, H)$ is extremal relative to $\Omega^{0,1}$. Since $\alpha < 0 < \beta$, relation 5.1(1) yields $l_0 = 0$ and, as before, 5.1(2) yields $l_1^T \bar{a}_1 = \max_{a_1 \in A_1} l_1^T a_1$. Q.E.D.

REFERENCES

[1] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[2] ———, *Optimal solutions to differential inclusions*, J. Optimization Theory Appl. to appear.

[3] R. V. GAMKRELIDZE, *On sliding optimal states*, Dokl. Akad. Nauk SSSR, 143 (1962), pp. 1243–1245. (In Russian.)

[4] H. HALKIN, *Implicit functions and optimization problems without continuous differentiability of the data*, this Journal, 12 (1974), pp. 229–236.

[5] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N. J., 1970.

[6] A. B. SCHWARZKOPF, *Controllability and tenability of nonlinear systems with state equality constraints*, this Journal, 13 (1975), pp. 695–705.

[7] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[8] ———, *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 18 (1975), pp. 41–62.
[9] ———, *Necessary conditions without differentiability assumptions in unilateral control problems*, Ibid., to appear.
[10] J. A. YORKE, *The maximum principle and controllability of nonlinear systems*, this Journal. 10 (1972), pp. 334–338.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 3, May 1976

STOCHASTIC CONVEX PROGRAMMING: RELATIVELY COMPLETE RECOURSE AND INDUCED FEASIBILITY*

R. T. ROCKAFELLAR† AND R. J-B. WETS‡

Abstract. The basic dual problem and extended dual problem associated with a two-stage stochastic program are shown to be equivalent, if the program is strictly feasible and satisfies a condition generalizing, in a sense, the condition of relatively complete recourse in stochastic linear programming. Combined with earlier results, this yields the fact that, under the same assumptions, solutions to the program can be characterized in terms of saddle points of the basic Lagrangian. A couple of examples are used to illustrate the salient points of the theory. The last section contains a review of the principal implications of the results of this paper combined with those of three preceding papers also devoted to stochastic convex programs.

1. Introduction. This is the fourth in a series of papers [1], [2], [3] devoted to the following two-stage model in stochastic programming. Let C_1 and C_2 be nonempty, closed convex sets in R^{n_1} and R^{n_2}, respectively, and let (S, Σ, σ) be a probability space. Let f_{1i} be a finite convex function on R^{n_1} for $i = 0, 1, \cdots, m_1$, and let $f_{2i}(s, \cdot, \cdot)$ be a finite convex function on $R^{n_1} \times R^{n_2}$ for $i = 0, 1, \cdots, m_2$ and $s \in S$. The problem is to minimize

$$f_{10}(x_1) + \int_S f_{20}(s, x_1, x_2(s))\sigma(ds) \tag{1.1}$$

over all $x_1 \in R^{n_1}$ and $x_2 \in \mathscr{L}^\infty_{n_2} = \mathscr{L}^\infty(S, \Sigma, \sigma; R^{n_2})$ (the Lebesgue space of equivalence classes) satisfying

$$x_1 \in C_1 \quad \text{and} \quad f_{1i}(x_1) \leqq 0 \quad \text{for } i = 1, \cdots, m_1, \tag{1.2}$$

and almost surely

$$x_2(s) \in C_2 \quad \text{and} \quad f_{2i}(s, x_1, x_2(s)) \leqq 0 \quad \text{for } i = 1, \cdots, m_2. \tag{1.3}$$

It is assumed that $f_{2i}(s, x_1, x_2)$ is measurable in s for each $(x_1, x_2) \in R^{n_1} \times R^{n_2}$, in fact summable if $i = 0$ and bounded if $i = 1, \cdots, m_2$. (From this it follows that for each $x_1 \in R^{n_1}$ and $x_2 \in \mathscr{L}^\infty_{n_2}$, $f_{2i}(s, x_1, x_2(s))$ is measurable in s, summable if $i = 0$ and essentially bounded if $i = 1, \cdots, m_2$.)

The *basic Lagrangian* function introduced for this problem in [1] is defined on the product of the sets

$$X_0 = \{(x_1, x_2) \in R^{n_1} \times \mathscr{L}^\infty_{n_2} | x_1 \in C_1 \text{ and almost surely } x_2(s) \in C_2\}, \tag{1.4}$$

$$Y_0 = \{(y_1, y_2) \in R^{m_1} \times \mathscr{L}^1_{m_2} | y_1 \geqq 0 \text{ and almost surely } y_2(s) \geqq 0\}, \tag{1.5}$$

* Received by the editors October 7, 1974, and in revised form May 29, 1975.
† Department of Mathematics, University of Washington, Seattle, Washington 98195.
‡ Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506.

by the formula

$$L(x_1, x_2, y_1, y_2) = f_{10}(x_1) + \sum_{i=1}^{m_1} y_{1i} f_{1i}(x_1) + \int_S [f_{20}(s, x_1, x_2(s)) + \sum_{i=1}^{m_2} y_{2i}(s) f_{2i}(s, x_1, x_2(s))]\sigma(ds). \tag{1.6}$$

The given problem can be identified with

P minimize $f(x_1, x_2)$ over all $(x_1, x_2) \in X_0$, where

$$f(x_1, x_2) = \sup_{(y_1, y_2) \in Y_0} L(x_1, x_2, y_1, y_2).$$

The *basic dual problem* is

D maximize $g(y_1, y_2)$ over all $(y_1, y_2) \in Y_0$, where

$$g(y_1, y_2) = \inf_{(x_1, x_2) \in X_0} L(x_1, x_2, y_1, y_2).$$

The relationship between P and D was studied in [1], and it was shown in particular that

$$\min \mathrm{P} = \sup \mathrm{D} \quad \text{if } C_1 \text{ and } C_2 \text{ are bounded.} \tag{1.7}$$

In cases where actually min P = max D, a pair $(\bar{x}_1, \bar{x}_2)$ solves P if and only if there exists $(\bar{y}_1, \bar{y}_2) \in Y_0$ such that $(\bar{x}_1, \bar{x}_2, \bar{y}_1, \bar{y}_2)$ is a saddle point of the Lagrangian. This saddle point property was reduced in [2] to a certain set of Kuhn–Tucker conditions involving a function $p \in \mathscr{L}^1_{n_1}$ which essentially associates prices with the constraint that x_1 must be chosen before the observation of s. The pairs $(\bar{y}_1, \bar{y}_2)$ are, of course, solutions to D.

To apply this basic duality theory at its fullest, one needs a simple criterion for the relation inf P = max D. But the latter does not hold in general, even if P is *strictly feasible* in the sense that for some $\varepsilon > 0$ the constraints (1.2) and (almost surely) (1.3) can be satisfied with $-\varepsilon$ in place of 0.

The goal of this paper is to obtain such a criterion in supplementing strict feasibility by a condition on the availability of second-stage recourse. The technique is to analyze the so-called induced constraints in the first stage in terms of the "extended duality" developed in [3]. The extended duality adjoins to the Lagrangian additional terms involving "singular" linear functionals on $\mathscr{L}^\infty_1$. It is interesting that, despite reliance on such esoteric objects in the proof, our main result on basic duality makes no mention of them in its statement.

Let K_1 be the set of all $x_1 \in R^{n_1}$ satisfying the first-stage constraints (1.2) and let K_2 be the set of all $x_1 \in R^{n_1}$ such that there exists an $x_2 \in \mathscr{L}^\infty_{n_2}$ satisfying the second-stage constraints (1.3) almost surely. It is evident that K_2 is convex. According to [1, proof of Thm. 1], we have $x_1 \in K_2$ if for the set

$$\Gamma(s, x_1) = \{x_2 \in C_2 | f_{2i}(s, x_1, x_2) \leqq 0, i = 1, \cdots, m_2\}, \tag{1.8}$$

there is a bounded region B with $\Gamma(s, x_1) \cap B \neq \varnothing$ almost surely.

We shall call K_2 the *induced feasible set* for the first stage of P, as opposed to the *explicit constraint set* K_1.

Still another set is of interest in this connection. Let us say that a function $\theta \in \mathscr{L}_1^\infty(S, \Sigma, \sigma)$ is *singularly nonpositive*, if for every $\varepsilon > 0$, there exists a (measurable) set $T \subset S$, comprised of a finite number of atoms with respect to σ (or empty), such that $\theta(s) \leqq \varepsilon$ for almost every $s \in S \setminus T$. The reason for this terminology will become clear in the next section. The *singularly induced feasible set* K_2° is defined as the set of all $x_1 \in R^{n_1}$ such that there exists an $x_2 \in \mathscr{L}_{n_2}^\infty$ with $x_2(s) \in C_2$ almost surely and $f_{2i}(\cdot, x_1, x_2(\cdot))$ singularly nonpositive for $i = 1, \cdots, m_2$. Like K_1 and K_2, the set K_2° is convex. Obviously

$$K_2 \subset K_2^\circ, \tag{1.9}$$

but in general the sets are not equal. The relations between these two sets is investigated further in § 4.

The main result is the following. (ri C denotes the relative interior of a set C, i.e., the interior of C relative to the smallest affine set containing C [10, § 6].)

THEOREM 1. *Suppose that* P *is strictly feasible and* ri $K_1 \subset K_2^\circ$. *Then*

$$\inf \mathrm{P} = \max \mathrm{D}, \tag{1.10}$$

so that solutions to P *and* D *correspond to saddle points of the basic Lagrangian* L.

In the last section (§ 4) of this paper we pursue the implications of this result and the significance of the hypothesis ri $K_1 \subset K_2^\circ$. We note, however, that this hypothesis is automatically satisfied whenever

$$K_2 \supset K_1. \tag{1.11}$$

Stochastic programs satisfying this last condition are known as stochastic programs with *relatively complete recourse*. Strictly speaking, this is the version of that condition for the class of stochastic programs under consideration here.

This is not an unusual property for stochastic programs. In fact, we might expect that for many stochastic programs arising from specific applications a stronger property will actually be satisfied, namely, the so-called *complete recourse* condition, which requires that for all $x_1 \in R^{n_1}$, there exists $x_2 \in \mathscr{L}_{n_2}^\infty$ satisfying the second stage constraints (1.3), or equivalently that $K_2 = R^{n_1}$; this implies that for all x_1, $\Gamma(s, x_1) \neq \varnothing$ almost surely.

The seminal papers on stochastic programming of G. Dantzig [4] and Beale [5] consider only stochastic programs with complete recourse. This restriction is not artificial, since the applications envisaged by these authors fall in this class. Actually, Beale's model [5, § 5] and one of the problems motivating Dantzig's work, described in [6], belong to an even more restrictive class, known as stochastic programs with *simple recourse*, which has received considerable attention (cf. [7] for a survey). Roughly speaking, for simple recourse the recourse decision is simply a way to record the "state of the system" after a first stage decision x_1 has been selected and a particular element s of S has been observed.

The term "*complete*" was first utilized by G. Dantzig in [4]. The more detailed classification sketched out above was introduced in [8]. Interest in the class of stochastic programs with relatively complete recourse—but not necessarily complete recourse—stems from theoretical considerations, but also from the

observation made in § 4 of [8] that some important allocation problems arising in agricultural economics and formulated by G. Tintner [9] are indeed members of this class and not of the more restrictive class of stochastic programs with complete recourse. Independently of the implications resulting from the theory developed here, stochastic programs with relatively complete recourse are also of interest from a computational viewpoint, since they usually possess special structures which can be exploited in the solution procedure; see, for example, [8, §§ 2 and 4].

2. Singular multipliers and induced feasibility. As in [3], we denote by Y_0° the set of all $y^\circ = (y_1^\circ, \cdots, y_{m_2}^\circ)$ such that y_i° is a nonnegative singular linear functional on $\mathscr{L}_1^\infty$. The latter means that y_i° is a continuous linear functional with $y_i^\circ(c) \geqq 0$ for every nonnegative $c \in \mathscr{L}_1^\infty$, and there exists an increasing sequence of measurable sets S_k with $\bigcup_{k=1}^\infty S_k = S$, such that $y_i^\circ(c) = 0$ if $c(s) = 0$ almost surely for $s \notin S_k$.

The *extended Lagrangian* associated with P is the function $\tilde{L}$ on $X_0 \times (Y_0 \times Y_0^\circ)$ defined by

$$\tilde{L}(x_1, x_2, y_1, y_2, y^\circ) = L(x_1, x_2, y_1, y_2) + \sum_{i=1}^{m_2} y_i^\circ(f_{2i}(\cdot, x_1, x_2(\cdot))). \tag{2.1}$$

The *extended dual problem* is

$\tilde{\mathrm{D}}$ maximize $\tilde{g}(y_1, y_2, y^\circ)$ over all $(y_1, y_2, y^\circ) \in Y_0 \times Y_0^\circ$, where

$$\tilde{g}(y_1, y_2, y^\circ) = \inf_{(x_1, x_2) \in X} \tilde{L}(x_1, x_2, y_1, y_2, y^\circ).$$

We have

$$\tilde{g}(y_1, y_2, 0) \equiv g(y_1, y_2), \tag{2.2}$$

so that D can be regarded as a "subproblem" of $\tilde{\mathrm{D}}$.

It was shown in [3] that strict feasibility in P implies inf P = max $\tilde{\mathrm{D}}$. We shall demonstrate in the next section that, in some cases, solving $\tilde{\mathrm{D}}$ is equivalent to solving D, and this will yield Theorem 1. The present section paves the way to this argument by developing a representation of the singularly induced feasible set K_2° in terms of the singular component of $\tilde{L}$ in (2.1). This representation, in the theorem which follows, explains the name we have given to K_2°.

THEOREM 2. *One has $x_1 \in K_2^\circ$ if and only if there exists $x_2 \in \mathscr{L}_{n_2}^\infty$ such that $x_2(s) \in C_2$ almost surely and*

$$\sum_{i=1}^{m_2} y_i^\circ(f_{2i}(\cdot, x_1, x_2(\cdot))) \leqq 0 \quad \textit{for all } y^\circ \in Y_0^\circ. \tag{2.3}$$

Proof. Clearly, the theorem will be proved if we establish that a function $\theta \in \mathscr{L}_1^\infty$ is singularly nonpositive if and only if $b^\circ(\theta) \leqq 0$ for every nonnegative singular functional b°.

Suppose first that θ is singularly nonpositive, and let b° be a nonnegative singular functional with an associated sequence of sets S_k, as per definition. Let $\varepsilon > 0$. Then there exists $T \subset S$, consisting of a finite number of atoms, such that $\theta(s) \leqq \varepsilon$ almost surely outside of T. Since $S_k \uparrow S$, we have $\sigma(S_k) \uparrow 1$. Hence for some k sufficiently large we have $S_k \supset T$ (except possibly for a subset of T of

measure zero), implying that $b^\circ(\theta)$ depends only on the restriction of θ to $S \setminus T$. Let e be the function in $\mathscr{L}_1^\infty$ with $e(s) \equiv 1$. Then $b^\circ(\theta) \leqq b^\circ(\varepsilon e) = \varepsilon b^\circ(e)$, because b° is nonnegative and $\theta(s) \leqq \varepsilon e(s)$ almost surely on $S \setminus T$. This is true for arbitrary $\varepsilon > 0$, so we conclude $b^\circ(\theta) \leqq 0$.

Assume now that the function $\theta \in \mathscr{L}_1^\infty$ is not singularly nonpositive. Thus for a certain $\varepsilon > 0$ the set

$$T = \{s \in S | \theta(s) > \varepsilon\}$$

is *not* comprised of a finite number of atoms (up to a set of measure zero). We shall construct a nonnegative singular functional b° such that $b^\circ(\theta) \geqq \varepsilon$. The assumed property of T implies the existence of a decreasing sequence of measurable sets $T_k \subset T$ such that $\sigma(T_k) > 0$ for all k and $\sigma(T_{k+1}) \leqq \frac{1}{2}\sigma(T_k)$. Then

$$0 = \lim_{k \to \infty} \sigma(T_k) = \sigma\Big(\bigcap_{k=1}^{\infty} T_k\Big).$$

Deleting the null set $T_\infty = \bigcap_{k=1}^\infty T_k$ from each set in the sequence, if necessary, we can suppose that $\bigcap_{k=1}^\infty T_k = \varnothing$. For each k, let b_k be the nonnegative linear functional on $\mathscr{L}_1^\infty$ defined by

$$b_k(c) = \frac{1}{\sigma(T_k)} \int_{T_k} c(s)\sigma(ds). \tag{2.4}$$

Observe that

$$b_k(e) = \|b_k\| = 1 \quad \text{for all } k, \tag{2.5}$$

where, as above, $e(s) \equiv 1$. The set $\{b_k | k = 1, 2, \cdots\}$ is thus bounded in the dual space $(\mathscr{L}_1^\infty)^*$ and hence has an accumulation point in the weak* topology. Let b° denote any such point. Then b° is again nonnegative, and $b^\circ(e) = 1$ by (2.5). Moreover, b° is singular: setting $S_k = S \setminus T_k$, we have $S = \bigcup_{k=1}^\infty S_k$, and for $l \geqq k$ the functional b_l has $b_l(c) = 0$ for all $c \in \mathscr{L}_1^\infty$ vanishing almost surely outside of S_k; thus $b^\circ(c) = 0$ for all $c \in \mathscr{L}_1^\infty$ vanishing almost surely outside of S_k. In particular, for $c(s) = \max\{\theta(s) - \varepsilon e(s), 0\} - [\theta(s) - \varepsilon e(s)]$ we have $c(s) = 0$ for all $s \in T$, and hence $b^\circ(c) = 0$. Therefore

$$b^\circ(\theta) - \varepsilon = b^\circ(\theta - \varepsilon e) = b^\circ(\max\{\theta - \varepsilon e, 0\}) \geqq 0,$$

and the proof is finished.

3. Equivalence of D and $\tilde{\mathbf{D}}$. We consider now, as in the extended Kuhn–Tucker conditions in [3], the function l on $R^{n_1} \times Y_0^\circ$ defined by

$$l(x_1, y^\circ) = \inf\left\{ \sum_{i=1}^{m_2} y_i^\circ(f_{2i}(\cdot, x_1, x_2(\cdot))) \,\middle|\, x_2 \in \mathscr{L}_{n_2}^\infty,\ x_2(s) \in C_2 \text{ a.s.} \right\}. \tag{3.1}$$

This is convex in x_1, concave in y°, and nowhere $+\infty$. Let

$$K_2^{\circ\circ} = \{x_1 \in R^{n_1} | l(x_1, y^\circ) \leqq 0 \text{ for all } y^\circ \in Y_0^\circ\}. \tag{3.2}$$

This is a closed convex set in R^{n_1}. (Each of the functions $l(\cdot, y^\circ)$ for $y^\circ \in Y_0^\circ$,

convex on R^{n_1} and nowhere $+\infty$, is continuous.) Moreover

$$K_2^\circ \subset K_2^{\circ\circ}, \tag{3.3}$$

in view of Theorem 2.

THEOREM 3. *Suppose there exists at least one* $x_1 \in C_1$ *with* $f_{1i}(x_1) < 0$ *for* $i = 1, \cdots, m_2$, *and that every such* x_1 *which is also in* ri C_1 *belongs to* $K_2^{\circ\circ}$. *Then the dual problems* D *and* $\tilde{\mathrm{D}}$ *are equivalent, in the sense that for every* $(y_1, y_2, y^\circ) \in Y_0 + Y_0^\circ$ *there exists* y_1' *such that* $(y_1', y_2) \in Y_0$ *and*

$$\tilde{g}(y_1, y_2, y^\circ) \leqq \tilde{g}(y_1', y_2, 0) = g(y_1', y_2). \tag{3.4}$$

Proof. Let $(y_1, y_2, y^\circ) \in Y_0 \times Y_0^\circ$. We assume $\tilde{g}(y_1, y_2, y^\circ)$ is not $-\infty$ (and hence is finite), since otherwise the conclusion of the theorem is trivial. In this case we have the following formula:

$$\tilde{g}(y_1, y_2, y^\circ) = \inf_{(x_1, x_2) \in X_0} \{L(x_1, x_2, y_1, y_2) + l(x_1, y^\circ)\}. \tag{3.5}$$

To see this, fix $(y_1, y_2, y^\circ) \in Y_0 \times Y_0^\circ$, and observe that for all $x_1 \in C_1$ we have that

$$\inf_{x_2 \in D} \left\{ \int_S L_2(s, x_1, x_2(s), y_2(s))\sigma(ds) + \sum_{i=1}^{m_2} y_i^\circ(f_{2i}(\cdot, x_1, x_2(\cdot))) \right\}$$

$$= \inf_{x_2 \in D} \int_S L_2(s, x_1, x_2(s), y_2(s))\sigma(ds) + \inf_{x_2 \in D} y_i^\circ(f_{2i}(\cdot, x_1, x_2(\cdot))),$$

where

$$\mathcal{D} = \{x_2 \in \mathcal{L}_{n_2}^\infty \mid x_2(s) \in C_2 \text{ almost surely}\}.$$

Since the inequality $\geqq$ certainly holds, equality will follow if we show that for arbitrary $x_2' \in \mathcal{D}$, $x_2'' \in \mathcal{D}$ and $\varepsilon > 0$, there exists $x_2 \in \mathcal{D}$ such that

$$\begin{aligned} &\int_S L_2(s, x_1, x_2(s), y_2(s))\sigma(ds) + \sum_{i=1}^{m_2} y_i^\circ(f_{2i}(\cdot, x_1, x_2(\cdot))) \\ &\leqq \int_S L_2(s, x_1, x_2''(s), y_2(s))\sigma(ds) + \sum_{i=1}^{m_2} y_i^\circ(f_{2i}(\cdot, x_1, x_2'(\cdot))) + \varepsilon. \end{aligned} \tag{3.6}$$

Now to each singular functional y_i°, there correspond an increasing sequence of measurable sets S_{ik} with $\cup S_{ik} = S$, such that $y_i^\circ(a) = 0$ if for some k, the function $a \in \mathcal{L}_1^\infty$ vanishes a.e. outside S_{ik}. The latter property implies that $y_i^\circ(b) = y_i^\circ(b')$ if b and b' agree almost everywhere outside of S_{ik}. Now for each index k define

$$x_s^k(s) = \begin{cases} x_s''(s) & \text{if } s \in S_{ik} \quad \text{for } i = 1, \cdots, m_2, \\ x_2'(s) & \text{for all other } s. \end{cases}$$

For each k, the function $x_2^k \in \mathcal{D}$ and

$$f_{2i}(s, x_1, x_2^k(s)) = f_{2i}(s, x_1, x_2'(s)) \quad \text{if } s \notin S_{ik}$$

so that

$$\sum_{i=1}^{m_2} y_i^\circ(f_{2i}(\cdot, x_1, x_2^k(\cdot))) = \sum_{i=1}^{m_2} y_i^\circ(f_{2i}(\cdot, x_1, x_2'(\cdot))).$$

On the other hand, since $\lim_{k\to+\infty}\sigma(S\setminus S_{ik})=0$, we get that

$$\lim_{k\to\infty}\int_S L_2(s,x_1,x_2^k(s),y_2(s))\sigma(ds)=\int_S L_2(s,x_1,x_2''(s),y_2(s))\sigma(ds).$$

From the two preceding equalities, it follows that (3.6) holds for $x_2=x_2^k$ if k is sufficiently large, which in turn directly yields (3.5).

Now, define the functions h and k on R^{n_1} by

$$\begin{aligned} h(x_1)&=\begin{cases}\inf\{L(x_1,x_2,y_1,y_2)\,|\,x_2\in\mathscr{L}^\infty_{n_2},\ x_2(s)\in C_2 \text{ a.s.}\} & \text{if } x_1\in C_1,\\ +\infty & \text{if } x_1\notin C_1,\end{cases}\\ k(x_1)&=-l(x_1,y^\circ).\end{aligned}\tag{3.7}$$

Then h is a convex function, not identically $+\infty$, while k is a concave function, nowhere $-\infty$, and

$$\tilde g(y_1,y_2,y^\circ)=\inf_{x_1\in R^{n_1}}\{h(x_1)-k(x_1)\}.\tag{3.8}$$

The finiteness of $\tilde g(y_1,y_2,y^\circ)$ implies k cannot be identically $+\infty$, and hence k is finite everywhere; furthermore h cannot have the value $-\infty$ and hence is proper. Fenchel's duality theorem [10, Thm. 31.1] is thus applicable to (3.8), and we obtain

$$\tilde g(y_1,y_2,y^\circ)=\max_{x_1^*\in R^{n_1}}\{k^*(x_1^*)-h^*(x_1^*)\},\tag{3.9}$$

where the conjugate functions k^* and h^* are defined by

$$h^*(x_1^*)=\sup_{x_1\in R^{n_1}}\{x_1\cdot x_1^*-h(x_1)\},\tag{3.10}$$

and

$$k^*(x_1^*)=\inf_{x_1\in R^{n_1}}\{x_1\cdot x_1^*-k(x_1)\}.\tag{3.11}$$

Fix x_1^* for which the maximum in (3.9) is attained. Then

$$-h^*(x_1^*)=\tilde g(y_1,y_2,y_0)-k^*(x_1^*),\tag{3.12}$$

and therefore by formula (3.10),

$$h(x_1)-x_1\cdot x_1^*\geqq\tilde g(y_1,y_2,y^\circ)-k^*(x_1^*)\quad\text{for all } x_1\in R^{n_1}.\tag{3.13}$$

Also from the definition of k and by formula (3.11),

$$l(x_1,y^\circ)+x_1\cdot x_1^*\geqq k^*(x_1^*)\quad\text{for all } x_1\in R^{n_1}.\tag{3.14}$$

The latter implies that $x_1\cdot x_1^*\geqq k^*(x_1^*)$ if $l(x_1,y^\circ)\leqq 0$, and thus, in particular, if $x_1\in K_2^{\circ\circ}$. Our hypothesis then yields that $x_1\cdot x_1^*\geqq k^*(x_1^*)$ for all x_1 in the set

$$K_1'=\{x_1\in\operatorname{ri}C_1\,|\,f_{1i}(x_1)<0,\ i=1,\cdots,m_1\}.\tag{3.15}$$

Define

$$K_1'' = \{x_1 \in C_1 | f_{1i}(x_1) < 0,\ i = 1, \cdots, m_1\}. \tag{3.16}$$

By hypothesis, K_1'' is nonempty. From this (and the finiteness, hence continuity, of the convex functions f_{1i}) it follows that $K_1' = \operatorname{ri} K_1''$, while on the other hand,

$$\operatorname{cl} K_1'' = \{x_1 \in C_1 | f_{1i}(x_1) \leqq 0,\ i = 1, \cdots, m_1\} = K_1. \tag{3.17}$$

Hence K_1 is in fact the closure of the set K_1', where the inequality $x_1 \cdot x_1^* \geqq k^*(x_1^*)$ holds, so that

$$k^*(x_1^*) \leqq \inf_{x_1 \in K_1} x_1 \cdot x_1^*. \tag{3.18}$$

The right side of (3.18) represents an ordinary convex program which, by our hypothesis, is strictly feasible. In consequence, there exist multipliers $\bar{y}_{1i} \geqq 0$, $i = 1, \cdots, m_1$, such that

$$k^*(x_1^*) \leqq \inf_{x_1 \in C_1} \left\{x_1 \cdot x_1^* + \sum_{i=1}^{m_1} \bar{y}_{1i} f_{1i}(x_1)\right\}.$$

The latter is better expressed, for our purposes, as

$$\sum_{i=1}^{m_1} \bar{y}_{1i} f_{1i}(x_1) \geqq k^*(x_1^*) - x_1 \cdot x_1^* \quad \text{for all } x_1 \in C_1. \tag{3.19}$$

Combining this inequality with (3.13) and reverting to the definition (3.7) of h, we see that

$$L(x_1, x_2, y_1, y_2) + \sum_{i=1}^{m_1} \bar{y}_{1i} f_{1i}(x_1) \geqq \tilde{g}(y_1, y_2, y^\circ) \quad \text{for all } (x_1, x_2) \in X_0. \tag{3.20}$$

But the left side of (3.20) is $L(x_1, x_2, y_1 + \bar{y}_1, y_2)$. Therefore, setting $y_1' = y_1 + \bar{y}_1$ we have $(y_1', y_2) \in Y_0$ and

$$\tilde{g}(y_1, y_2, y^\circ) \leqq \inf_{(x_1, x_2) \in X_0} L(x_1, x_2, y_1', y_2) = g(y_1', y_2),$$

which is the desired relation.

Proof of Theorem 1. Since P is strictly feasible, we know that $\inf \mathrm{P} = \max \tilde{\mathrm{D}}$ [3, Thm. 2], and also that the set K_1'', as defined in (3.16), is nonempty. But then, as in the proof above, the set K_1' in (3.15) is $\operatorname{ri} K_1''$ while $\operatorname{cl} K_1'' = K_1$. Therefore

$$\operatorname{ri} K_1 = \operatorname{ri}(\operatorname{cl} K_1'') = \operatorname{ri} K_1'' = K_1'.$$

Our assumption that $\operatorname{ri} K_1 \subset K_2^\circ$ then gives us, by way of (3.3), that $K_1' \subset K_2^{\circ\circ}$. Thus the hypothesis of Theorem 3 is fulfilled, yielding the conclusion that $\max \tilde{\mathrm{D}} = \max \mathrm{D}$.

4. Analysis of induced feasibility. We turn now to investigating further the relations between the induced feasible set K_2, the singularly induced feasible set K_2° and a related set K_2^σ, which consists of all vectors $x_1 \in R^{n_1}$ such that for almost all $s \in S$ there exists a vector $x_2 \in C_2 \subset R^{n_2}$ such that

$$f_{2i}(s, x_1, x_2) \leqq 0 \quad \text{for } i = 1, \cdots, m_2. \tag{4.1}$$

We shall call K_2^σ the σ-induced feasible set. It is evident that

$$K_2^\sigma \supset K_2.$$

One can view K_2^σ as the set of all (first-stage) decisions x_1 with which we can associate at least one feasible recourse decision for almost any "observed value" of s in S. In order for x_1 to be also in K_2, one must be able to string these recourse decisions together so as to form an essentially bounded measurable function of s.

The singularly induced feasible set K_2° is not so easily amenable to physical interpretation. However, the main results do not refer to K_2 but to the larger set K_2^σ, or even (in Theorem 3) to a still larger set $K_2^{\circ\circ}$. At least in part, this is due to technical reasons which we examine in this section. We concentrate our attention on two "extreme" cases: at one end the *discrete case*, where the support of the random variable consists of a *finite* number of atoms, and at the other end the *nonatomic case*, where the probability space contains *no* atoms. (This latter case includes the one of $S \subset R^N$, N finite, and σ absolutely continuous with respect to Lebesgue measure). These two situations seem to cover nearly all applications of practical interest. By abuse of language we shall refer to (S, Σ, σ) as being a discrete or nonatomic probability space in the respective cases.

Recall that for $s \in S$ and $x_1 \in R^{n_1}$ one has

$$\Gamma(s, x_1) = \{x_2 \in C_2 | f_{2i}(s, x_1, x_2) \leqq 0 \text{ for } i = 1, \cdots, m_2\}. \tag{4.2}$$

As already pointed out in [1, Proof of Thm. 1], the multifunction

$$s \mapsto \Gamma(s, x_1)$$

is measurable. This follows from [11, Corollary 4.3], since for fixed x_1 the functions

$$(s, x_2) \mapsto f_{2i}(s, x_1, x_2) \quad \text{for } i = 1, \cdots, m_2$$

are normal convex integrands [12, Lemma 2]. Thus for each $x_1 \in R^{n_1}$, the set

$$\omega(x_1) = \{s \in S | \Gamma(s, x_1) \neq \varnothing\} \tag{4.3}$$

is a measurable set. Moreover if $x_1 \in K_2^\sigma$, then $\omega(x_1)$ is a set of measure 1, i.e., $\sigma[\omega(x_1)] = 1$. We also define

$$\omega^{-1}(s) = \{x_1 \in R^{n_1} | \Gamma(s, x_1) \neq \varnothing\}, \tag{4.4}$$

which is clearly a convex set. With this notation we have that

$$K_2^\sigma = \{x_1 \in R^{n_1} | \sigma[\omega(x_1)] = 1\}. \tag{4.5}$$

PROPOSITION. *Suppose that for all s in S, $\omega^{-1}(s)$ is closed. Then the σ-induced feasible set K_2^σ is closed and convex.*

Proof. It suffices to show that the σ-induced feasible set can be written as

$$K_2^\sigma = \bigcap_{s \in S'} \omega^{-1}(s), \tag{4.6}$$

where S' is some subset of S of measure 1. The proposition is clearly true if $K_2^\sigma = \varnothing$. Assume otherwise and let D be a countable dense subset of K_2^σ. Such a set exists, since K_2^σ is a subset of the separable metric space R^{n_1}. Take S'

$= \bigcap_{x_1 \in D} \omega(x_1)$. Clearly $\sigma(S') = 1$ and $K_2^\sigma \supset \bigcap_{s \in S'} \omega^{-1}(s)$. Now for all $s \in S'$, we also have that $\omega^{-1}(s) \supset D$ and thus $\omega^{-1}(s) \supset K_2^\sigma$ since $\omega^{-1}(s)$ is closed, i.e., $K_2^\sigma \subset \bigcap_{s \in S'} \omega^{-1}(s)$.

COROLLARY A. *Suppose that C_2 is compact. Then K_2^σ is closed and convex.*

Proof. In this case, $\omega^{-1}(s)$ is closed for every $s \in S$, since C_2 is compact and the functions $f_{2i}(s, \cdot, \cdot)$ are lower semicontinuous.

COROLLARY B ([13, Thm. 3.5]). *Suppose that C_2 is polyhedral and that for $i = 1, \cdots, m_2$ and all $s \in S$ the functions $(x_1, x_2) \mapsto f_{2i}(s, x_1, x_2)$ are affine. Then K_2^σ is closed and convex.*

Proof. For each fixed s, the set

$$W(s) = \{(x_1, x_2) | f_{2i}(s, x_1, x_2) \leqq 0 \quad \text{for } i = 1, \cdots, m_2, x_1 \in R^{n_1}, x_2 \in C_2\}$$

is a polyhedral convex set, and its projection in the x_2-coordinates is $\omega^{-1}(s)$. Thus $\omega^{-1}(s)$ is polyhedral convex and consequently closed.

With some additional assumptions, it is also possible to show that $K_2^\sigma = \bigcap_{s \in S} \omega^{-1}(s)$. This essentially requires embedding S in a topological space (with S then the support of σ) and subjecting the maps $s \mapsto f_{2i}(s, x_1, x_2)$ to continuity conditions (cf. [14, Thm. 2]).

The following two theorems establish the relations between the various induced feasible sets in the discrete and nonatomic cases.

THEOREM 4. *Suppose that (S, Σ, σ) is a discrete probability space. Then*

$$R^{n_1} = K_2^\circ = K_2^{\circ\circ} \supset K_2 = K_2^\sigma. \tag{4.7}$$

Proof. When (S, Σ, σ) is a discrete probability space, every function in $\mathscr{L}^\infty$ is singularly nonpositive, since the criterion for singular nonpositivity allows us to ignore a finite number of atoms; thus $K_2^\circ = R^{n_1}$. The first string of equalities now follows from the known inclusions $K_2^\circ \subset K_2^{\circ\circ} \subset R^{n_1}$. The equality of $K_2 = K_2^\sigma$ is a direct consequence of the definition of these sets when the underlying probability space is discrete.

THEOREM 5. *Suppose that (S, Σ, σ) is a nonatomic probability space. Then*

$$K_2 = K_2^\circ. \tag{4.8}$$

Moreover, if to every $x_1 \in K_2^\sigma$ there corresponds a bounded region $B \subset R^{n_2}$ such that for almost all s, $\Gamma(s, x_1) \cap B \neq \varnothing$, then

$$K_2^\sigma = K_2 = K_2^\circ. \tag{4.9}$$

Proof. When (S, Σ, σ) is nonatomic, a function in $\mathscr{L}^\infty$ is singularly nonpositive if and only if it is nonpositive. This yields (4.8). We have already observed that, in general, $K_2^\sigma \supset K_2$. Thus to prove (4.9) it only remains to show inclusion in the other direction. Fix $x_1 \in K_2^\sigma$. The multifunction $s \mapsto \Gamma(s, x_1)$ is closed-convex-valued and measurable, and thus the multifunction $s \mapsto \Gamma(s, x_1) \cap \operatorname{cl} B$ is compact-convex-valued and measurable. Furthermore, by assumption, $\Gamma(s, x_1) \cap \operatorname{cl} B$ is almost surely nonempty. Thus there exists a measurable selector $\bar{x}_2$ with $\bar{x}_2(s) \in \Gamma(s, x_1) \cap \operatorname{cl} B$ for almost all s [12, Cor. 1.1]. Since B is bounded, $\bar{x}_2$ is in $\mathscr{L}_{n_2}^\infty$; hence $x_1 \in K_2$ and consequently $K_2^\sigma \subset K_2$.

These two theorems have immediate implications as to the class of dual variables we need to consider in obtaining an inf-max duality theorem.

COROLLARY 4. *Suppose that* P *is strictly feasible and* (S, Σ, σ) *is a discrete probability space (finitely many points). Then*

$$\inf P = \max D. \tag{4.10}$$

Proof. Theorems 4 and 1.

COROLLARY 5A. *Suppose that* (S, Σ, σ) *is a nonatomic probability space. Then* $x_1 \in K_2$ *if and only if there exists* $x_2 \in \mathscr{L}^\infty_{n_2}$ *such that* $x_2(s) \in C_2$ *almost surely and*

$$\sum_{i=1}^{m_2} y_i^\circ(f_{2i}(\cdot, x_1, x_2(\cdot)) \leqq 0 \quad \text{for all } y^\circ \in Y_0^\circ. \tag{4.11}$$

Proof. Theorems 5 and 2.

COROLLARY 5B. *Suppose that* P *is strictly feasible,* (S, Σ, σ) *is a nonatomic probability space, and to each* $x_1 \in K_2^\sigma$ *there corresponds a bounded region* B *with* $\Gamma(s, x_1) \cap B \neq \varnothing$ *almost surely. Suppose also that* $\omega^{-1}(s)$ *is closed for all* $s \in S$. *Then* ri $K_1 \subset K_2^\circ$ *if and only if* P *is a stochastic program with relatively complete recourse, in which case*

$$\inf P = \max D.$$

Proof. Theorem 5 with the Proposition above and Theorem 1.

COROLLARY 5C. *Suppose that* P *is a stochastic program with relatively complete recourse, strictly feasible with* C_2 *compact and* (S, Σ, σ) *is nonatomic. Then*

$$\inf P = \max D.$$

Proof. Corollary 5B with Corollary A of the above Proposition.

One of the implications of Corollaries 5B and 5C is that under those assumptions K_2 and K_2° are closed.

Corollaries 5A and 5B assert that when (S, Σ, σ) is nonatomic, the "singular multipliers" result from the presence of induced constraints. The singular multipliers y_2° appearing in the extended Kuhn–Tucker conditions [3] correspond—figuratively speaking—to a singular subset T of S which determines the critical points in S. These multipliers can not be $\mathscr{L}^1$ functions, since these critical points have mass 0, yet they do play a crucial role in the optimization problem.

On the other hand, if (S, Σ, σ) is discrete, Corollary 4 indicates that we never need to use "singular multipliers" to obtain the strong form of the duality result. Thus the basic Kuhn–Tucker conditions [2] are in fact *necessary* and sufficient, assuming strict feasibility. This does not mean that we can ignore the induced constraints, but more simply that the multipliers associated to these constraints will be represented by $\mathscr{L}^1$ functions on the probability space. (In the discrete case the dual of $\mathscr{L}^\infty_{n_2}$ *is* $\mathscr{L}^1_{n_2}$.) We illustrate this by a couple of examples.

Example 1. Find $x_1 \in R^{n_1}$, $x_2 \in \mathscr{L}^\infty_1$ such that

$$x_1 \geqq 0,$$

$$x_2(s) \geqq 0 \quad \text{and} \quad s - x_1 + x_2(s) \leqq 0 \quad \text{for almost all } s,$$

and one has the minimum of the expression

$$2x_1 - \frac{1}{n}\sum_{s\in S} x_2(s),$$

where $S = \{s = (k-1)/n,\ k = 1, \cdots, n\}$ with $\sigma(s) = 1/n$. There are no first-stage constraints; $C_1 = \{x_1 | x_1 \geqq 0\}$. The induced feasible set is

$$K_2 = \{x_1 | x_1 \geqq 1\},$$

whereas $K_2^\circ = R$ (Theorem 4). From Corollary 4A we know that the basic Kuhn–Tucker conditions are necessary *and* sufficient for this problem. From the differentiable form of these conditions with C_1 and C_2 the nonnegative orthants, we obtain using [2, Cor. B] that a pair $((\bar{x}_1, \bar{x}_2), \bar{y}_2) \in (R \times \mathscr{L}_1^\infty) \times \mathscr{L}_1^1$ determines optimal solutions to the program (4.8), $\cdots$, (4.10) and its dual if there exists a function $\rho \in \mathscr{L}_1^1$ satisfying:

(a) $\bar{x}_1 \geqq 0$;

(b) $\bar{x}_2(s) \geqq 0$, $\bar{y}_2(s) \geqq 0$, $s - \bar{x}_1 + \bar{x}_2(s) \leqq 0$, $\bar{y}_2(s)[s - \bar{x}_1 + \bar{x}_2(s)] = 0$ for all $s \in S$;

($c_\oplus$) $2 \geqq (1/n)\sum_{s\in S} \rho(s)$ and $2x_1 = (\bar{x}_1/n)\sum_{s\in S} \rho(s)$;

($d_\oplus$) $\rho(s) = -\bar{y}_2(s)$, $\bar{y}_2(s) \geqq 1$ and $\bar{x}_2(s)[-1 + \bar{y}_2(s)] = 0$ for all $s \in S$.

One verifies easily that the values

$$\bar{x}_1 = 1, \qquad \bar{x}_2(s) = 1 - s \quad \text{for } s = \frac{k}{n-1}, \quad k = 0, 1, \cdots, n-1,$$

and

$$\bar{y}_2(s) = -\rho(s) = 1 \quad \text{for } s = \frac{k}{n}, \quad k = 0, 1, \cdots, n-2, \quad \bar{y}_2(1) = -\rho(1) = n+1$$

satisfy the above conditions. It is striking that the "price" $y_2(s)$ associated with the constraint

$$s - x_1 + x_2(s) \leqq 0$$

is much larger when $s = 1$ than when $s < 1$.

Example 2. We consider the same problem as in Example 1, except that the probability space is now nonatomic. Specifically: S is the interval [0, 1] and σ is the Lebesgue measure. As before, the induced feasible set is

$$K_2 = \{x_1 | x_1 \geqq 1\}.$$

This is also the singularly induced feasible set K_2° (Theorem 5), and as can be verified, it is also the set $K_2^{\circ\circ}$ defined by (3.2) and utilized in Theorem 3. Corollary 5A directs us to use in this case the extended Kuhn–Tucker conditions [3, § 5]. Thus, we have that a pair $((\bar{x}_1, \bar{x}_2), (\bar{y}_2, \bar{y}^\circ)) \in (R \times \mathscr{L}_1^\infty) \times (\mathscr{L}_1^\infty \times \mathscr{S}_1)$ determines optimal solutions to program (4.8), $\cdots$, (4.10) and its extended dual (with s uniform on [0, 1]) if there exists $\rho \geqq \in \mathscr{L}_1^1$ satisfying

(a) $\bar{x}_1 \geqq 0$;

(b) $\bar{x}_2(s) \geqq 0$, $\bar{y}_1(s) \geqq 0$, $s - \bar{x}_1 + \bar{x}_2(s) \leqq 0$, $\bar{y}_2(s)[s - \bar{x}_1 + \bar{x}_2(s)] = 0$ for $s \in [0, 1]$;

(c°) $\bar{x}_1$ minimizes $(2x_1 + \int \rho(s)\sigma(ds) + l(x_1, \bar{y}^\circ))$ subject to $x_1 \geqq 0$;

($d_\oplus$) $\rho(s) = -\bar{y}_2(s)$, $\bar{y}_2(s) \geqq 1$ and $\bar{x}_2(s)[-1 + \bar{y}_2(s)] = 0$ for $s \in [0, 1]$;

(e) $\bar{y}^\circ \geqq 0$, $\bar{y}^\circ(\cdot - \bar{x}_1 + \bar{x}_2(\cdot)) = 0$ and $0 = \inf\{\bar{y}^\circ(\cdot - \bar{x}_1 + \bar{x}_2(\cdot)) | x_2 \in \mathscr{L}_1^\infty \cdot([0, 1], \Sigma, \sigma), x_2(s) \geqq 0$ almost surely$\}$.

Conditions (a), (b) and ($d_\oplus$) are the same as before, but this time a term involving the singular multipliers $l(x_1, y^\circ)$ appears in (c°), and these multipliers must satisfy the condition (e). The functional $\bar{y}^\circ$ is a continuous linear functional on $\mathscr{L}_1^\infty$ and can be expressed as an integral with respect to a purely finitely additive measure ν on S. Let ν be the measure on S which assigns measure 1 to a set A if A is (Lebesgue) measurable and 1 is a point of density of A; otherwise the measure of A is 0. (Such a measure can be generated on the Borel field by a construction similar to the one used in the proof of Theorem 4.1 of [16] starting by simply specifying $\nu(B) = 0$ for every set B of Lebesgue measure 0 and $\nu(B) = 1$ if B is (relatively) open in [0, 1] and contains 1). One can verify that the values

$$\bar{x}_1 = 1, \quad \bar{x}_2(s) = 1 - s \quad \text{for } s \in [0, 1]$$

and

$$\bar{y}_2(s) = -\rho(s) = 1 \quad \text{for } s \in [0, 1] \quad \text{and} \quad \bar{y}^\circ(\cdot) = \int \cdot \, \nu(ds)$$

satisfy the above conditions.

The solutions to the problems in Examples 1 and 2 resemble each other in many ways, except for the presence in the case of Example 2 of the singular function $\bar{y}^\circ$, and on the other hand the "jump" in the $\bar{y}_2$ multiplier when $s = 1$ in the case of Example 1. In fact, if we allow n to go to $+\infty$ in Example 1, it is clear that $\bar{y}_2(1)$ also tends toward $+\infty$. In other words, in the limit there will be an "infinite" price associated with the second-stage constraint when $s = 1$. We know from the derivation in Example 2 that this unusual behavior at $s = 1$ is due to the presence of induced constraints. The relations between these two examples give an illustration of the content of Theorem 1 of [3].

One can also view Theorem 1 as an enticement to introduce the induced constraints explicitly among the first-stage constraints (1.2). If this is done, *every* stochastic program becomes a stochastic program with relatively complete recourse and Theorem 1 becomes applicable to *every* stochastic program.

This, however, requires the actual determination of these induced constraints. The general theory of optimization indicates that merely a finite number of these will be sufficient to represent the binding constraints at the minimum. But this is only of relative comfort since, in general, the constraints in question are not especially easy to identify. Practically, we expect that the appropriate constraints will be generated as needed. By this it is meant that the algorithm builder will use some test to verify if a given $x_1 \in K_1$ is or is not a member of K_2, and in the latter case he will generate certain induced constraints—to be added to the constraints determining K_1—which would "cut out" that particular x_1. This procedure is already used for stochastic *linear* programming [15, § 5], although in that case fairly complete and concrete characterizations of the induced feasible set K_2 are known [15, § 4].

We conclude this paper by illustrating the effect on the dual variables of introducing the induced constraints as first-stage constraints in the case of the examples appearing above.

Example 1′. Same as Example 1, except that the induced constraint

$$x_1 \geqq 1$$

is now explicitly introduced as a first-stage constraint. The same Kuhn–Tucker conditions yield optimality criteria, except that (a) must be changed to

(a′) $\bar{x}_1 \geqq 0, 1-\bar{x}_1 \leqq 0, \bar{y}_1 \geqq 0, (1-\bar{x}_1)\bar{y}_1 = 0.$

With this modification, it can be seen that the following yield optimal solution to P and its dual:

$$\bar{x}_1 = 1, \quad \bar{x}_2(s) = 1-s \quad \text{for } s \in S$$

and

$$\bar{y}_1 = 1, \quad \bar{y}_2(s) = -\rho(s) = 1 \quad \text{for } s \in S.$$

The "curious" behavior of $\bar{y}_2(s)$ at $s = 1$ in Example 1 has now disappeared.

Example 2′. Same as Example 2 except that the induced constraint is explicitly introduced as a first-stage constraint. The new problem 2′ is now a stochastic program with relatively complete recourse. We can thus turn to the *basic* Kuhn–Tucker conditions to obtain optimality criteria. They are (a′) as above, (b) and ($d_{\oplus}$) as in Example 2, but from [2, Cor. B] we also have

(c′) $2+\int \rho(s)\sigma(ds) \geqq 0$ and $x_1[2+\int \rho(s)\sigma(ds)] = 0.$

This shows that the values

$$x_1 = 1, \quad \bar{x}_2(s) = 1-s \quad \text{for } s \in [0, 1]$$

and

$$\bar{y}_1 = 1, \quad \bar{y}_2(s) = -\rho(s) = 1 \quad \text{for } s \in [0, 1]$$

yield optimal solutions to Example 2′ and its dual. Observe that the $(\bar{y}_2, \bar{y}^\circ)$ solution obtained in Example 2 is actually an optimal solution to the extended dual $\tilde{D}$ of Example 2′, but so is the solution obtained here (with $y^\circ = 0$), giving us a concrete illustration of Theorem 3.

If in P the set C_1 is replaced by $C_1 \cap K_2$ (or $C_1 \cap K_2^\circ$, or $C_1 \cap K_2^{\circ\circ}$), then every problem so generated is also a stochastic program with relatively complete recourse. But this time the relation between the dual variables associated with the original problem and those of the new problem is no longer as easy to establish.

Finally, we observe that from the proofs of Theorems 1 and 3 it follows that we could actually use the larger set $K_2^{\circ\circ}$ in place of K_2°. This gives a more general result, but $K_2^{\circ\circ}$ is at the same time "less concrete". We have not succeeded in proving any more intimate relationship between $K_2^{\circ\circ}$ and K_2° than the inclusion

$$K_2^{\circ\circ} \supset K_2^\circ,$$

except in the discrete case, when evidently equality holds.

5. Conclusion. The objective of [1], [2], [3] and this paper is to develop necessary and sufficient optimality conditions for stochastic convex programs. The model chosen P (see §1) demands that the recourse (or second-stage) decision as a function of the random elements be measurable (an inconsequential restriction) and essentially bounded. This last condition is a definite restriction, in general,

(not if the second-stage feasibility region is bounded [1, Thm. 2]) but it is not a significant restriction [1, Thm. 1] since the main concern is not with the existence of optimal solutions. The approach is through general duality theory: we first embed the original problem in a class of perturbed problems (the natural choice turns out to be to perturb the constraints by elements of $R^{m_1} \times \mathscr{L}^{\infty}_{m_2}$), then set up a *Lagrangian* L associated with the system of perturbations and finally from L derive a dual problem D. Saddle points of L are characterized by the so-called Kuhn–Tucker conditions. These Kuhn–Tucker conditions always provide sufficient optimality conditions for P; moreover they become also necessary if it can be shown that inf P = max D (and not just inf P = sup D). To guarantee the existence of optimal solutions to D, the standard requirement is to demand that P satisfies a constraint qualification (e.g., strict feasibility).

This is precisely what happens [3, Thm. 2] if the space associated with perturbations is sufficiently "large", viz., if the multiplier space is selected to be $R^{m_1} \times (\mathscr{L}^{\infty}_{m_2})^*$. The extended Kuhn–Tucker conditions [3, § 5] are then necessary and sufficient. The choice of $R^{m_1} \times (\mathscr{L}^{\infty}_{m_2})^*$ as the multiplier space is however rather unsatisfactory since calculations involving elements of $(\mathscr{L}^{\infty}_{m_2})^*$ are generally unmanageable unless one can handle "separately" the singular part and the $\mathscr{L}^1$-part of every such $(\mathscr{L}^{\infty}_{m_2})^*$ multiplier.

This paper shows that the singular parts of the optimal multipliers correspond basically to the induced constraints (Theorem 2), more precisely to the singularly induced feasibility set. Consequently, if there are no induced constraints (relatively complete recourse) or, more generally, if the induced constraints do not determine binding constraints at the optimum, we may restrict the multiplier space to $R^{m_1} \times \mathscr{L}^1_{m_2}$ and still obtain the necessity of the Kuhn–Tucker conditions (Theorem 1). Note also that every stochastic program can be transformed into a stochastic program with relatively complete recourse by the inclusion of the induced constraints in the first-stage constraints. In this case the basic duality theory [1, § 4] is applicable, and the necessary and sufficient conditions for optimality are given by the (basic) Kuhn–Tucker conditions [2] involving only $\mathscr{L}^1$-functions as multipliers.

REFERENCES

[1] R. T. ROCKAFELLAR AND R. WETS, *Stochastic convex programming: Basic duality*, Pacific J. Math., to appear.

[2] ———, *Stochastic convex programming: Kuhn–Tucker conditions*, J. Mathematical Economics, to appear.

[3] ———, *Stochastic convex programming: Extended duality and singular multipliers*, Pacific J. Math., to appear.

[4] G. B. DANTZIG, *Linear programming under uncertainty*, Management Sci., 1 (1955), pp. 197–206.

[5] E. BEALE, *On minimizing a convex function subject to linear inequalities*, J. Roy. Statist. Soc. Ser. B, 17 (1955), pp. 173–184.

[6] A. R. FERGUSON AND G. DANTZIG, *The allocation of aircraft to routes: An example of linear programming under uncertain demand*, Management Sci., 3 (1956), pp. 45–73.

[7] W. ZIEMBA, *Stochastic programs with simple recourse*, Mathematical Programming in Theory and Practice, P. Hammer and G. Zoutendijk, eds., North-Holland, Amsterdam, 1974.

[8] D. WALKUP AND R. WETS, *Stochastic programs with recourse: Special forms*, Proc. Princeton Symposium on Mathematical Programming, H. Kuhn, ed., Princeton University Press, Princeton, N. J., 1970, pp. 139–161.
[9] G. TINTNER, *A note on stochastic linear programming*, Econometrica, 28 (1960), pp. 490–495.
[10] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N. J., 1969.
[11] ———, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.
[12] ———. *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.
[13] D. WALKUP AND R. WETS, *Stochastic programs with recourse*, SIAM J. Appl. Math, 15 (1967), pp. 1299–1314.
[14] R. WETS, *Induced constraints for stochastic optimization problems*, Techniques of Optimization, A. Balakrishnan, ed., Academic Press, New York, 1972, pp. 433–443.
[15] ———, *Stochastic programs with fixed recourse: The equivalent deterministic program*, SIAM Rev., 16 (1974), pp. 309–339.
[16] K. YOSIDA AND E. HEWITT, *Finitely additive measures*, Trans. Amer. Math. Soc., 72 (1952), pp. 46–66.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

SINGULARLY PERTURBED OPTIMAL CONTROL PROBLEMS. I: CONVERGENCE*

PAUL BINDING†

Abstract. The problem studied is as follows: when does the full solution of minimizing $x^0(T)$, given

$$\dot{x}(t) = f(x(t), y(t), u(t)), \qquad u(t) \in U,$$
$$\varepsilon \dot{y}(t) = g(x(t), y(t), u(t)), \qquad 0 \leqq t \leqq T,$$

with boundary conditions on x and y, converge in some sense to the reduced solution of minimizing $x_0^0(T_0)$, given

$$\dot{x}_0(t) = f(x_0(t), y_0(t), u_0(t)), \qquad u_0(t) \in U,$$
$$0 = g(x_0(t), y_0(t), u_0(t)), \qquad 0 \leqq t \leqq T_0,$$

with boundary conditions on x_0 as $\varepsilon \to 0$? Without the minimization, this is a standard topic in o.d.e. theory which essentially covers the case where $u = u_0$ is smooth. The corresponding methods need considerable modification for the control problem and, in the end, are closer to those of optimal existence theory. Assuming Lipschitz dependent right sides for the full model, we see that various additional hypotheses give convergence in modes varying from weak L_1 to strong AC. In particular, if controls are prerestricted to a fixed compact set in L_1 (e.g., of uniformly bounded variation), or if the model is linear in y and the reduced solution is normal, then y and u converge in L_1 to y_0 and u_0 while x converges in AC to x_0, and thus the full optimal costs converge to the reduced one.

Introduction. There are many reasons for considering perturbations, among them the determination of a confidence (i.e., error) estimate for the use of a mathematical model which is in some sense oversimplified. Rather general theories have been advanced for the use of models which are averaged or those with smoothed coefficients, but the question of order reduction is less clear. Most of the known work concerns linear models, and the object here is to analyze nonlinear ordinary differential control problems in which some of the state variables are "parasitic" or "fast", i.e., evolve on a faster time scale than the others. Such effects were investigated long ago, e.g., by Prandtl for p.d.e. (partial differential equations) and Nagumo for o.d.e. (ordinary differential equations). In many cases, these fast variables are virtually constant outside their "boundary layer" of evolution; this leads to a "reduced" model involving the other, slow variables. The full model can then be viewed as the reduced one perturbed by the parasitic effects.

Such perturbations fit into a general class termed "singular". Singular perturbations have been applied a good deal of late to control problems by, e.g., Kokotovic and O'Malley, but such investigations have been confined mostly to linear dynamics with quadratic or time optimal cost functions and have made use of explicit formulas for the optimal controls.

* Received by the editors December 27, 1974, and in final revised form June 9, 1975.

† Department of Mathematics, Statistics and Computing Science, University of Calgary, Calgary, Alberta, Canada T2N 1N4. This work was supported by a grant from the National Research Council of Canada.

In the absence of such a priori knowledge, the singularly perturbed control problem possesses two differences from more conventional o.d.e. formulations. The first is the unsmoothness of the fast variables: in fact, the better behaved the full model is, in terms of differentiability and simplicity, the more the fast variables reflect the character of the *control* variables. When the latter are piecewise continuous, a "boundary layer" occurs at each discontinuity. In general, controls are only measurable and the resulting "perpetual" boundary layer requires different estimates. A piecewise theory is also developed here, but only for cases where controls are of a priori bounded variation. The second difference from the usual o.d.e. case is that the two-point boundary problems (b.p.) appearing in the control formulation are frequently known to be soluble. For example, suppose that the state variables must satisfy conditions at both ends of the trajectory; it is often the case that an improvement over a known (perhaps default) control strategy is desired, and that the resulting known trajectory does satisfy the end conditions. Similar remarks hold concerning solubility of the b.p. arising from the maximum principle. The works [10], [15] make use of this b.p., but here it is used sparingly, the approach being mostly via the model equations alone.

The first phase of the analysis, in §§ 2–4, simply gives convergence of the full trajectories to the reduced ones. Convergence can mean many things, of course, but the minimal objective here is that the slow (including cost) variables should converge uniformly. Existence theory also arises naturally here. The level of model smoothness used is Lipschitz, and the assumptions are discussed in detail in § 4. Section 2 contains the key estimates and is a contribution to the initial value o.d.e. theory under reduced smoothness assumptions.

The second phase, published separately as §§ 5–7, estimates the difference between full and reduced solutions. With bounded variation (rather than measurable, as earlier) controls, Lipschitz models can be controlled to differences in response (hence cost) of the same order as the ratio ε of slow and fast time constants. The further problems of (i) eliminating the boundary-layer discrepancy at the fast variable boundary conditions and (ii) reducing the response differences to a smaller order than ε are also tackled. Finally several of the results are exemplified in a problem with relay dynamics limited to a time constant $1/\varepsilon$.

1. Basic assumptions.

Summary. The full model d.e. are Lipschitz in the state variables, and the fast d.e. satisfy a stability assumption in the fast variables alone. An assumption of existence of (weak) minimizing solutions and their continuous dependence on nonsingular perturbations is made, but only for the reduced problem.

1.1. Problem specification. We are to minimize $x^0(T)$ subject to

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{y}_\varepsilon(t), \mathbf{u}(t)), \qquad \mathbf{u}(t) \in U \subset \mathbb{R}^m, \tag{1.1}$$

$$\mathbf{x}(0) \in X_0 = (\{0\}, \hat{X}_0)\,[=\{0\}\times\hat{X}_0], \qquad \mathbf{x}(T) \in X_1 = (\mathbb{R}, \hat{X}_1) \tag{1.2}$$

for almost all $t \in [0, T]$ and $\mathbf{u} \in L_1([0, T] \to \mathbb{R}^m)$. Here $\hat{X}_i \subset \mathbb{R}^n$, x^0 is the *cost* variable, the other x^i are the *slow* variables and $\mathbf{u}$ is the control. T is the least time for which the terminal condition (1.2) is satisfied; its dependence on $\mathbf{x}(0)$ and ε will

be restricted below. The *fast* disturbance $\mathbf{y}_\varepsilon$ satisfies

$$\mathbf{y}'(\tau) = \mathbf{g}(\mathbf{x}(\tau), \mathbf{y}(\tau), \mathbf{u}(\tau)), \quad \mathbf{y}_\varepsilon(t) = \mathbf{y}(\tau), \quad ' \equiv d/d\tau, \quad \tau = t/\varepsilon; \tag{1.3}$$

that is,

$$\varepsilon \dot{\mathbf{y}}_\varepsilon(t) = \mathbf{g}(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t)) \quad \text{a.e. } t \in [0, T], \tag{1.3*}$$

$$\mathbf{y}_\varepsilon(0) \in Y_0, \qquad \mathbf{y}_\varepsilon(T) \in Y_1, \tag{1.4}$$

where $Y_i \subset \mathbb{R}^k$ and ε is a small positive constant. For notational convenience we set

$$\mathbf{z} = (\mathbf{x}, \mathbf{y}), \quad Z_i = (X_i, Y_i), \quad \mathbf{h} = (\mathbf{f}, \mathbf{g}),$$

and we assume that Z_0 is bounded and $\mathbf{h}$ is independent of x^0.

We also assume for the moment that $Y_1 = \mathbb{R}^k$ and U is bounded; these simplify matters and amendments necessary to cope with more general situations are discussed in § 4. The cost function has been taken in Lagrange form but modification to general Bolza form (with $a \circ \mathbf{x}(T)$ instead) follows standard techniques (cf. [5, § 69]). Likewise no real difficulties are incurred in extending dependence of the end sets Z_i to t. The question of explicit dependence of the d.e. (1.1), (1.3) on t and τ is taken up in § 4.

1.2. Restrictions on the d.e. First we assume that

$$|\mathbf{f}(\xi, \nu, \rho) - \mathbf{f}(\xi_*, \nu_*, \rho)| \leqq \alpha|\xi - \xi_*| + \beta|\nu - \nu_*|, \tag{1.5}$$

$$|\mathbf{g}(\xi, \nu, \rho) - \mathbf{g}(\xi_*, \nu, \rho_*)| \leqq \gamma|\xi - \xi_*| + \delta|\rho - \rho_*|, \tag{1.6}$$

$$[\nu - \nu_*][\mathbf{g}(\xi, \nu, \rho) - \mathbf{g}(\xi, \nu_*, \rho)] \leqq -\kappa|\nu - \nu_*|^2, \qquad \kappa > 0, \tag{1.7}$$

where $|\cdot|$ denotes appropriate Euclidean norm and adjacent vectors are multiplied scalarly. Double bars will be used for function space norms, and vectors may be row or column depending on context. In general, italic letters denote functions and Greek ones denote constants and variables.

We further assume a uniform Lipschitz condition on $\mathbf{g}$ in ν (1.6) and continuity of $\mathbf{f}$ in ρ. These guarantee that the problem is well-posed in the following sense. For any admissible (see § 1.1) control $\mathbf{u}$ on $[0, T]$ and initial value $\mathbf{z}(0)$ there is just one absolutely continuous (AC) state function $\mathbf{z} = \mathbf{z}_\varepsilon = (\mathbf{x}_\varepsilon, \mathbf{y}_\varepsilon)$ satisfying (1.1) and (1.3) on $[0, T]$.

We also assume that the endtime $T = T_\varepsilon$ for an optimal solution is uniformly bounded in ε, say,

$$T_\varepsilon \leqq T_\infty. \tag{1.8}$$

Fixed time formulations are thus covered, and it also suffices if f^0 has a positive lower bound; the latter can be weakened [1] or removed altogether if the state variables are known to lie in a fixed compact set [6, p. 391], but state constraints will not be imposed here.

1.3. The reduced model. Formally setting $\varepsilon = 0$ in (1.1)–(1.3) gives

$$\dot{\mathbf{x}}_0(t) = \mathbf{f}(\mathbf{x}_0(t), \mathbf{y}_0(t), \mathbf{u}(t)), \tag{1.9}$$

$$\mathbf{x}_0(0) \in X_0, \qquad \mathbf{x}_0(T_0) \in X_1, \tag{1.10}$$

$$\mathbf{0} = \mathbf{g}(\mathbf{x}_0(t), \mathbf{y}_0(t), \mathbf{u}(t)), \tag{1.11}$$

which is the *reduced* model. It is shown in § 2.1 that the assumptions so far guarantee a unique reduced solution $\mathbf{z}_0 = (\mathbf{x}_0, \mathbf{y}_0)$ for each $\mathbf{u} \in L_1$. Note that $\mathbf{y}_0(0)$ and $\mathbf{y}_0(T_0)$ are now determined by (1.10) and (1.11).

We now make two well-posing assumptions on the reduced optimization problem; from these it will follow that for small enough ε the full model is similarly well-posed (e.g., has feasible solutions). First, there is an admissible triple $\mathbf{u}$, $\mathbf{x}_0(0)$, T_0 so that the terminal condition $\mathbf{x}_0(T) \in X_1$ (hence $\mathbf{z}_0(T) \in Z_1$) is satisfied.

The other assumption requires a preliminary comment. Eliminating $\mathbf{y}_0(t)$ from (1.9), (1.11) makes the resulting right side of (1.9) Lipschitz dependent on $x_0(t)$ (see § 2.1), so minimizing weak ("relaxed" [19]) reduced solutions $\mathbf{x}_0^*$ exist. Our assumption is now that any such $\mathbf{x}_0^*$ has uniformly continuously varying endpoints; i.e., given $\eta > 0$, there is $\theta > 0$ so that if $|\mathbf{x}(t) - \mathbf{x}_0^*(t)| < \theta$ for all t, then

$$\mathbf{x}(s) \in X_0, \qquad \mathbf{x}(T_0) \in X_1 \quad \text{with } |s| < \eta, \, |T_0 - T_0^*| < \eta. \tag{1.12}$$

Here T_0^* is the endtime for $\mathbf{x}_0^*$ which is extended over $\Lambda = [-\eta, T_0^* + \eta]$, say, by fixing $\mathbf{u}$ at some constant value, t ranges over Λ and $\mathbf{x} \in \mathrm{AC}$ on Λ.

There are standard conditions for (1.12), for example, nontangency of $\mathbf{x}_0^*$ to full-dimensional tangent cones to X_i at the endpoints (this is justified in § 6). Weaker conditions can be given using Lyapunov-like functions (cf. [11, Chap. 4]), or local controllability conditions as in [12] in case X_i are "thin".

2. Singularly perturbed o.d.e.

Summary. In § 2.1 we obtain $\mathbf{y}_0 \in L_\infty$ for each $\mathbf{u} \in L_1$ with $\mathbf{y}_0(t)$ Lipschitz dependent on $\mathbf{x}_0(t)$ from (1.11), and this guarantees unique reduced solutions $\mathbf{z}_0$. We then consider *fixed* measurable control $\mathbf{u}$ and endpoints $\mathbf{x}(0)$, $\mathbf{x}(T)$ and suppress them for most of the section. This corresponds to the o.d.e. case with right sides measurable in t. The basic result, that $\mathbf{z}_\varepsilon \to \mathbf{z}_0$ as $\varepsilon \to 0$, is obtained via three differential inequality estimates and features different norms, uniform for $\mathbf{x}$ and L_1 for $\mathbf{y}$. Section 2.2 contains the first two estimates for $\mathbf{x}_\varepsilon$ and $\mathbf{y}_\varepsilon$ and gives bounds which are uniform in ε and t. Using these and the third estimate of § 2.3, we see that $\mathbf{y}_\varepsilon$ are relatively compact in L_1 as ε varies, and the desired result follows from standard connections between L_1 and almost everywhere convergence in § 2.4.

2.1. Reduced solutions. From (1.7) and [16, Thm. 6.4.4] a unique function $\mathbf{e}$ exists so that

$$\mathbf{g}(\boldsymbol{\xi}, \mathbf{e}(\boldsymbol{\xi}, \boldsymbol{\rho}), \boldsymbol{\rho}) = \mathbf{0}. \tag{2.1}$$

Now let $\boldsymbol{\eta} = \mathbf{e}(\boldsymbol{\xi}, \boldsymbol{\rho}) - \mathbf{e}(\boldsymbol{\xi}_*, \boldsymbol{\rho}_*)$ so

$$\begin{aligned} \kappa|\boldsymbol{\eta}|^2 &\leqq |\boldsymbol{\eta}[-\mathbf{g}(\boldsymbol{\xi}_*, \mathbf{e}(\boldsymbol{\xi}, \boldsymbol{\rho}), \boldsymbol{\rho}_*)]| \\ &\leqq |\boldsymbol{\eta}|[\gamma|\boldsymbol{\xi} - \boldsymbol{\xi}_*| + \delta|\boldsymbol{\rho} - \boldsymbol{\rho}_*|], \end{aligned} \tag{2.2}$$

using (2.1) and its $_*$ equivalent, and (1.6), (1.7). With $\boldsymbol{\xi}_* = \boldsymbol{\rho}_* = 0$, we obtain

$$|\mathbf{e}(\boldsymbol{\xi}, \boldsymbol{\rho})| \leqq \kappa^{-1}(\gamma|\boldsymbol{\xi}| + \delta|\boldsymbol{\rho}|) + |\mathbf{e}(\mathbf{0}, \mathbf{0})|,$$

so $\mathbf{e}$ is bounded when $\boldsymbol{\xi}$ and $\boldsymbol{\rho}$ are.

LEMMA 2A. *For each admissible control* $\mathbf{u}$ *on* $[0, T]$, $T < \infty$, (1.9), (1.11) *has a unique solution continuable over* $[0, T]$ *with* $\mathbf{x}_0 \in AC$ *and* $\mathbf{y}_0 \in L_\infty$. *Further,* $\mathbf{y}_0(t)$ *has a Lipschitz constant* γ/κ *with respect to* $\mathbf{x}_0(t)$ *for each* t.

Proof. The Lipschitz constant γ/κ is immediate from (2.2), while measurability of $\mathbf{y}_0 : t \to \mathbf{e}(\mathbf{x}_0(t), \mathbf{u}(t))$ is trivial. Since U is bounded, the continuity and Lipschitz assumptions of § 1.2 show that

$$\mathbf{f}^r(\boldsymbol{\xi}, \mathbf{u}(t)) = \mathbf{f}(\boldsymbol{\xi}, \mathbf{e}(\boldsymbol{\xi}, \mathbf{u}(t)), \mathbf{u}(t)) \tag{2.3}$$

satisfies a Lipschitz condition in $\boldsymbol{\xi}$ and hence a bound of the form $O(|\boldsymbol{\xi}| + 1)$. Thus $\mathbf{x}_0 \in AC$ is unique and bounded over $[0, T]$ from standard exponential estimates [11, Chap. 2]. We now conclude that $\mathbf{y}_0$ is also uniformly bounded, and hence L_∞. Q.E.D.

2.2. Norm estimates. For the rest of this section, $\mathbf{u}$, T and $\mathbf{x}(0) = \mathbf{x}_\varepsilon(0)$ will be fixed and $\mathbf{u}$ suppressed in general. Define the norms

$$\|\mathbf{x}\| = \max\{|\mathbf{x}(t)| : 0 \leqq t \leqq T\}, \qquad \|\mathbf{y}\|_\pi = \left(\int_0^T |\mathbf{y}(t)|^\pi \, dt\right)^{1/\pi}, \tag{2.4}$$

where $\pi \geqq 1$, and let

$$b(t) = |\mathbf{f}(\mathbf{0}, \mathbf{0})|, \qquad c(t) = |\mathbf{g}(\mathbf{0}, \mathbf{0})|.$$

Then using (1.5) we have

$$\begin{aligned} \frac{d}{dt}|\mathbf{x}_\varepsilon(t)| \leqq |\dot{\mathbf{x}}_\varepsilon(t)| &\leqq |\mathbf{f}(\mathbf{x}_\varepsilon(t), \mathbf{y}_\varepsilon(t)) - \mathbf{f}(\mathbf{0}, \mathbf{y}_\varepsilon(t))| \\ &\quad + |\mathbf{f}(\mathbf{0}, \mathbf{y}_\varepsilon(t)) - \mathbf{f}(\mathbf{0}, \mathbf{0})| + b(t) \\ &\leqq \alpha|\mathbf{x}_\varepsilon(t)| + \beta|\mathbf{y}_\varepsilon(t)| + b(t), \end{aligned} \tag{2.5}$$

while from (1.6), (1.7) we have

$$\begin{aligned} \varepsilon|\mathbf{y}_\varepsilon(t)|\frac{d}{dt}|\mathbf{y}_\varepsilon(t)| &\leqq \frac{\varepsilon}{2}\frac{d}{dt}[\mathbf{y}_\varepsilon(t)\mathbf{y}_\varepsilon(t)] \\ &= \mathbf{y}_\varepsilon(t)[\mathbf{g}(\mathbf{x}_\varepsilon(t), \mathbf{y}_\varepsilon(t)) - \mathbf{g}(\mathbf{0}, \mathbf{y}_\varepsilon(t)) \\ &\quad + \mathbf{g}(\mathbf{0}, \mathbf{y}_\varepsilon(t)) - \mathbf{g}(\mathbf{0}, \mathbf{0}) + \mathbf{g}(\mathbf{0}, \mathbf{0})] \\ &\leqq |\mathbf{y}_\varepsilon(t)|[\gamma\mathbf{x}_\varepsilon(t)| - \kappa|\mathbf{y}_\varepsilon(t)| + c(t)]. \end{aligned} \tag{2.6}$$

Now set

$$v(t)=|\mathbf{x}_\varepsilon(t)|, \qquad w(t)=|\mathbf{y}_\varepsilon(t)|.$$

Then

$$\dot v-\alpha v-\beta w-b\leqq 0,$$
$$w(\varepsilon\dot w-\gamma v+\kappa w-c)\leqq 0,$$

and standard o.d.e. theory [11, Chap. 1] shows that v and w are pointwise dominated by the maximal solutions v_* and w_* of the corresponding linear equations unless $w_*(t)=0$ for some $t>0$. The latter is easily prevented by amending c.

We now triangulate the linear system by introducing $r=\omega v+w$, where $\omega=\omega_\varepsilon$ is the solution branch of

$$\alpha\varepsilon\omega+\gamma=\omega(\beta\varepsilon\omega-\kappa) \tag{2.7}$$

which remains finite as $\varepsilon\to 0$; in fact $\omega_\varepsilon=-\gamma/\kappa+O(\varepsilon)$. With these substitutions: (2.5) and (2.6) become

$$\dot v\leqq\lambda v+\beta r+b, \tag{2.8}$$

$$\varepsilon\dot r\leqq-\mu r+c, \tag{2.9}$$

where $\lambda=\alpha-\beta\omega$ and $\mu=\kappa-\beta\varepsilon\omega$ are both positive for small ε. Now we can integrate (2.9) and (2.8) to give

$$r(t)\leqq e^{-\mu T/\varepsilon}r(0)+\|c\|/\mu, \qquad \tau=t/\varepsilon,$$
$$\|r\|\leqq r(0)+\|c\|/\mu, \tag{2.10}$$

$$\|v\|\leqq e^{\lambda T}(\beta\|r\|+\|b\|)\lambda. \tag{2.11}$$

Thus $\|\mathbf{x}_\varepsilon\|$ and $\|\mathbf{y}_\varepsilon\|$ are bounded uniformly in ε. An alternative direct approach, eliminating v and using Fubini's theorem, is also possible (cf. § 4.3).

2.3. Compactness of $\mathbf{y}_\varepsilon$. Since $\mathbf{y}_\varepsilon$ are uniformly bounded, it is enough to prove that

$$\int_0^{T-s}|\mathbf{y}_\varepsilon(t+s)-\mathbf{y}_\varepsilon(t)|\,dt\to 0 \quad \text{as } s\to 0 \tag{2.12}$$

uniformly in ε, in order to conclude relative compactness of $\mathbf{y}_\varepsilon$ in L_1 [7, IV.8.20].

LEMMA 2C. *Let* $\varepsilon\dot{\mathbf{q}}(t)=\mathbf{g}(\mathbf{p}(t),\mathbf{q}(t))$, $\varepsilon\dot{\mathbf{q}}_*(t)=\mathbf{g}(\mathbf{p}_*(t),\mathbf{q}_*(t))$ *and* $\mathbf{p},\mathbf{p}_*\in L_1$. *Then*

$$\kappa\|\mathbf{q}-\mathbf{q}_*\|_1\leqq\varepsilon|\mathbf{q}(0)-\mathbf{q}_*(0)|+\gamma\|\mathbf{p}-\mathbf{p}_*\|_1.$$

Proof. Let $\mathbf{a}=\mathbf{q}-\mathbf{q}_*$, $\mathbf{b}=\mathbf{p}-\mathbf{p}_*$. Then

$$\varepsilon|\mathbf{a}(t)|\frac{d}{dt}|\mathbf{a}(t)|\leqq\frac{\varepsilon}{2}\frac{d}{dt}[\mathbf{a}(t)\mathbf{a}(t)]$$
$$\leqq|\mathbf{a}(t)|[\gamma|\mathbf{b}(t)|-\kappa|\mathbf{a}(t)|]. \tag{2.13}$$

Integrating (cf. the treatment of (2.6)), we have

$$|\mathbf{a}(t)| \leqq e^{-\kappa\tau}\left[|\mathbf{a}(0)| + \gamma \int_0^t e^{\kappa s/\varepsilon}|\mathbf{b}(s)|\, ds/\varepsilon\right].$$

Now using Fubini's theorem [7, III.11.9], we get

$$\begin{aligned}\kappa\|\mathbf{a}\|_1 &\leqq |\mathbf{a}(0)|\varepsilon(1-e^{-\kappa\tau}) + \gamma\kappa \int_0^T e^{\kappa s/\varepsilon}|\mathbf{b}(s)| \int_s^T e^{-\kappa t/\varepsilon}\, dt\, ds/\varepsilon \\ &\leqq \varepsilon|\mathbf{a}(0)| + \gamma \int_0^T |\mathbf{b}(s)|\, ds. \qquad \text{Q.E.D.}\end{aligned} \tag{2.14}$$

At this point we use the control dependence of $\mathbf{g}$ (1.6). Fix s and set

$$\mathbf{q}(t) = \mathbf{y}_\varepsilon(t+s), \qquad \mathbf{q}_*(t) = \mathbf{y}_\varepsilon(t),$$
$$b(t) = |\mathbf{x}_\varepsilon(t+s) - \mathbf{x}_\varepsilon(t)|, \qquad c(t) = |\mathbf{u}(t+s) - \mathbf{u}(t)|.$$

By an obvious amendment of the lemma,

$$\begin{aligned}\kappa\|\mathbf{q}-\mathbf{q}_*\|_1 &\leqq \varepsilon|\mathbf{y}_\varepsilon(s) - \mathbf{y}_\varepsilon(0)| + \gamma\|b\|_1 + \delta\|c\|_1 \\ &\leqq s[\max|\mathbf{g}(\cdot)| + \gamma \max|\mathbf{f}(\cdot)|] + \delta\|c\|_1.\end{aligned}$$

Since the last expression is independent of ε and tends to zero as $s \to 0$, the $\mathbf{y}_\varepsilon$ do indeed form a relatively L_1 compact set.

2.4. Convergence of $\mathbf{z}_\varepsilon$ to $\mathbf{z}_0$. We shall use sequences $\varepsilon_i \downarrow 0$, denote $\mathbf{z}_\varepsilon$ for $\varepsilon = \varepsilon_i$ by $\mathbf{z}_i$, and relabel subsequences with the same notation.

THEOREM 2D. *Under the assumptions of* § 1 *with* $\mathbf{u}$, T *and* $\mathbf{x}_\varepsilon(0) = \mathbf{x}(0)$ *fixed,* $\|\mathbf{x}_\varepsilon - \mathbf{x}_0\|$ *and* $\|\mathbf{y}_\varepsilon - \mathbf{y}_0\|_1$ *converge to zero with* ε, *while* $\mathbf{y}_i \to \mathbf{y}_0$ *almost everywhere for some sequence* $\varepsilon_i \downarrow 0$.

Proof. Starting with any sequence $\varepsilon_i \downarrow 0$, we see that § 2.3 guarantees a subsequence and $\mathbf{y}_* \in L_1$ with $\|\mathbf{y}_i - \mathbf{y}_*\|_1 \to 0$. Thus $\mathbf{y}_i \to \mathbf{y}_*$ a.e. [7, III.3.6, III.6.13] for a new subsequence. It follows that

$$\mathbf{x}_i : t \to \mathbf{x}(0) + \int_0^t \mathbf{f}(\mathbf{x}_i(s), \mathbf{y}_i(s))\, ds$$

converge uniformly over $[0, T]$ to

$$\mathbf{x}_* : t \to \mathbf{x}(0) + \int_0^t f(\mathbf{x}_*(s), \mathbf{y}_*(s))\, ds, \tag{2.15}$$

since $\mathbf{y}_i$ are bounded (§ 2.2) and $\mathbf{f}$ satisfies a (Lipschitz) uniqueness condition in the first $\mathbb{R}^n$ argument. Finally by boundedness of $\mathbf{y}_i$ and $\mathbf{g}$ and the dominated convergence theorem [7, III.6.16],

$$\begin{aligned}\int_0^t \mathbf{g}(\mathbf{x}_*(s), \mathbf{y}_*(s))\, ds &= \lim \int_0^t \mathbf{g}(\mathbf{x}_i(s), \mathbf{y}_i(s))\, ds \quad \text{as } i \to \infty \\ &= \lim [\varepsilon_i \mathbf{y}_i(t)]_0^t = \mathbf{0}.\end{aligned}$$

This establishes

$$\mathbf{g}(\mathbf{x}_*(t), \mathbf{y}_*(t)) = \mathbf{0} \quad \text{a.e. } t, \tag{2.16}$$

so $\mathbf{x}_* = \mathbf{x}_0$ and $\mathbf{y}_* = \mathbf{y}_0$ a.e. Finally it is easily seen that these are the *only* possible limit points (modulo null sets for $\mathbf{y}_0$), so $\|\mathbf{x}_\varepsilon - \mathbf{x}_0\|$ and $\|\mathbf{y}_\varepsilon - \mathbf{y}_0\|_1$ tend to zero with ε. Q.E.D.

It will be observed in addition that $\|\dot{\mathbf{x}}_i - \dot{\mathbf{x}}_0\|_1 \to 0$ by pointwise a.e. convergence of $\mathbf{y}_i$ and the dominated convergence theorem, so $\mathbf{x}_i \to \mathbf{x}_0$ in AC. Further, if we assume instead that $\mathbf{x}_\varepsilon(0) \to \mathbf{x}_0(0)$ as $\varepsilon \to 0$, then the conclusions still go through, the only change being a term in $v(0)$ in (2.11).

3. Convergence of optimal cost.

Summary. In § 3.1 the convergence theory of § 2.4 is combined with the endpoint assumption (1.12) to give upper semicontinuity (usc) of infimal cost at the reduced solution. Lower semicontinuity (lsc) is more tricky, and various results, corresponding to the three "standard" optimal control existence assumptions, are given in §§ 3.2–3.4. The first requires the full optimal controls to be uniformly well-behaved in ε; this condition will be developed for quantitative error estimates later. Strong convergence is obtained for all variables. The second imposes a convexity condition in both $\mathbf{y}$ and $\mathbf{u}$, and the third requires linearity in $\mathbf{y}$. These, and combinations of them, give (§ 3.3) strong convergence of $\mathbf{x}_\varepsilon$, but at best weak convergence of $\dot{\mathbf{x}}_\varepsilon$ and $\mathbf{y}_\varepsilon$. Using the maximum principle in § 3.4, we recover strong convergence under existence and normality assumptions on optimal reduced solutions. Finally, § 3.5 contains examples illustrating the hypotheses. *The assumptions of* § 1 *remain in force throughout* § 3.

3.1. Upper semicontinuity of infimal costs.

LEMMA 3A. *The infimal costs J_ε of* (1.1)–(1.4) *and J_0 of* (1.9)–(1.11) *satisfy* $\overline{\lim}_{\varepsilon \downarrow 0} J_\varepsilon \leqq J_0$.

Proof. Let $\mathbf{u}_0$ and $\mathbf{z}_0$ be the (perhaps weak) optimal control and solution, and T_0 be the corresponding endtime, for the reduced problem. From Theorem 2D, if $\bar{\mathbf{x}}_\varepsilon$ satisfies (1.1) with $\mathbf{u} = \mathbf{u}_0$ and $\bar{\mathbf{x}}_\varepsilon(0) = \mathbf{x}_0(0)$, then there is $\varepsilon_1 > 0$ so that

$$|\bar{\mathbf{x}}_\varepsilon(T_0) - \mathbf{x}_0(T_0)| < \varphi(\varepsilon_2) \quad \text{whenever } \varepsilon < \varepsilon_2 < \varepsilon_1, \tag{3.1}$$

where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ increases and is continuous at zero. By continuous dependence of endpoints (see (1.12)) there is $\bar{T}_\varepsilon$ so that

$$\bar{\mathbf{x}}_\varepsilon(\bar{T}_\varepsilon) \in X_1, \qquad |\bar{T}_\varepsilon - T_0| < \chi(|\bar{\mathbf{x}}_\varepsilon(T_0) - \mathbf{x}_0(T)|) \tag{3.2}$$

with χ as per φ. Choose ε_1 so that $\chi \circ \varphi(\varepsilon_1) < \infty$; this ensures that full trajectories are close enough to hit X_1.

By uniform boundedness (§ 2.2) of $\mathbf{z}_\varepsilon$ on $[0, T_0 + \chi \circ \varphi(\varepsilon_1)]$, there is a (linear) function ψ so that

$$|\bar{\mathbf{x}}_\varepsilon(\bar{T}_\varepsilon) - \bar{\mathbf{x}}_\varepsilon(T_0)| < \psi(|\bar{T}_\varepsilon - T_0|). \tag{3.3}$$

Combining (3.1)–(3.3), we obtain

$$J_\varepsilon \leqq \bar{\mathbf{x}}^0_\varepsilon(\bar{T}_\varepsilon) \leqq \mathbf{x}^0_0(T_0) + \psi \circ \chi \circ \varphi(\varepsilon_2) = J_0 + \psi \circ \chi \circ \varphi(\varepsilon_2)$$

whenever $\varepsilon < \varepsilon_2$. Q.E.D.

It will be observed that this also gives existence of feasible full solutions for each $\varepsilon < \varepsilon_1$, an assumption not made explicitly.

Since the object of the theory is to discern to what extent the reduced model can approximate the full one, usc is a necessary property to ensure against undue optimism. The endpoint variation assumption, seen to be necessary in § 3.5, seems a very reasonable one.

3.2. Compact control set. In this subsection we introduce the following additional assumption.

I. *There is a sequence of controls* $\mathbf{u}_i$, *relatively compact in* L_1, *so that if* $\mathbf{z}_i$ *and* T_i *are the response and endtime for* (1.1)–(1.4) *with control* $\mathbf{u}_i$ *and* $\varepsilon = \varepsilon_i$, *then*

$$\mathbf{x}_i^0(T_i) - J_i \to 0 \quad \text{as } i \to \infty.$$

Here J_i is the infimal cost for the $\varepsilon = \varepsilon_i$ full problem, and $\mathbf{z}_i$ could be taken as (perhaps weak) optimal solutions. Assumption I is satisfied if the optimal controls belong to the required class by virtue of model "smoothness" (e.g., linear time optimal problems), or if the minimum is sought over a restricted control class to start with.

By boundedness and [7, III.3.6, III.6.12, 13 and 16] it is no different to assume compactness in the topologies of μ-uniform (almost everywhere) convergence or convergence in measure; this in turn contains the case where the controls $\mathbf{u}_i$ have uniformly bounded variation, a case to be treated in §§ 5 and 6. Berkovitz [2, Thm. 4.1] also uses compactness with respect to convergence in measure to give an a.e. convergent control subsequence and an existence result for unbounded controls.

Conventions. For the remainder of § 3, the symbols $\mathbf{z}_i$ and $\mathbf{u}_i$ will denote the solution and control for $\varepsilon = \varepsilon_i$ and similarly for $\mathbf{z}_0$ and $\mathbf{u}_0$ at $\varepsilon = 0$. Subsequences will in general be extracted without comment and the following standard device for dealing with variable endtimes will be employed.

Let $\mathbf{z}_i(T_i) \in Z_1$: by assumption (1.8), we may extend $\mathbf{z}_i$ over all $[0, T_\infty]$ by setting

$$\mathbf{u}_i(t) = \mathbf{u}_i(T_i) \quad \text{for } t \in [T_i, T_\infty].$$

We also denote

$$\mathbf{f}_i(t) = \mathbf{f}(\mathbf{z}_i(t), \mathbf{u}_i(t)) \tag{3.4}$$

and similarly for $\mathbf{g}_i$, $\mathbf{h}_i$, $\mathbf{f}_0$, $\mathbf{g}_0$ and $\mathbf{h}_0$. Finally, weakly sequentially compact will be abbreviated to wsc.

THEOREM 3B. *If* I *holds, then* $J_i \to J_0$, *the infimal reduced cost. Further,* $\mathbf{x}_i \to \mathbf{x}_0$ *in AC and* $(\mathbf{y}_i, \mathbf{u}_i) \to (\mathbf{y}_0, \mathbf{u}_0)$ *in* L_1, *where* $\mathbf{z}_0$ *is a (perhaps weak) optimal reduced solution.*

Proof. By Lemma 3A it is enough to show that $\mathbf{z}_i \to \mathbf{z}_0$, a reduced solution with cost $\geqq J_0$. By assumption I, there is $\mathbf{u}_0$ so that $\mathbf{u}_i \to \mathbf{u}_0$ in L_1 and hence a.e. (see above). By uniform boundedness and (1.1), $\mathbf{x}_i$ are equicontinuous, so let $\|\mathbf{x}_i \to \mathbf{x}_0\| \to 0$ ((2.4)) and

$$\varepsilon_i \dot{\mathbf{y}}_{i*}(t) = \mathbf{g}(\mathbf{x}_0(t), \mathbf{y}_{i*}(t), \mathbf{u}_0(t)), \qquad \mathbf{y}_{i*}(0) = \mathbf{y}_i(0).$$

From Lemma 2C,

$$\kappa\|\mathbf{y}_i - \mathbf{y}_{i*}\|_1 \leqq \gamma\|\mathbf{x}_i - \mathbf{x}_0\|_1 + \delta\|\mathbf{u}_i - \mathbf{u}_0\|_1$$

so

$$\mathbf{y}_i - \mathbf{y}_{i*} \to 0 \quad \text{in } L_1. \tag{3.5}$$

Now arguing as in § 2.4, the $\mathbf{y}_{i*}$ are L_1 compact so have an a.e. limit $\mathbf{y}_0$, say, where (cf. (2.16))

$$\mathbf{g}(\mathbf{x}_0(t), \mathbf{y}_0(t), \mathbf{u}_0(t)) = 0 \quad \text{a.e. } t.$$

Thus since $\mathbf{y}_{i*} \to \mathbf{y}_0$ by Theorem 2D, (3.5) gives $\mathbf{y}_i \to \mathbf{y}_0$ in L_1.

It remains to discuss $\mathbf{x}_0$. We have $(\mathbf{y}_i, \mathbf{u}_i) \to (\mathbf{y}_0, \mathbf{u}_0)$ a.e., so (cf. (2.15))

$$\mathbf{x}_i(t) \to \mathbf{x}_{0*}(t) = \mathbf{x}_0(0) + \int_0^t \mathbf{f}(\mathbf{x}_{0*}(s), \mathbf{y}_0(s), \mathbf{u}(s))\, ds$$

for each t, and so $\mathbf{x}_{0*} = \mathbf{x}_0$ *is* a reduced slow solution. Finally $(\mathbf{z}_i, \mathbf{u}_i) \to (\mathbf{z}_0, \mathbf{u}_0)$ a.e. gives $\mathbf{x}_i \to \mathbf{x}_0$ in AC by the dominated convergence theorem (cf. end of § 2.4). Q.E.D.

Technically, we do not assume existence of any optimal controls, though I guarantees existence of the reduced optimum J_0 if II is satisfied.

II. *U and X_i are closed.*

3.3. Weak convergence results. We now forsake assumption I and use a priori hypotheses of convexity and linearity instead. The first conditions are variants of one originally due to McShane [13] and recently explored a good deal by control theorists. Let $\mathbf{h} = (f^0, \hat{\mathbf{h}})$ and $\mathbf{x} = (x^0, \hat{\mathbf{x}})$.

III. *There exist compact $V \subset \mathbb{R}^{n+1}$, $W \subset \mathbb{R}^k$ and $\varepsilon_1 > 0$ so that full optimal trajectories $\mathbf{z}_\varepsilon$ for $\varepsilon < \varepsilon_1$ satisfy $\mathbf{z}_\varepsilon(t) \in (V, W)$ and for all $\boldsymbol{\xi} \in V$,*

$$\Gamma(W) = \{(\varphi, \boldsymbol{\chi}) : \varphi \geqq f^0(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho}), \boldsymbol{\chi} = \hat{\mathbf{h}}(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho}), \boldsymbol{\nu} \in W, \boldsymbol{\rho} \in U\} \tag{3.6}$$

is convex.

IV. *This is the same as for* III *except that $\Gamma(\{\boldsymbol{\nu}\})$ is convex for all $(\boldsymbol{\xi}, \boldsymbol{\nu}) \in (V, W)$.*

Assumption IV is a standard optimal existence condition. Existence of a compact set containing $\mathbf{z}_\varepsilon$ follows from § 2.2, so the only new assumption is that of convexity.

LEMMA 3C. *Assume* III *and* II. *Then $J_\varepsilon \to J_0$, and given any sequence $\mathbf{z}_i$ as per the conventions of* § 3.2, *there is a reduced solution $\mathbf{z}_0$ so that $\hat{\mathbf{x}}_i \to \hat{\mathbf{x}}_0$ weakly in AC (thus $\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_0\| \to 0$) and $x_0^0(t) \leqq \lim x_i^0(t)$ as $i \to \infty$ for all t.*

Proof. By Lemma 3A the assertion about J_ε follows from that about $\mathbf{z}_i$. Now § 2.2 and boundedness of U give uniform boundedness of $\mathbf{z}_i$ and hence of the d.e. right sides $\mathbf{h}_i$ (see (3.4)). Thus $\mathbf{h}_i$ are equicontinuous, so we can assume that

$\|\mathbf{x}_i - \mathbf{x}_*\| \to 0$ and $\mathbf{h}_i \to \mathbf{h}_*$ weakly in L_1. Taking limits and using boundedness of $\mathbf{y}_i$,

$$\mathbf{x}_*(t) = \lim \int_0^t \mathbf{f}_i = \int_0^t \mathbf{f}_*, \tag{3.7}$$

$$\mathbf{0} = \lim [\varepsilon_i \mathbf{y}_i(t)]_0^t = \lim \int_0^t \mathbf{g}_i = \int_0^t \mathbf{g}_*. \tag{3.8}$$

Now III gives

$$\mathbf{h}_*(t) = \mathbf{h}(\mathbf{x}_*(t), \mathbf{y}_*(t), \mathbf{u}_*(t))$$

for some (W, U)-valued function $(\mathbf{y}_*, \mathbf{u}_*)$ which can be taken integrable by Filippov's lemma [8]. Define $\hat{\mathbf{x}}_0 = \hat{\mathbf{x}}_*$ and

$$\dot{x}_0^0(t) = \min \{f^0(\mathbf{x}_*(t), \boldsymbol{\nu}, \boldsymbol{\rho}) : \boldsymbol{\nu} \in W, \boldsymbol{\rho} \in U, \hat{\mathbf{h}}(\mathbf{x}_*(t), \boldsymbol{\nu}, \boldsymbol{\rho}) = (\dot{\hat{\mathbf{x}}}_0, \mathbf{0})\}.$$

This exists by II, and it is a simple consequence of convexity in III that $\dot{\mathbf{x}}_0 \in L_1$ (cf. [17, Thm. 10.7]). Thus $\mathbf{z}_0 = (\mathbf{x}_0, \mathbf{y}_*)$ is a reduced solution, $\hat{\mathbf{x}}_i \to \hat{\mathbf{x}}_0$ weakly in AC by (3.7) and $x_0^0(t) \leqq x_*^0(t)$ is immediate. Q.E.D.

It may be noted that there is no reason for $(\mathbf{y}_*, \mathbf{u}_*)$ to coincide with any weak limits of $(\mathbf{y}_i, \boldsymbol{u}_i)$. We now turn to an alternative assumption involving linearity.

V.

$$\mathbf{h}(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho}) = H(\boldsymbol{\xi})\boldsymbol{\nu} + \bar{\mathbf{h}}(\boldsymbol{\xi}, \boldsymbol{\rho}),$$

where $H = (F, G)$ is an $(n+1+k) \times k$ matrix-valued function.

THEOREM 3C. *Assume* V. *Then $J_\varepsilon \to J_0$ and a given sequence $\mathbf{z}_i$ converges to a (perhaps weak) reduced solution $\mathbf{z}_0$ in the following sense: $\hat{\mathbf{x}}_i \to \hat{\mathbf{x}}_0$ weakly in AC, $\mathbf{y}_i \to \mathbf{y}_0$ weakly in L_1 and $x_0^0(t) \leqq \lim x_i(t)$.*

Proof. The proof of the Lemma can be repeated, replacing $\mathbf{h}$ by $\bar{\mathbf{h}}$, down to (3.7). Since $\mathbf{y}_i$ are also wsc in L_1 (because uniformly bounded), let $\mathbf{y}_i \to \mathbf{y}_*$ weakly in L_1. Denote $F \circ \mathbf{x}_i = F_i$, etc. Then (3.7) becomes

$$\begin{aligned} \mathbf{x}_*(t) &= \lim \int_0^t [F_* \mathbf{y}_i + (F_i - F_*)\mathbf{y}_i + \bar{\mathbf{f}}_i] \\ &= \int_0^t [F_* \mathbf{y}_* + \bar{\mathbf{f}}_*] \end{aligned} \tag{3.9}$$

since $F_* \in L_\infty \cong L_1^*$ [7, IV.8.5] and $\mathbf{y}_i$ are bounded. Likewise,

$$\mathbf{0} = \int_0^t [G_* \mathbf{y}_* + \bar{\mathbf{g}}_*].$$

Differentiating and using Mazur's theorem [7; V.3.13], we obtain

$$\begin{aligned} \dot{\mathbf{x}}_*(s) &= \bar{\mathbf{f}}_*(s) - F_*(s) G_*(s)^{-1} \bar{\mathbf{g}}_*(s) \\ &\in \operatorname{co} \{\bar{\mathbf{f}}(\mathbf{x}_*(s), \boldsymbol{\rho}) - F_*(s) G_*(s)^{-1} \bar{\mathbf{g}}(\mathbf{x}_*(s), \boldsymbol{\rho}) : \boldsymbol{\rho} \in U\} \equiv \Delta, \end{aligned} \tag{3.10}$$

where co denotes convex hull.

The rest of the proof is as per Lemma 3C, weak convergence of $\hat{\mathbf{x}}_i$ to $\hat{\mathbf{x}}_0 = \hat{\mathbf{x}}$ in AC coming from (3.9) and setting, instead,

$$\dot{x}_0^0(t) = \min \{f^0(\mathbf{x}_*(t), \mathbf{y}_*(t), \boldsymbol{\rho}) : \boldsymbol{\rho} \in U, \hat{\mathbf{h}}(\mathbf{x}_*(t), \mathbf{y}_*(t), \boldsymbol{\rho}) = (\dot{\hat{\mathbf{x}}}_0, \mathbf{0})\}$$

to give a reduced solution $\mathbf{z}_0 = (\mathbf{x}_0, \mathbf{y}_*)$. Q.E.D.

Combining this with standard optimal existence theory (cf. [6], [14]), we get the following corollaries.

COROLLARY 3C1. *Assuming* II, IV *and* V, *we can take* $\mathbf{z}_0$ *optimal if* $\mathbf{z}_i$, *are.*

COROLLARY 3C2. *Assume* II *and that*

$$\mathbf{h}(\xi, \nu, \rho) = H_1\xi + H_2\nu + \tilde{\mathbf{h}}(\rho).$$

Then $\mathbf{z}_0$ *achieves optimum cost* J_0 *if* $\mathbf{z}_i$ *achieve optima* J_i.

In Corollary 3C2, $\mathbf{z}_0$ may still not be a (strong) reduced solution. Finally we may combine the above ideas with an L_1-compactness condition (cf. assumption I) as follows.

COROLLARY 3C3. *Assume* II, IV, *that* $\mathbf{z}_i$ *are optimal and that* $\mathbf{y}_i$ *are relatively* L_1*-compact. Then* $\mathbf{z}_0$ *can be taken optimal with* $\mathbf{y}_i \to \mathbf{y}_0$ *in* L_1 *as well.*

This follows because we can take $\mathbf{y}_i \to \mathbf{y}_*$ in L_1 hence a.e. Thus continuity of $\mathbf{h}$ gives

$$\mathbf{h}_*(t) \in \operatorname{co}\{\mathbf{h}(\mathbf{x}_*(t), \mathbf{y}_*(t), \rho) : \rho \in U\}$$

as the analogue of (3.9), and the argument then follows that of the Theorem.

3.4. Strong convergence results. From now on we assume V; the object is to give additional conditions ensuring convergence as in Theorem 3B.

VI. *H and* $\bar{\mathbf{h}}$ *have continuous derivatives with respect to the slow variables.*

These will be denoted by H_x and $\bar{\mathbf{h}}_x$, with $H_x \circ \mathbf{x}_i$ contracted to H_{xi}, etc. By carefully analyzing the reachable set, VI can be avoided, but the treatment given (via the maximum principle which is a statement about the reachable set) seems complicated enough anyway.

We start with a revised version of § 2, appropriate for the Euler multiplier system to be employed later. $H_i = (F_i, G_i)$ retains its previous meaning: in particular all the eigenvalues of G_i have real parts $\leqq -\kappa$ by (1.7).

LEMMA 3D. *Hypotheses*:

$$\dot{\mathbf{p}}_i = \mathbf{p}_i(A_i\mathbf{y}_i + B_i) + \mathbf{q}_i(C_i\mathbf{y}_i + D_i), \tag{3.11}$$

$$\varepsilon_i\dot{\mathbf{q}}_i = -\mathbf{p}_iF_i - \mathbf{q}_iG_i, \tag{3.12}$$

where capital symbols denote uniformly bounded linear operator-valued functions of t. $(A_i, C_i, F_i, G_i) \to (A_0, C_0, F_0, G_0)$ *pointwise while* $(B_i, D_i, \mathbf{y}_i) \to (B_0, D_0, \mathbf{y}_0)$ *weakly in* L_1. *Terminal conditions* $\mathbf{p}_i(T_i) \to \boldsymbol{\pi}_0$, $\mathbf{q}_i(T_i) = \mathbf{0}$, $T_i \to T_0$ *are given.*

Conclusions: $\mathbf{r}_i = (\mathbf{p}_i, \mathbf{q}_i) \to \mathbf{r}_0 = (\mathbf{p}_0, \mathbf{q}_0)$ *as per Theorem* 2D; *i.e.*, $\mathbf{p}_i \to \mathbf{p}_0$ *in AC*, $\mathbf{q}_i \to \mathbf{q}_0$ *in* L_1 *while* $\mathbf{r}_0$ *satisfies* (3.11), (3.12) *and the boundary conditions obtained by formally setting* $\varepsilon = 0$, $i = \infty$.

Proof. Examination of § 2 reveals no essential changes until § 2.4, except to extend solutions over $[0, T_\infty]$ (cf. § 3.2 conventions) and to observe that $\mathbf{r}_i(T_0) \to (\boldsymbol{\pi}_0, \mathbf{0})$ by boundedness of derivatives.

From 2C, $\mathbf{q}_i$ are L_1-compact and $\mathbf{p}_i$ are equicontinuous, so take

$$\|\mathbf{p}_i - \mathbf{p}_0\| \to 0 \quad \text{and} \quad \mathbf{q}_i \to \mathbf{q}_0 \quad \text{a.e.}$$

Integrating and taking limits of (3.11), we have

$$\begin{aligned}\mathbf{p}_0(t) &= \lim \int_0^t \{\mathbf{p}_0 A_0 \mathbf{y}_i + (\mathbf{p}_i A_i - \mathbf{p}_0 A_0)\mathbf{y}_i + \mathbf{p}_0 B_i + (\mathbf{p}_i - \mathbf{p}_0)B_i \\ &\quad + \mathbf{q}_0 C_0 \mathbf{y}_i + (\mathbf{q}_i C_i - \mathbf{q}_0 C_0)\mathbf{y}_i + \mathbf{q}_0 D_i + (\mathbf{q}_i - \mathbf{q}_0)D_i\} \\ &= \int_0^t \{\mathbf{p}_0(A_0 \mathbf{y}_0 + B_0) + \mathbf{q}_0(C_0 \mathbf{y}_0 + D_0)\}\end{aligned}$$

using uniform boundedness of $\mathbf{y}_i$ and $\mathbf{p}_0 A_0$ (etc.) $\in L_\infty$ (cf. (3.8)). The derivation of

$$\mathbf{0} = -\mathbf{p}_0 F_0 - \mathbf{q}_0 G_0$$

follows that of (2.16). Q.E.D.

We can now set up the principal results.

VII. *$N : \xi \to \{\pi \in \mathbb{R}^{n+1} : \pi$ is normal to X_1 at $\xi\}$ is upper semicontinuous.*

Thus if $\xi_i \in X_1$ with corresponding normals π_i and $(\xi_i, \pi_i) \to (\xi, \pi)$, then π is normal to X_1 at ξ. This prevents X_1 from having notches, but is not a great restriction.

VIII. *With $\mathbf{z}_i$ as usual, the normals $\pi_i = N(\mathbf{x}_i(T_i))$ are uniformly bounded in i.*

It is assumed here that $\pi_i^0 = 1$, so VIII is a local controllability assumption on the *full* model; cf. (1.12) which refers to the reduced model.

DEFINITION. An *extremal* is a (perhaps weak) solution satisfying the maximum principle.

THEOREM 3D. *Assume* II, VI, VII *and* VIII *and that $\mathbf{z}_i$ are (perhaps weak) optimal solutions. Then the function $\mathbf{z}_0$ of Theorem* 3C *is a reduced extremal.*

Proof. For convenience we display

$$\begin{aligned}\dot{\mathbf{x}}_i &= F_i \mathbf{y}_i + \bar{\mathbf{f}}_i, \\ \varepsilon_i \dot{\mathbf{y}}_i &= G_i \mathbf{y}_i + \bar{\mathbf{g}}_i.\end{aligned} \tag{3.13}$$

Define

$$\begin{aligned}&M_i(t) = \mathbf{r}_i(t)\bar{\mathbf{h}}(\mathbf{x}_i(t), \boldsymbol{\rho}), \\ &\mathbf{p}_i(T_i) = \pi_i, \qquad \dot{\mathbf{p}}_i = \mathbf{p}_i(F_{xi}\mathbf{y}_i + \bar{\mathbf{f}}_{xi}) - \mathbf{q}_i(G_{xi}\mathbf{y}_i + \bar{\mathbf{g}}_{xi}), \\ &\mathbf{q}_i(T_i) = \mathbf{0}, \qquad \varepsilon_i \dot{\mathbf{q}}_i = -\mathbf{p}_i F_i - \mathbf{q}_i G_i.\end{aligned} \tag{3.14}$$

From the maximum principle,

$$\mathbf{r}_i(t)\bar{\mathbf{h}}_i(t) = \max\{M_i(t) : \boldsymbol{\rho} \in U\} \equiv m_i(t). \tag{3.15}$$

Using VIII we can assume that $\pi_i \to \pi_0$, so from VII and Lemma 3D, $\mathbf{r}_i \to \mathbf{r}_0$ a.e. with $\mathbf{p}_0(T_0) \in N(\mathbf{x}_0(T_0))$.

A rather tedious calculation shows that the reduced form of (3.14) ($\varepsilon = 0$, $i = \infty$) is the multiplier system corresponding to the reduced form of (3.13), while

convergence of $\mathbf{r}_i$ gives $M_i \to M_0$ a.e. whence $m_i \to m_0$ a.e. Now,

$$\int_0^t m_0 = \lim \int_0^t \mathbf{r}_i \bar{\mathbf{h}}_i$$

$$= \lim \int_0^t \mathbf{r}_0 \bar{\mathbf{h}}_i + \lim \int_0^t (\mathbf{r}_i - \mathbf{r}_0)\bar{\mathbf{h}}_i$$

$$= \int_0^t \mathbf{r}_0 \bar{\mathbf{h}}_*$$

using weak convergence of $\bar{\mathbf{h}}_i$ on the first integral and the dominated convergence theorem [7, III.6.16] on the second. It follows that

$$\mathbf{r}_0(t)\bar{\mathbf{h}}_*(t) = m_0(t) \quad \text{a.e. } t. \qquad \text{Q.E.D.}$$

We are now in a position to deduce strong convergence from a normality condition along optimal reduced extremals, i.e., those giving infimal reduced cost J_0.

IX. *Every optimal reduced extremal* $\mathbf{z}_*$ *has a control* $\mathbf{u}_*$ *determined uniquely by the maximum principle.*

Equivalently, the set Δ (see (3.10)) is strictly convex along the chosen direction

$$\bar{\mathbf{f}}(\mathbf{x}_*(t), \mathbf{u}_*(t)) - F_*(t)G_*(t)^{-1}\bar{\mathbf{g}}(\mathbf{x}_*(t), \mathbf{u}_*(t))$$

for almost every t. Note that $\mathbf{x}_*$ is then automatically a *strong* reduced slow solution and so gives *minimal* reduced cost.

COROLLARY 3D. *Assume* IX *and Theorem* 3D. *Then* $\mathbf{z}_0$ *is an optimal reduced solution,* $\mathbf{x}_i \to \mathbf{x}_0$ *in AC and* $(\mathbf{y}_i, \mathbf{u}_i) \to (\mathbf{y}_0, \mathbf{u}_0)$ *in* L_1.

Proof. $\mathbf{z}_i$ are full extremals by the maximum principle, and $\mathbf{z}_0$ is a reduced extremal by Theorem 3D. Assumption IX now gives $\mathbf{u}_i \to \mathbf{u}_0$ a.e. hence in L_1 [7; III.6.16], so Theorem 3B completes the proof. Q.E.D.

3.5. Examples. The first shows the necessity of a continuous dependence condition for Lemma 3A to be valid. No control is needed.

Example 3E1.

$$\dot{x}^0 = x^1, \quad \dot{x}^1 = y, \quad x^1(0) = x^1(1) = 0,$$

$$\varepsilon\dot{y} = -y, \quad y(0) = 1, \quad T = 1.$$

Obviously $y_\varepsilon(t) > 0$ for $t > 0$, so $x_\varepsilon^1(1) > 0$, and no full solutions exist. On the other hand, the reduced system is

$$y_0 = 0, \quad \dot{x}_0^0 = x_0^1, \quad \dot{x}_0^1 = 0, \quad x_0^1(0) = x_0^1(1) = 0,$$

so $x_0^0(1) = 0$ is the optimal (and only possible) cost.

With minor variations, one could obtain a finite difference between reduced and all full optimal costs.

Example 3E2.

$$\dot{x}^0 = (y)^2, \quad U = \{-1, 1\}, \quad T = 1,$$

$$\varepsilon\dot{y} = u - y, \qquad y(0) = 0.$$

Clearly, $u \in L_1$ gives $y_\varepsilon \neq 0$ so $x^0(1) > 0$; thus the infimal full cost of zero is attained by the weak solution

$$x^0_\varepsilon = y_\varepsilon = u_\varepsilon = 0. \tag{3.16}$$

The reduced model is

$$y_0 = u, \quad \dot{x}^0_0 = (u)^2 = 1,$$

so $x^0_0(1) = 1$ is the optimal reduced cost, and we do not have lsc at $\varepsilon = 0$. Theorem 3B and Corollary 3C3 fail because L_1-approximations to the weak control and fast solution (3.16) are not strongly convergent. Lemma 3C fails because U is not convex and Theorem 3C fails because of the nonlinear $(y)^2$ term.

Example 3E3.

$$\dot{x}^0 = uy^1, \quad U = [-1, 1], \quad T = 1,$$

$$\varepsilon \dot{\mathbf{y}} = \begin{bmatrix} 0 & 1 \\ -2 & -2 \end{bmatrix} \mathbf{y}, \qquad \mathbf{y}(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Then $y^1_\varepsilon(t) = e^{-\tau} \sin \tau$ ($\tau = t/\varepsilon$), and the optimal control satisfies

$$u_\varepsilon(t) = -\operatorname{sgn}[y^1_\varepsilon(t)].$$

It follows that u_ε are not L_1 compact, so Theorem 3B does not apply, but y_ε *are* L_1-compact, and Corollary 3C3 gives a positive result. In fact, the reduced model is

$$\mathbf{y}_0 = \mathbf{0}, \qquad \dot{x}^0_0 = 0$$

so all reduced variables vanish. It is easily checked that $\mathbf{y}_\varepsilon \to \mathbf{0}$ pointwise on $]0, 1]$ and that $x^0_\varepsilon(1) = \|y^1\|_1 \to 0$ so, in fact, $\mathbf{z}_0$ and $\mathbf{z}_* = (x_*, \mathbf{y}_*)$ coincide here (as one would predict from linearity in u).

4. Relaxing the assumptions.

Summary. Comments will be made regarding only § 1 and are designed to enable interested readers to carry out the indicated extensions themselves. In § 4.1, we discuss more general "nonlinear Lipschitz" conditions. Section 4.2 considers the implication of time dependence to the difference between (1.3) and (1.3*). Time-dependent Lipschitz conditions form the topic of § 4.3. The questions of relaxing the stability side of (1.7) to "nonsingularity" and of terminal conditions on $\mathbf{y}$, i.e., bounded Y_1, are connected and considered in § 4.4. Finally § 4.5 contains comments on unbounded U.

4.1. Lipschitz conditions. We assume (1.7) for the present. The other Lipschitz conditions have four functions: to provide solution boundedness, solution uniqueness, norm estimates in § 2.2 and the solution difference estimate of Lemma 2C. It may be pointed out that the analysis is virtually unchanged if the condition on $\mathbf{f}$ in the fast variable is taken monotonic, i.e.,

$$[\boldsymbol{\xi} - \boldsymbol{\xi}_*][\mathbf{f}(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho}) - \mathbf{f}(\boldsymbol{\xi}_*, \boldsymbol{\nu}, \boldsymbol{\rho})] \leqq \alpha |\boldsymbol{\xi} - \boldsymbol{\xi}_*|^2.$$

Concerning boundedness, there are nonlinear estimates like

$$|\mathbf{f}(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho})| \leqq a(|\boldsymbol{\xi}| + |\boldsymbol{\nu}|), \tag{4.1}$$

with $\int_0^\theta 1/a$ divergent for $\theta > 0$, from the theory of o.d.e. Similarly for uniqueness, and conditions involving Kamke- or Lyapunov-like functions are possible (cf. [11, Chap. 2]). Boundedness and uniqueness for the reduced model must also be ensured, of course (these will not automatically follow from those for the full model, at least in t-dependent cases or nonlinear versions of (1.7)). There is a singular perturbation theory for nonunique reduced solutions (cf. [15, p. 86]), and one could derive an analogue here, but it presumably would have only theoretical value.

The differential inequality theory of § 2.2 remains unchanged with continuous nonlinear "moduli of continuity" for $\mathbf{h}$ provided certain monotonicity conditions are satisfied. For example,

$$\begin{aligned} &|\mathbf{f}(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho}) - \mathbf{f}(\boldsymbol{\xi}_*, \boldsymbol{\nu}_*, \boldsymbol{\rho})| \leqq d(|\boldsymbol{\xi} - \boldsymbol{\xi}_*|, |\boldsymbol{\nu} - \boldsymbol{\nu}_*|), \\ &|\mathbf{g}(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho}) - \mathbf{g}(\boldsymbol{\xi}_*, \boldsymbol{\nu}, \boldsymbol{\rho})| \leqq e(|\boldsymbol{\xi} - \boldsymbol{\xi}_*|) \end{aligned} \tag{4.2}$$

with both sides increasing in $|\boldsymbol{\xi} - \boldsymbol{\xi}_*|$ will suffice for § 2.2. Indeed one reaches

$$\begin{aligned} \dot{v}(t) &\leqq d(v(t), w(t)) + b(t), \\ \varepsilon \dot{w}(t) &\leqq e \circ v(t) - \kappa w(t) + c \end{aligned}$$

with v and w dominated by solutions of the corresponding equations [11, Chap. 1]. The triangulation process can still be carried out with

$$r = w + \kappa^{-1} e \circ v + a,$$

where $a \to 0$ as $\varepsilon \to 0$, to give

$$\begin{aligned} \dot{v}(t) &\leqq d_1(v(t), r(t)) + b(t), \\ \varepsilon \dot{r}(t) &\leqq -\kappa r(t) + \kappa a(t) + c(t). \end{aligned}$$

As $\varepsilon \to 0$, d_1 tends to a limit d_*, say, and it thus suffices if

$$\dot{v}(t) = d_*(v(t), r(t)) + b(t)$$

is L_1-stable for perturbations r. Again appropriate conditions can be given from o.d.e. theory [11, Chap. 3].

The other point where the Lipschitz bounds are used is in Lemma 2C; the applications are in §§ 2.3 and 3.2. With e as in (4.2), (2.13) becomes

$$\dot{\varepsilon} a \leqq w - \kappa a, \quad w(t) = e(|\mathbf{b}(t)|), \quad a(t) = |\mathbf{a}(t)|.$$

Now $\|w\|_1 \to 0$ as $\mathbf{p} \to \mathbf{p}_*$ in L_1 by continuity of e and the dominated convergence theorem, so Lemma 2C gives directly

$$\kappa \|w\|_1 \leqq \varepsilon a(0) + \|w\|_1.$$

It is easily seen that if δ (see (1.6)) is constant, then this suffices for the two applications. Nonlinear bounds $\delta(|\rho - \rho_*|)$ can be treated as follows. Assuming that δ is a modulus of continuity, we can take

$$\delta(0) = 0, \qquad \delta(\varphi + \psi) \leqq \delta(\varphi) + \delta(\psi).$$

The standard argument (cf. [7, IV.8.20]) now gives

$$\begin{aligned}\int_0^{T-s} &\delta(|u(t+s)-u(t)|)\,dt \\ &\leqq \int_0^{T-s} [\delta(|u(t+s)-v(t+s)|)+\delta(|v(t+s)-v(t)|)+\delta(|v(t)-u(t)|)]\,dt \\ &< 3\eta,\end{aligned}$$

where v is a simple function with $\|v-u\|_1 < \eta$ and s is small enough for the middle term to be less than η.

4.2. Reduced model dependence on *t*. The standard o.d.e. formulations seem to be couched in terms of (1.1), (1.3*) and to obtain fast solutions of the form $\mathbf{y}_\varepsilon(t) = \mathbf{y}_0(t) + \mathbf{y}_r(\tau)$, where $\tau = t/\varepsilon$. These are generally achieved by keeping (1.7) linear and independent of t (in the dissipative case considered so far, this means that the Jacobian $\partial\mathbf{g}/\partial\mathbf{y}(\cdot)$ has eigenvalues with real part $\leqq -\kappa$). We show below what can happen when (1.7) does depend on t.

There are two possible versions of (1.3), (1.3*) in the t-dependent case, viz.,

$$\mathbf{y}'(\tau) = \mathbf{g}(\mathbf{x}(\tau), \mathbf{y}(\tau), \mathbf{u}(\tau), \tau), \qquad \mathbf{y}_\varepsilon(t) = \mathbf{y}(\tau), \tag{4.3}$$

and

$$\varepsilon\dot{\mathbf{y}}_\varepsilon(t) = \mathbf{g}(\mathbf{x}(t), \mathbf{y}_\varepsilon(t), \mathbf{u}(t), t). \tag{4.4}$$

In (4.3), the parasitic dynamics are assumed to be externally forced on the fast time scale, while in (4.4) external effects act on the slow scale. Note that (4.3) means

$$\varepsilon\dot{\mathbf{y}}(t) = \mathbf{g}(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t), \tau),$$

so the two corresponding reduced d.e. are

$$\mathbf{0} = \mathbf{g}(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t), \infty)$$

and

$$\mathbf{0} = \mathbf{g}(\mathbf{x}(t), \mathbf{y}(t), \mathbf{u}(t), t).$$

Mathematically, the difficulties in extending the earlier analysis to (4.3) rest principally with (2.16). It is easily seen that, if

$$\lim \mathbf{g}(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho}, t) = \mathbf{g}(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho}, \infty) \quad \text{as } t \to \infty$$

exists uniformly over $(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho})$, then

$$\begin{aligned}\mathbf{0} &= \lim \int_0^t \mathbf{g}(\mathbf{x}_i(s), \mathbf{y}_i(s), \mathbf{u}(s), \infty)\,ds \quad \text{as } i \to \infty \\ &= \lim \int_0^t g(\mathbf{x}_i(s), \mathbf{y}_i(s), \mathbf{u}(s), \infty)\,ds \\ &= \int_0^t g(\mathbf{x}_*(s), \mathbf{y}_*(s), \mathbf{u}(s), \infty)\,ds.\end{aligned}$$

Without such uniformity, however, the conclusion may fail.

On the other hand, the difficulties in dealing with (4.4) basically stem from the fact that estimates as per § 2.2 will not be available in terms of τ unless uniformity assumptions in t are made on (1.7). For example, suppose that κ there is replaced by $\kappa t^{1/2}$. Then the differential inequality estimates involve equations of the form

$$\varepsilon \dot{w}(t) = -\kappa w(t) t^{1/2} + a(t) \tag{4.5}$$

instead of just $\varepsilon \dot{w} = -\kappa w + a$. Thus where previously we could use

$$\int_0^1 e^{-\kappa t/\varepsilon} a(t)\, dt/\varepsilon = O(\|a\|_\infty),$$

we now have to consider

$$\int_0^1 \exp(-\kappa t^{3/2}\eta) a(t)\, dt/\varepsilon, \qquad \eta = 2/(3\varepsilon). \tag{4.6}$$

Set $a(t) = 1$, for illustration, and $s = \eta t^{3/2}$, so (4.6) becomes

$$\eta^{1/3} \int_0^\eta e^{-\kappa s} s^{-1/3}\, ds \geqq \eta^{1/3} \int_0^1 e^{-\kappa s} s^{-1/3}\, ds$$

for large η, i.e., small ε. Thus (4.6) is $O(\varepsilon^{-1/3})$ not $O(1)$, and the whole analysis breaks down. One can make headway by using nonlinear functions of $|\boldsymbol{\nu} - \boldsymbol{\nu}_*|$ in (1.7) to counter the behavior in t, but the analysis is much more complicated and does not seem justified here.

4.3. General t-dependence. We now examine the effect of t-dependent Lipschitz conditions, again leaving (1.7) alone. It is easily seen that, apart from the point about uniformity as $t \to \infty$ if (4.3) is chosen, the analysis of § 2 (and hence of § 3 with assumptions like III taken pointwise) goes through unchanged if $\mathbf{h}$ is measurable in t and obeys

$$|\mathbf{h}(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho}, t)| \leqq d(t) \tag{4.7}$$

with $d \in L_\infty$ but $\alpha, \cdots, \delta$ independent of t. Further, if $\alpha, \cdots, \delta$ are L_∞-functions of t, then only minor alterations are necessary, except that the triangulation of § 2.2 needs modifying in one of two ways. *Either* retain (2.7) so that (2.9) becomes

$$\varepsilon \dot{r}(t) \leqq -\mu(t) r(t) + c(t) + \varepsilon \dot{\omega}(t) v(t),$$

where $\dot{\omega} \in L_\infty$ is easily checked for small enough ε. Then (2.10) takes the form

$$\begin{aligned} r(t) &\leqq r(0) + \mu^{-1}\|c\| + \int_0^t \varepsilon \dot{\omega}(s) \int_0^s \exp\left(\int_\sigma^s \lambda\right) [\beta(\sigma) r(\sigma) + b(\sigma)]\, d\sigma\, ds \\ &\leqq O[\varepsilon(\|r\| + 1)]. \end{aligned} \tag{4.8}$$

Thus $\|r\| = O(1)$. *Or* redefine $\omega = \omega_\varepsilon$ to satisfy

$$\varepsilon[\dot{\omega}(t) + \alpha(t)\omega(t) - \beta(t)\omega(t)^2] + \gamma(t) + \kappa\omega(t) = 0,$$

with $\omega_\varepsilon(0)$ chosen so that $\omega_\varepsilon(t) \to -\gamma(t)/\kappa$ uniformly as $\varepsilon \to 0$. Then the analysis of

§ 2.2 is virtually unchanged; for example, (2.10) becomes

$$r(t) \leqq r(0) \exp\left(-\varepsilon^{-1} \int_0^t \mu\right) + \|c\|_\infty \kappa^{-1} \exp(\kappa + \|\mu\|_\infty).$$

In fact, $d \in L_1$, corresponding to Carathéodory's o.d.e. condition, is possible with changes as follows. In Lemma 2A, $\mathbf{y}_0 \in L_1$ while only a $\|\mathbf{y}_\varepsilon\|_1$-estimate is available in § 2.2:

$$\|r\|_1 \leqq Tr(0) + \|c\|_1 \|\mu\|_\infty^{-1} \exp(\kappa + \|\mu\|_\infty) + O[\varepsilon(\|r\|_1 + \|b\|_1 \|\omega\|_1)]. \tag{4.9}$$

This is proved from (2.9) with $c \in L_1$ using the technique of Lemma 2C; the final term comes from the analogue of (4.8), using a convenient upper bound for $\|\omega\|_1$ as ε varies. The estimate for $\mathbf{x}_\varepsilon$ is still uniform, since

$$\|v\| \leqq (\|\beta\|_\infty \|r\|_1 + \|b\|_1) \exp(\|\lambda\|_1).$$

Inequality (4.9) (with $\|r\|_1$ terms collected) is enough for relative L_1-compactness of $\mathbf{y}_\varepsilon$ for fixed $\mathbf{u}$.

The other uses of the boundedness of d (see (4.7)) were for convergence of $\varepsilon_i \mathbf{y}_i$ pointwise to zero, domination of $\mathbf{h}_i$ (for the convergence theorem [7, III.3.16]) and wsc of $\mathbf{y}_i$ and $\mathbf{h}_i$ in L_1. With $d \in L_1$ the second of these still holds, and wsc of $\mathbf{h}_i$ is clear from the criterion [7, IV.8.10, 11]

$$\lim \int_\Sigma |\mathbf{h}_i(s)|\, ds = 0 \quad \text{as } |\Sigma| \to 0$$

uniformly in i, $|\Sigma|$ being the Lebesgue measure of $\Sigma \subset [0, T]$.

In order to establish wsc of $\mathbf{y}_i$, or equivalently $r = r_\varepsilon$ (see (2.9)) for variable $\mathbf{u}$, define

$$e(\Sigma, \varepsilon, s) = \int_{\Sigma \cap [s, T]} e^{\kappa(s-t)/\varepsilon} dt/\varepsilon.$$

Using the technique of Lemma 2C, we have

$$\kappa \int_\Sigma r_\varepsilon \leqq |r(0)| \int_\Sigma e^{-\kappa t/\varepsilon}\, dt + \kappa \int_0^T c(s) e(\Sigma, \varepsilon, s)\, ds, \tag{4.10}$$

so it suffices if $e(\Sigma, \varepsilon, s) \to 0$ uniformly in (ε, s) as $|\Sigma| \to 0$. Choose a finite union Λ of intervals with $|\Sigma \sim \Lambda| < \varepsilon |\Sigma|$ to give

$$e(\Sigma, \varepsilon, s) \leqq \kappa^{-1} |\Lambda| + \varepsilon^{-1} |\Sigma \sim \Lambda| = O(|\Sigma|)$$

as required.

The analysis is now as before, but with $\varepsilon_i \mathbf{y}_i$ *weakly* convergent to zero, so if $\mathbf{g}_i \to \mathbf{g}_0$ weakly in L_1, then

$$\int_0^t \int_0^s \mathbf{g}_0(s)\, ds\, dt = 0$$

and $\mathbf{g}_0 = \mathbf{0}$ again (cf. (3.8)). Finally, nonlinear Lipschitz conditions (cf. § 4.1) are possible if d in (4.7) is multiplied by an appropriate function of $|\boldsymbol{\xi}| + |\boldsymbol{\nu}|$ (cf. (4.1)). We note $c(s) = |\mathbf{g}(\mathbf{0}, \mathbf{0}, \mathbf{u}(s), s)|$ in (4.10); thus r_ε and $\mathbf{y}_i$ remain wsc in L_1. The only

essential difference is that $\mathbf{h}_i$ are no longer dominated, but they are wsc (because "Lipschitz" in the fast variables $\mathbf{y}_i$), so the fundamental convergence theorem below will apply. Incidentally this result also shows why subsequences have been treated so informally.

THEOREM 4C. *Let $\mathbf{y}_i$ be wsc in L_1. Then L_1 and a.e. convergence essentially coincide; i.e., if $\mathbf{y}_i \to \mathbf{y}_0$ in one mode, then all limit functions in the other mode differ from $\mathbf{y}_0$ at most on a null set.*

Proof. Let $\mathbf{y}_i \to \mathbf{y}_0$ in L_1, so we can assume $\mathbf{y}_i \to \mathbf{y}_*$ a.e. by [7, III.3.6, III.6.13(a)] for a subsequence. Thus from [7, III.6.13(b)], $\mathbf{y}_i \to \mathbf{y}_*$ in measure, and noting that only wsc (not weak convergence) is needed in [7, IV.8.12] we obtain $\mathbf{y}_i \to \mathbf{y}_*$ in L_1.

Thus $\mathbf{y}_* = \mathbf{y}_0$ a.e. Conversely, if $\mathbf{y}_i \to \mathbf{y}_0$ a.e., then the previous sentence, with $\mathbf{y}_0$ replacing $\mathbf{y}_*$, gives $\mathbf{y}_i \to \mathbf{y}_0$ in L_1. Q.E.D.

For further comments on ensuring wsc in L_1, see § 4.5.

4.4. End conditions and stability. The dissipative action of (1.7) has been seen to ensure that, with bounded Y_0, the endpoints $\mathbf{y}_\varepsilon(T)$ are close to $\mathbf{y}_0(T)$. Since $\mathbf{y}_0(T)$ is determined by

$$\mathbf{g}(\mathbf{x}_0(T), \mathbf{y}_0(T), \mathbf{u}(T)) = \mathbf{0}$$

with $(\mathbf{x}_0(T), \mathbf{u}(T)) \in (X_1, U)$. the question arises as to whether the assumption of $Y_1 = \mathbb{R}^k$ can be relaxed without losing convergence.

We suppose that the full problem (1.1)–(1.4) has an admissible (hence an infimal) solution for each $\varepsilon > 0$. Pick $\varepsilon_i \downarrow 0$ and $\mathbf{z}_i, \mathbf{u}_i$ as corresponding (perhaps weak) optimal trajectory and control with endtime T_i. Then under the conditions of Theorem 3B, it follows that

$$\mathbf{g}(\mathbf{z}_*(t), \mathbf{u}_*(t)) = \mathbf{0} \quad \text{a.e. } t \in [0, T_*], \tag{4.11}$$

where $_*$ denotes subsequential limit. If (4.11) holds, in particular, at $t = T_*$, then we have the following necessary condition.

$$\cap(Y_1, \mathbf{g}, \mathbf{z}_*(T_*), \mathbf{u}_*(T_*)) \begin{cases} \textit{The closures of } Y_1 \textit{ and} \\ \{\boldsymbol{\nu} : \mathbf{g}(\mathbf{x}_*(T_*), \boldsymbol{\nu}, \mathbf{u}_*(T_*)) = \mathbf{0}\} \textit{ should intersect.} \end{cases}$$

The notation $\cap(\cdot)$ will be used below: the condition is clearly necessary whenever a sequence of controls $\mathbf{u}_i$ is uniformly continuous on a neighborhood of T_*. On the other hand, some weakening is possible in general, e.g., if $\mathbf{u}_*$ is discontinuous at t_*, but points $\mathbf{y}_i(t)$ are controllable to Y_i for t close to T_i. An example illustrating this appears in § 6.

Having seen what restrictions the bounding of Y_1 imposes, we turn to the relaxation it permits. Working with t reversed, (1.7) becomes an *instability* assumption under which bounded Y_0 (the reversed terminal set) gives convergence under singular perturbation. Returning to the original t-direction, it follows that bounding Y_1 instead will handle the case where (1.7) is replaced by

$$(\boldsymbol{\nu} - \boldsymbol{\nu}_*)[\mathbf{g}(\boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\rho}) - \mathbf{g}(\boldsymbol{\xi}, \boldsymbol{\nu}_*, \boldsymbol{\rho})] \geqq \kappa |\boldsymbol{\nu} - \boldsymbol{\nu}_*|^2. \tag{4.12}$$

Combining these ideas, we can obtain a generalization of a standard o.d.e. singular perturbation condition involving block diagonalization of $\partial \mathbf{g}/\partial \mathbf{y}$. We first

state the result for fixed control $\mathbf{u}$ and endpoints (i.e., the setting of § 2). Explicit t-dependence would also be possible (cf. § 4.3). Let $P : \mathbb{R}^k \to \mathbb{R}^j$ denote the projection onto the first j coordinates, P^c that onto the last $k-j$ and Q the reflection $(P, -P^c)$.

THEOREM 4D. *Suppose there is a homeomorphism* $M : \mathbb{R}^k \to \mathbb{R}^k$ *so that*

$$PM^{-1}\mathbf{g}(\boldsymbol{\xi}, M\boldsymbol{\nu}, \boldsymbol{\rho}) = \mathbf{a}(\boldsymbol{\xi}, P\boldsymbol{\nu}, \boldsymbol{\rho}), \qquad P^cM^{-1}\mathbf{g}(\boldsymbol{\xi}, M\boldsymbol{\nu}, \boldsymbol{\rho}) = \mathbf{b}(\boldsymbol{\xi}, P^c\boldsymbol{\nu}, \boldsymbol{\rho}),$$

and $\mathbf{a}$ *satisfies* (1.7), $\mathbf{b}$ *satisfies* (4.12); *i.e.,* $\boldsymbol{\nu} \to QM^{-1}\mathbf{g}(\boldsymbol{\xi}, M\boldsymbol{\nu}, \boldsymbol{\rho})$ *satisfies* (1.7). *If* $\cap(P^cY_0, P^c\mathbf{g}, \mathbf{x}_0(0), \mathbf{u}(0))$ *and* $\cap(PY_1, P\mathbf{g}, \mathbf{x}_0(T), \mathbf{u}(T))$ *both hold, then so do the conclusions of Theorem* 2D.

This follows by considering $QM\mathbf{y}$ instead of $\mathbf{y}$. If there is no boundary condition on $P^cM\mathbf{y}(0)$ [$PM\mathbf{y}(T)$], then the first [second] intersection condition is unnecessary. Returning to the control problem, we see that § 3.1 will continue to hold provided that the $\cap$-conditions are valid with $\mathbf{x}_0$ as the (perhaps weak) reduced optimal solution and with U replacing $\mathbf{u}(0)$, $\mathbf{u}(T)$. If not, then we may get strict lsc in cost at $\varepsilon = 0$ (cf. Example 3E1). Conversely the results of the rest of § 3 are valid under $\cap(P^cY_0, P^c\mathbf{g}, \mathbf{x}_*(0), U)$ and $\cap(PY_1, P\mathbf{g}, \mathbf{x}_*(T_*), U)$ where, as usual, $\mathbf{x}_i \to \mathbf{x}_*$. Quantitative illustrations of § 3.1 with boundary corrections when the $\cap$-conditions are satisfied can be found in [10] and § 6.

4.5. Unbounded controls. The work of § 3, in particular, has much in common with optimum existence theory for variational problems, and unbounded derivative sets form one of the standard topics in this theory.

For most practical purposes, the relaxations in § 4.1 and here may be viewed as gilding the lily since non-Lipschitz d.e. and unbounded controls will be necessary only in rather special models. It may thus be worth commenting on some situations involving unbounded U; aside from the purely mathematical question, possible models might (i) involve convexity, (ii) be simple enough, (iii) admit impulsive control. Case (i) refers to representation of constraints by infinite-valued convex functions, a device popularized by Rockafellar and others. It may be noted that the earlier viewpoint of defining a function via its domain is an alternative. Case (ii) refers to the situation where physical bounds are present, but an unbounded (near) optimal solution is computable cheaply enough to justify an a posteriori check on whether the bounds are met. This is generally confined to linear dynamics and quadratic or time optimal cost functions, and, as mentioned in introduction, singular perturbations of such models have already been studied. Case (iii) is relevant to singular perturbations because, under certain assumptions, the reduced and full solutions can be made to *coincide*. If the optimal reduced control and trajectory are AC, then we obtain discrepancies or "boundary layers" only at the endpoints (cf. §§ 5 and 6) and, subject to impulsive controllability, the full endpoints can be moved to the reduced ones. The problem then essentially becomes of the usual nonsingular perturbation type. In general, (iii) is not covered by the analysis here, but the necessary treatment of Stieltjes integral systems can be found in [4] or [18], and the work of §§ 5 and 6 is also similar.

Assuming the positive lower bound on $f^0(\cdot)$ (see § 1.2), we may continue to assume that $(\mathbf{z}(t), t)$ belongs to a fixed compact set (cf. [6, pp. 395–6]). Section 2 then follows as previously because $\mathbf{u} \in L_1$ is fixed. Section 3, however, uses wsc in

L_1 of the o.d.e. right sides $\mathbf{h}_i$ (cf. § 4.3), and this involves "growth" assumptions as in [3], [9] and their references. While these cover §§ 3.1 and 3.2, the analysis of § 3.3 relies on $\mathbf{h}_*$, the weak L_1-limit of $\mathbf{h}_i$, taking values a.e. t in the set Γ (3.6) evaluated at $\xi = \mathbf{x}_*(t)$. This requires a uniformity condition in the absence of compactness, and two such may be found in [3, pp. 32, 36]. Finally the argument of Theorem 3D makes use of boundedness of $\mathbf{h}_i$ at (3.15): the extra assumptions needed to cover § 3.4 seem complicated and not worth detailing.

Acknowledgment. I would like to thank K. W. Chang and D. R. Westbrook for their help, and P. Kokotovic for introducing me to the topic.

REFERENCES

[1] M. ANVARI AND R. DATKO, *The existence of optimal control for a performance index with a positive integrand*, this Journal, 4 (1966), pp. 372–381.
[2] L. BERKOVITZ, *An existence theorem for optimal control*, J. Optimization Theory Appl., 4 (1969), pp. 77–86.
[3] ———, *Existence and lower closure theorems for abstract control problems*, this Journal, 12 (1974), pp. 27–42.
[4] P. BINDING, *Bounded variation evolution equations*, J. Math. Anal. Appl., 48 (1974), pp. 70–94.
[5] G. BLISS, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, Ill., 1946.
[6] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints I*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412.
[7] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Part I, Interscience, New York, 1958.
[8] A. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.
[9] A. IOFFE, *An existence theorem for problems in the calculus of variations*, Soviet Math. Dokl., 13 (1972), pp. 919–923.
[10] P. KOKOTOVIC AND A. HADDAD, *Controllability and time-optimal control of systems with slow and fast modes*, IEEE Decision and Control Conference, Phoenix, Arizona, 1974.
[11] V. LAKSHMIKANTHAM AND S. LEELA, *Differential and Integral Inequalities*, vol. 1, Academic Press, New York, 1969.
[12] E. LEE AND L. MARKUS, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.
[13] E. MCSHANE, *Existence theorems for Bolza problems in the calculus of variations*, Duke Math. J., 7 (1940), pp. 28–61.
[14] L. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
[15] R. O'MALLEY, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.
[16] J. ORTEGA AND W. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[17] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.
[18] W. SCHMAEDEKE, *Optimal control theory for nonlinear vector differential equations containing measures*, this Journal, 3 (1965), pp. 231–280.
[19] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.
[20] K. W. CHANG, *Singularly perturbed general boundary value problems*, SIAM J. Math. Anal., 3 (1972), 520–526.
[21] L. CESARI, *Lower semicontinuity and lower closure theorems without semi-normality conditions*, Ann. Mat. Pura Appl., 98 (1974), pp. 381–397.
[22] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

A DIRECT SUFFICIENT CONDITION FOR FREE FINAL TIME OPTIMAL CONTROL PROBLEMS*

P. M. MEREAU† AND W. F. POWERS‡

Abstract. A general sufficient condition for global optimality in terms of inequalities between the functions involved in the definition of the problem is developed. This condition is an extension of the sufficient condition obtained by Leitmann and Stalford [9] and can handle problems with constraints on the control and/or the state as well as problems with free final time. Simplified forms are obtained for the particular cases of minimum time and fixed final time problems. Simple examples which illustrate the applicability of the condition to problems with several extremal solutions and/or singular subarcs are also presented.

1. Problem formulation and necessary conditions. Let the following quantities be given: (t_0, x_0), a fixed point in $R^1 \times R^n$; $[T_1, T_2]$, $T_2 \geqq T_1 \geqq t_0$, a compact interval in R^1; X and U, open sets in R^n and R^m, $m \leqq n$; $X_t \subset X$ and $U_t \subset U$, arbitrary sets defined for each t in $[t_0, T_2]$; $f(t, x, u)$ and $L(t, x, u)$, continuous functions from $[t_0, T_2] \times X \times U$ into, respectively, R^n and R^1; $g(t_f, x_f)$, $\psi(t_f, x_f)$, and $\phi(t_f, x_f)$, continuous functions from $[T_1, T_2] \times X$ into, respectively, R^1, R^p, $p \leqq n+1$ and R^q.

A control $u : [t_0, t_1] \to R^m$ is said to be admissible if it is measurable and if it satisfies

$$u(t) \in U_t \quad \text{a.e. on } [t_0, t_1]. \tag{1}$$

Given an admissible control $u(t)$, $t \in [t_0, t_1]$, a trajectory $x(t)$, $t \in [t_0, t_1]$, is said to be admissible if it is an absolutely continuous function from $[t_0, t_1]$ into X, if it satisfies

$$x(t_0) = x_0, \tag{2}$$

$$\dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. on } [t_0, t_1], \tag{3}$$

$$x(t) \in X_t \quad \text{a.e. on } [t_0, t_1] \tag{4}$$

and if there exists $t_f \in [t_0, t_1] \cap [T_1, T_2]$ such that

$$\psi(t_f, x(t_f)) = 0, \tag{5}$$

$$\phi(t_f, x(t_f)) \leqq 0 \quad (\text{i.e., } \phi_i \leqq 0,\ i = 1, \cdots, q). \tag{6}$$

When the control u and the corresponding trajectory x are admissible, we shall say that the pair (u, x) is admissible or, when the final time t_f is mentioned, that the triple (u, x, t_f) is admissible. Given a cost functional

$$J = g(t_f, x(t_f)) + \int_{t_0}^{t_f} L(t, x, u)\, dt, \tag{7}$$

the problem is to determine the optimal triple (u, x, t_f) which minimizes the cost functional over the set of admissible triples.

* Received by the editors September 17, 1974, and in revised form April 17, 1975. This research was supported by the National Science Foundation under Grant GK-30115.

† Department of Aerospace Engineering, University of Michigan, Ann Arbor, Michigan. Now at ADERSA/GERBIOS, Velizy 78104, France.

‡ Department of Aerospace Engineering, University of Michigan, Ann Arbor, Michigan 48105.

Define the following functions:

$$K(t, x, u, \lambda) = L(t, x, u) + \lambda^T(t) f(t, x, u) + \dot{\lambda}^T(t) x, \tag{8}$$

$$G(t_f, x_f, \nu, \mu, \lambda_f, C) = g(t_f, x_f) + \nu^T \psi(t_f, x_f) + \mu^T \phi(t_f, x_f) - \lambda^T(t_f) x_f + C t_f, \tag{9}$$

where $\lambda(t)$ is an absolutely continuous function from $[t_0, T_2]$ into R^n and $\dot{\lambda}(t)$ is the derivative of λ with respect to t, ν, μ and C are constant quantities with values in R^p, R^q and R^1, respectively. In order to simplify some of the comments to be made later, the necessary conditions of the maximum principle [13] for the problem considered will now be given in terms of the functions K and G.

THEOREM 1 (Necessary conditions). *If the functions $f(t, x, u)$ and $L(t, x, u)$ are continuously differentiable with respect to (t, x) on $[t_0, T_2] \times X$, and if the functions $g(t_f, x_f)$, $\psi(t_f, x_f)$ and $\phi(t_f, x_f)$ are continuously differentiable with respect to (t_f, x_f) on $[T_1, T_2] \times X$, necessary conditions for an admissible triple (u^*, x^*, t_f^*) satisfying $x^*(t) \in \text{Int. } X_t$, $t \in [t_0, t_f^*]$, and the normality condition,[1] to be optimal are that there exist an absolutely continuous function $\lambda^*(t) \in R^n$, $t \in [t_0, t_f^*]$, and constant vectors $\nu^* \in R^p$, $\mu^* \in R^q$ such that*

$$\frac{\partial K}{\partial x}(t, x^*(t), u^*(t), \lambda^*(t)) = 0 \quad \text{a.e. on } [t_0, t_f^*], \tag{10}$$

$$K(t, x^*(t), u^*(t), \lambda^*(t)) \leqq K(t, x^*(t), u, \lambda^*(t)) \quad \text{for all admissible } u \text{ a.e. on } [t_0, t_f^*], \tag{11}$$

$$\frac{\partial G}{\partial t_f}(t_f^*, x^*(t_f^*), \nu^*, \mu^*, \lambda^*(t_f^*), K^*(t_f^*)) = 0 \quad (\text{with } C \equiv K^*(t_f^*)), \tag{12}$$

$$\frac{\partial G}{\partial x_f}(t_f^*, x^*(t_f^*), \nu^*, \mu^*, \lambda^*(t_f^*), K^*(t_f^*)) = 0, \tag{13}$$

$$\mu^* \geqq 0; \qquad \mu^{*T} \phi(t_f^*, x^*(t_f^*)) = 0, \tag{14}$$

where $K^*(t_f^*)$ is the value of K evaluated at t_f^* when x, u and λ take the value of the corresponding starred quantities.

In particular, if $u^*(t) \in \text{Int. } U_t$, $t \in [t_0, t_f^*]$, and if the functions $f(t, x, u)$ and $L(t, x, u)$ are continuously differentiable with respect to u on U, (11) implies

$$\frac{\partial K}{\partial u}(t, x^*(t), u^*(t), \lambda^*(t)) = 0 \quad \text{a.e. on} [t_0, t_f^*]. \tag{15}$$

2. A general sufficient condition. We shall first state and prove a sufficient condition for the optimality of a given triple and then make remarks concerning the implementation of the condition and present an illustrative example.

THEOREM 2 (Sufficient condition). *A sufficient condition for an admissible (u^*, x^*, t_f^*) to be optimal is that there exist:*

1. *A function $\tilde{\lambda}(t) \in R^n$, absolutely continuous on $[t_0, T_2]$;*
2. *Two functions $\bar{u}(t) \in U$ and $\bar{x}(t) \in X$ defined on $[t_f^*, T_2]$ and satisfying $\bar{x}(t_f^*) = x^*(t_f^*)$ and (3) a.e. on $[t_f^*, T_2]$;*

[1] That is, the constant multiplier in front of $L(t, x, u)$ in the expression of K is nonzero and can be taken equal to 1 by proper scaling of the other multipliers $\lambda_i(t)$, $i = 1, \cdots, n$. (See [1].) Note that normality is not needed for necessity but is assumed because we are interested in sufficiency.

3. *Two constants* $\tilde{\nu} \in R^p$ *and* $\tilde{\mu} \in R^q$ *satisfying* $\tilde{\mu} \geqq 0$ *and* $\tilde{\mu}^T \phi(t_f^*, x^*(t_f^*)) = 0$ *such that the following quantities*

$$\tilde{x}(t) = \begin{cases} x^*(t) & t \in [t_0, t_f^*], \\ \bar{x}(t), & t \in [t_f^*, T_2], \end{cases}$$

$$\tilde{u}(t) = \begin{cases} u^*(t), & t \in [t_o, t_f^*], \\ \bar{u}(t), & t \in [t_f^*, T_2], \end{cases}$$

$$\tilde{K}(t) = K(t, \tilde{x}(t), \tilde{u}(t), \tilde{\lambda}(t)),$$

$$\tilde{G}(t_f, x_f) = G(t_f, x_f, \tilde{\nu}, \tilde{\mu}, \tilde{\lambda}(t_f), \tilde{K}(t_f^*)) \quad (\text{with } C = \tilde{K}(t_f^*))$$

satisfy

(16) (i) $\tilde{K}(t) \leqq K(t, x, u, \tilde{\lambda}(t))$ *for all* $x \in X_t$, $u \in U_t$, *a.e. on* $[t_0, T_2]$,

(17) (ii) $\tilde{G}(t_f^*, x^*(t_f^*)) \leqq \tilde{G}(t_f, x_f)$ *for all* $x_f \in X_{t_f}$, $t_f \in [T_1, T_2]$,

(18) (iii) $[\tilde{K}(t_f) - \tilde{K}(t_f^*)](t_f - t_f^*) \geqq 0$ *for all* $t_f \in [T_1, T_2]$.

Moreover, if at least one of the above inequalities holds strictly, when $(u, x) \not\equiv (\tilde{u}, \tilde{x})$, (u^*, x^*, t_f^*) *is a proper (i.e., unique) optimal triple.*

Proof. Let (u, x, t_f) be an admissible triple. Then (16), (8) and (3) imply

$$L(t, \tilde{x}(t), \tilde{u}(t)) - L(t, x, u) \leqq \frac{d}{dt}[\tilde{\lambda}^T(t)(x - \tilde{x}(t))] \quad \text{a.e. on } [t_0, T_2].$$

Integrating both sides between t_0 and t_f and using (7) and (2) gives

$$J^* - g(t_f^*, x^*(t_f^*)) + \int_{t_f^*}^{t_f} L(t, \tilde{x}(t), \tilde{u}(t))\, dt - J + g(t_f, x(t_f)) \leqq \tilde{\lambda}^T(t_f)[x(t_f) - \tilde{x}(t_f)],$$

where J^* and J are the values of the cost functional (7) given by (u^*, x^*, t_f^*) and (u, x, t_f), respectively. Adding to the right the nonnegative quantity

$$\tilde{\nu}^T[\psi(t_f^*, x^*(t_f^*)) - \psi(t_f, x(t_f))] + \tilde{\mu}^T[\phi(t_f^*, x^*(t_f^*)) - \phi(t_f, x(t_f))] \geqq 0$$

and using (9) yields

$$J^* - J \leqq \tilde{G}(t_f^*, x^*(t_f^*)) + \tilde{\lambda}^T(t_f^*)x^*(t_f^*) - \tilde{K}(t_f^*)t_f^* - \tilde{G}(t_f, x(t_f)) - \tilde{\lambda}^T(t_f)x(t_f)$$
$$+ \tilde{K}(t_f^*)t_f + \tilde{\lambda}^T(t_f)[x(t_f) - x^*(t_f)] + \int_{t_f}^{t_f^*} L(t, \tilde{x}(t), \tilde{u}(t))\, dt,$$

but from (8) and (3),

$$\int_{t_f}^{t_f^*} L(t, \tilde{x}(t), \tilde{u}(t))\, dt = \int_{t_f}^{t_f^*} (\tilde{K}(t) - \frac{d}{dt}[\tilde{\lambda}^T(t)\tilde{x}(t)])\, dt$$

and it follows that

$$J^* - J \leqq \tilde{G}(t_f^*, x^*(t_f^*)) - \tilde{G}(t_f, x(t_f)) + \int_{t_f}^{t_f^*} [\tilde{K}(t) - \tilde{K}(t_f^*)]\, dt.$$

Then (17) and (18) imply

$$J^* \leqq J \quad \text{for all admissible triples } (u, x, t_f).$$

If at least one of the inequalities (i), (ii), (iii) holds strictly when $(u, x) \not\equiv (\tilde{u}, \tilde{x})$, the same arguments would give $J^* < J$ for all admissible triples and (u^*, x^*, t_f^*) would be a proper optimal triple. □

Remarks. 1. If $x^* \in \text{Int. } X_t$, $t \in [t_0, t_f^*]$, and the triple (u^*, x^*, t_f^*) satisfies necessary conditions for optimality, then the quantities $\lambda^*(t)$, $t \in [t_0, t_f^*]$, ν^* *and* μ^* of Theorem 1 are known (normality is assumed) and one can choose $\tilde{\nu} = \nu^*$, $\tilde{\mu} = \mu^*$, $\tilde{\lambda}(t) = \lambda^*(t)$ when $t \in [t_0, t_f^*]$ and $\tilde{\lambda}(t) = \bar{\lambda}(t)$ when $t \in [t_f^*, T_2]$, where $\tilde{\lambda}(t)$ is any absolutely continuous function from $[t_f^*, T_2]$ into R^n satisfying $\tilde{\lambda}(t_f^*) = \lambda^*(t_f^*)$. Obvious choices for the functions $\bar{u}(t)$, $\bar{x}(t)$ and $\bar{\lambda}(t)$, $t \in [t_f^*, T_2]$ are, for example,

(a) $\bar{u}(t) = \text{const. a.e. on } [t_f^*, T_2]$,
$x(t)$, solution of (3) with $u = \bar{u}(t)$ and initial condition $\bar{x}(t_f^*) = x^*(t_f^*)$,
$\bar{\lambda}(t) - \lambda^*(t_f^*)$ a.e. on $[t_f^*, T_2]$,

(b) $\bar{u}(t)$, solution of $f(t, x^*(t_f^*), u) = 0$, $t \in [t_f^*, T_2]$, if such a solution exists.
$\bar{x}(t) = x^*(t_f^*)$, $t \in [t_f^*, T_2]$, and $\bar{\lambda}(t) = \lambda^*(t_f^*)$, $t \in [t_f^*, T_2]$.

2. Since the triple (u, x, t_f) considered in the proof of Theorem 2 is admissible, the sufficient condition is obviously valid if inequalities (16) and (17) are verified with x in the set $X_t \cap R_t$, $t \in [t_0, T_2]$, and x_f in the set $X_{t_f} \cap R_{t_f} \cap \theta_{t_f}$ (whenever this is possible), where R_t is the reachable set at time t of the system (2)–(3) when $u(\tau) \in U_\tau$, $\tau \in [t_0, t]$, and θ_{t_f} is the target set at t_f, $\theta_{t_f} = \{x_f : \psi(t_f, x_f) = 0, \phi(t_f, x_f) \leqq 0\}$. Example 1 given below illustrates this remark.

3. It should be noted that the approach of this paper and [9], [11] is somewhat related to the sufficiency approach of Krotov[2] [5]–[7]. In [5], it is shown that (x^*, u^*) is optimal if there exists a function $\varphi(t, x)$ such that

$$R[t, x^*(t), u^*(t)] \leqq R[t, x(t), u(t)] \quad \forall\, x(t) \in X_t, \quad u(t) \in U_t,$$

$$\Phi[x_f^*] \leqq \Phi[x_f] \quad \forall\, x_f \in X_{t_f},$$

where

$$R = L(t, x, u) + \varphi_x f(t, x, u) + \varphi_t,$$

$$\Phi = g(x_f) + \varphi(t_f, x_f).$$

Note that R corresponds to K with $\varphi(t, x) \equiv \lambda^T(t)x$ and Φ is closely related to G. The main differences are that this paper attacks the free final time problem (which is the main extension of the results in [9], [11]), and involves an explicit method for a restricted class of problems (whereas the Krotov method applies to a larger class of problems but with the requirement of guessing the $\varphi(t, x)$-function).

[2] The authors are indebted to Professor Jack Warga for indicating the possibility of a relation to Krotov's method.

Example 1. Consider the following problem proposed in [14]:

$$\text{minimize } J = t_f$$

$$\begin{aligned}
\text{subject to } \dot{x}_1 &= \cos x_3 \quad x_1(0) = 0 \quad x_1(t_f) = \sqrt{2} \\
\dot{x}_2 &= \sin x_3 \quad x_2(0) = 0 \quad x_2(t_f) = 1 \\
\dot{x}_3 &= u \qquad\quad x_3(0) = 0 \quad x_3(t_f) \text{ free} \\
|u| &\leqq 1, \qquad t_f \in [1, 10].
\end{aligned}$$

This is a variation of Zermelo's problem, where it is desired to move a boat from a given point to another given point (in the x_1x_2-plane) in minimum time with a bound on the rate of change of the steering angle (x_3). The upper bound on t_f is chosen arbitrarily large and the lower bound is a reasonable choice since the optimal time without the constraint on u is $\sqrt{3}(>1)$, and certainly the constraint will cause the optimal time to increase.

Consider the following admissible triple:

$$\text{when } t \in [0, \pi/4]; \quad u^*(t) = 1; \quad x_1^*(t) = \sin t; \quad x_2^*(t) = 1 - \cos t; \quad x_3^*(t) = t;$$

$$\text{when } t \in \left[\frac{\pi}{4}, t_f^*\right]; \quad t_f^* = \frac{\pi}{4} + 1; \quad u^*(t) = 0; \quad x_1^*(t) = \frac{\sqrt{2}}{2}\left(t + 1 - \frac{\pi}{4}\right);$$

$$x_2^*(t) = 1 + \frac{\sqrt{2}}{2}\left(t - 1 - \frac{\pi}{4}\right); \quad x_3^*(t) = \frac{\pi}{4},$$

together with

$$\tilde{\lambda}_1(t) = \tilde{\lambda}_2(t) = -\sqrt{2}/2, \qquad t \in [0, 10],$$

$$\tilde{\lambda}_3(t) = \begin{cases} -1 + \cos\left(\frac{\pi}{4} - t\right), & t \in \left[0, \frac{\pi}{4}\right], \\ 0, & t \in \left[\frac{\pi}{4}, 10\right], \end{cases}$$

$$\bar{u}(t) = 0, \qquad t \in [t_f^*, 10],$$

$$\bar{x}_1(t) = \frac{\sqrt{2}}{2}\left(t + 1 - \frac{\pi}{4}\right); \quad \bar{x}_2(t) = 1 + \frac{\sqrt{2}}{2}\left(t - 1 - \frac{\pi}{4}\right); \quad \bar{x}_3(t) = \frac{\pi}{4}; \quad t \in [t_f^*, 10],$$

$$\tilde{\nu}_1 = \tilde{\nu}_2 = -\sqrt{2}/2.$$

Then we have, after some manipulation,

$$K(t, x, u, \tilde{\lambda}(t)) = \begin{cases} -\cos\left(\frac{\pi}{4} - x_3\right) - \left[1 - \cos\left(\frac{\pi}{4} - t\right)\right] u + x_3 \sin\left(\frac{\pi}{4} - t\right), & t \in \left[0, \frac{\pi}{4}\right], \\ -\cos\left(\frac{\pi}{4} - x_3\right), & t \in \left[\frac{\pi}{4}, 10\right], \end{cases}$$

$$\tilde{G}(t_f, x_f) = 1 + (\sqrt{2}/2), \qquad t_f \in [1, 10],$$

$$\tilde{K}(t_f) = -1, \qquad t_f \in [1, 10].$$

Let us now check inequality (16) of Theorem 2. When $t \in [0, \pi/4]$,

$$K(t, x, u, \tilde{\lambda}(t)) - \tilde{K}(t) = \left[1 - \cos\left(\frac{\pi}{4} - t\right)\right](1 - u) + \cos\left(\frac{\pi}{4} - t\right) - \alpha(t, x_3),$$

where

$$\alpha(t, x_3) = \cos\left(\frac{\pi}{4} - x_3\right) + (t - x_3) \sin\left(\frac{\pi}{4} - t\right).$$

We have $(\partial\alpha/\partial x_3)(t, x_3) = \sin(\pi/4 - x_3) - \sin(\pi/4 - t)$; but any admissible x_3 is such that $x_3 = \int_0^t u\, dt$, and since $|u| \leqq 1$, $|x_3| \leqq t$ and it follows that $(\partial\alpha/\partial x_3)(t, x_3) \geqq 0$ for any admissible x_3. Then $\sup_{x_3 \text{ admissible}} \alpha(t, x_3) = \alpha(t, \sup x_3) = \alpha(t, t) = \cos(\pi/4 - t)$ and we have,

$$K(t, x, u, \tilde{\lambda}(t)) - \tilde{K}(t) \geqq [1 - \cos(\pi/4 - t)](1 - u) \geqq 0$$

for all admissible pairs (u, x). When $t \in [\pi/4, 10]$, $K(t, x, u, \tilde{\lambda}(t)) - \tilde{K}(t) = 1 - \cos(\pi/4 - x_3) \geqq 0$. Therefore inequality (16) is satisfied (strictly when $u \neq u^*$). Inequalities (17) and (18) are also satisfied since $\tilde{G}(t_f, x_f) - \tilde{G}(t_f^*, x^*(t_f^*)) = 0$ and $\tilde{K}(t) = \tilde{K}(t_f^*) = -1$, and it follows from Theorem 2 that the triple (u^*, x^*, t_f^*) is properly optimal. It is interesting to note that besides the starred triple, there exist two other obvious extremal triples (i.e., triples which satisfy the necessary conditions of Theorem 1), the controls of which are,

$$\bar{u}(t) = \begin{cases} +1, & t \in \left[0, \frac{\pi}{4} + 2\pi\right], \\ 0, & t \in \left[\frac{\pi}{4} + 2\pi, \bar{t}_f = \frac{\pi}{4} + 2\pi + 1\right], \end{cases}$$

$$\bar{\bar{u}}(t) = \begin{cases} +1, & t \in \left[0, \frac{\pi}{4}\right], \\ -1, & t \in \left[\frac{\pi}{4}, \frac{\pi}{4} + 2\pi\right], \\ 0, & t \in \left[\frac{\pi}{4} + 2\pi, \bar{\bar{t}}_f = \frac{\pi}{4} + 2\pi + 1\right]. \end{cases}$$

The corresponding extremal trajectories in the x_1, x_2-plane are shown in Fig. 1.

This example shows that the proposed sufficient condition can be successfully applied to problems having several extremal solutions, contrary to similar sufficient conditions presented in references [8] and [10], and to problems with free final time, contrary to the sufficient conditions presented in [3], [4], [8]–[10], [12].

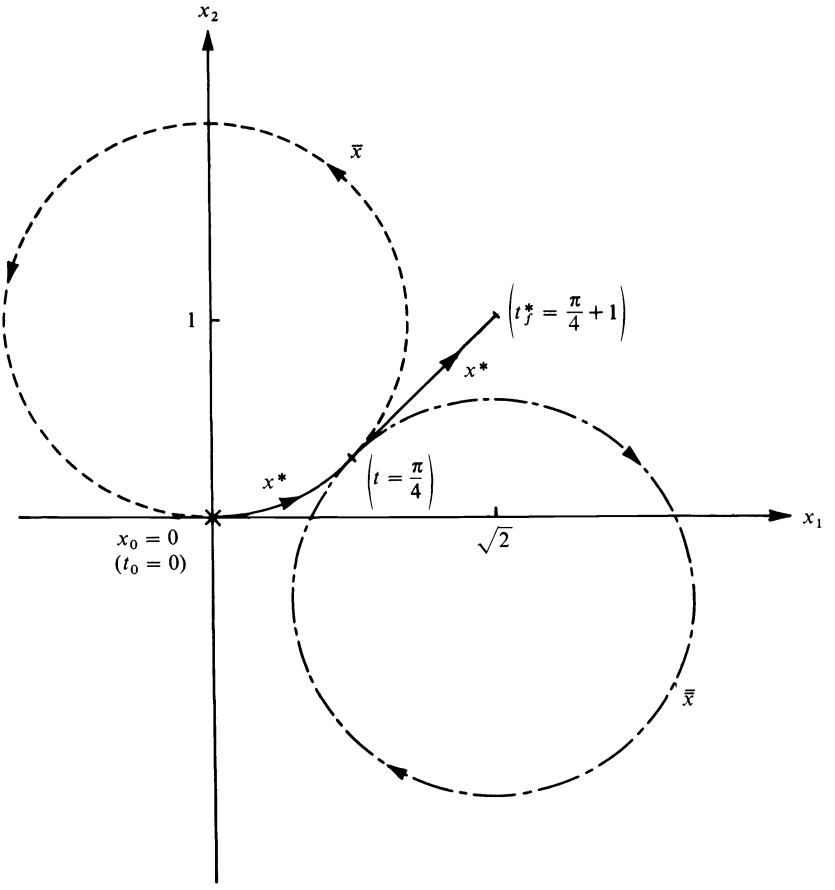


FIG. 1. *Extremal trajectories of Example* 1

3. Comments and corollaries. The sufficient condition proposed in Theorem 2 applies to a relatively general problem and thus appears somewhat complicated. When the problem considered has particular features, conditions (i), (ii) and (iii) of Theorem 2 may be simplified as shown below.

Comment 1. If inequalities (16), (17) and (18) of Theorem 2 are considered for $t \in [t_0, t_f^*]$ and $t_f \in [T_1, t_f^*]$, then the functions $\bar{u}(t)$ and $\bar{x}(t)$ are no longer necessary and Theorem 2 states that (u^*, x^*, t_f) is better (i.e., gives a smaller value to the cost functional (7)) than any admissible triple (u, x, t_f) such that $t_f \leqq t_f^*$. In particular, for a minimum time problem, $L(t, x, u) \equiv 1$, $g(t_f, x_f) = 0$ and we have the following.

COROLLARY 1 (Sufficient condition for minimum time problems). *A sufficient condition for an admissible triple* (u^*, x^*, t_f^*) *to be optimal is that there exist*:

1. *A function* $\tilde{\lambda}(t) \in R^n$, *absolutely continuous on* $[t_0, t_f^*]$;

2. *Two constants* $\tilde{\nu} \in R^p$ *and* $\tilde{\mu} \in R^q$ *satisfying* $\tilde{\mu} \geqq 0$ *and* $\tilde{\mu}^T \phi(t_f^*, x^*(t_f^*)) = 0$ *such that the following inequalities hold,*[3]

$$\text{(i)} \qquad \tilde{\lambda}^T(t)[f(t, x, u) - f(t, x^*(t), u^*(t))] + \tilde{\lambda}^T(t)[x - x^*(t)] \geqq 0$$

for all $x \in X_t$, $u \in U_t$, *a.e. on* $[t_0, t_f^*]$;

$$\text{(ii)} \qquad \tilde{\nu}^T \psi(t_f, x_f) + \tilde{\mu}^T \phi(t_f, x_f) - \tilde{\lambda}^T(t_f) x_f + \tilde{K}(t_f^*) t_f \geqq -\tilde{\lambda}^T(t_f^*) x^*(t_f^*) + \tilde{K}(t_f^*) t_f^*$$

for all $x \in X_{t_f}$, $t_f \in [t_0, t_f^*]$;

(iii) $\tilde{K}(t_f) \leqq \tilde{K}(t_f^*)$ *for all* $t_f \in [T_1, t_f^*]$.

Moreover, if at least one of the above inequalities holds strictly when $(u, x) \neq (u^*, x^*)$, (u^*, x^*, t_f^*) *is a proper optimal triple.*

Proof. Conditions (i), (ii) and (iii) above imply that inequalities (16), (17) and (18) of Theorem 2 are satisfied for $t_f \leqq t_f^*$ when $L(t, x, u) = 1$ and $g(t_f, x_f) = 0$. The case $t_f > t_f^*$ need not be considered since $J^* < J$ is automatically satisfied. □

Comment 2. When the final time t_f is prescribed, $T_1 = T_2 = T$ and condition (iii) of Theorem 2 is satisfied trivially. The sufficient condition reduces to the following.

COROLLARY 2 (Sufficient condition for fixed final time problems.) *A sufficient condition for an admissible pair* (u^*, x^*) *to be optimal is that there exist*:

1. *A function* $\tilde{\lambda}(t) \in R^n$, *absolutely continuous on* $[t_0, T]$;

2. *Two constants* $\tilde{\nu} \in R^p$ *and* $\tilde{\mu} \in R^q$ *satisfying* $\tilde{\mu} \geqq 0$ *and* $\tilde{\mu}^T \phi(x^*(T)) = 0$; *such that the quantities*

$$\tilde{K}(t) = K(t, x^*(t), u^*(t), \lambda^*(t)),$$

and

$$\tilde{G}(x_f) = \tilde{G}(t_f = T, x_f)$$

satisfy

(i) $\tilde{K}(t) \leqq K(t, x, u, \tilde{\lambda}(t))$ *for all* $x \in X_t$, $u \in U_t$, *a.e. on* $[t_0, T]$,

(ii) $\tilde{G}(x^*(T)) \leqq \tilde{G}(x_f)$ *for all* $x_f \in X_T$.

Moreover if at least one of the above inequalities hold strictly when $(u, x) \neq (u^*, x^*)$, (u^*, x^*) *is a proper optimal pair.*

Example 2. Consider the fish harvest problem proposed in [2],

$$\text{minimize } J = -\int_0^4 xu \, dt$$

$$\text{subject to } \dot{x} = x - x^2 - xu; \quad x(0) = .25; \quad x(4) \text{ free},$$

$$0 \leqq u \leqq 1; \quad 0 \leqq x \leqq 1,$$

[3] $\tilde{K}(t_f)$ is defined as in Theorem 2 and reduces to $K(t_f, x^*(t_f), u^*(t_f), \tilde{\lambda}(t_f))$ since $t_f \leqq t_f^*$.

where x, u and xu are proportional, respectively, to the fish population, the amount of effort in harvesting and the rate of fish removal.

The particular values $T=4$ and $x(0)=.25$ have been chosen in order to avoid complicated expressions, but the results hold for general values. Consider the following admissible pair:

$$\begin{aligned}
&\text{when } t\in[0,\log 3], \quad u^*(t)=0; \quad && x^*(t)=e^t/(e^t+3),\\
&\text{when } t\in[\log 3,2], \quad u^*(t)=.5; \quad && x^*(t)=.5,\\
&\text{when } t\in[2,4], \quad u^*(t)=1; \quad && x^*(t)=1/t,
\end{aligned}$$

together with

$$\tilde{\lambda}(t)=\begin{cases} -(e^{-t}/12)(e^t+3)^2, & t\in[0,\log 3],\\ -1, & t\in[\log 3,2],\\ -(t/4)(4-t), & t\in[2,4]. \end{cases}$$

Then we have after simplifications:

$$K(t,x,u,\tilde{\lambda}(t))=\begin{cases} \dfrac{e^{-t}}{12}(3+e^t)^2x^2+\dfrac{e^{-t}}{12}(3-e^t)^2xu-\frac{1}{6}(3+e^t)x, & t\in[0,\log 3]\\ x^2-x, & t\in[\log 3,2]\\ \dfrac{t}{4}(4-t)x^2-\frac{1}{4}(2-t)^2xu-\frac{1}{4}[t(2-t)+4]x, & t\in[2,4], \end{cases}$$

$\tilde{G}(x_f)=0$.

Consider inequality (i) of Corollary 2: when $t\in[0,\log 3]$, $\tilde{K}(t)=-e^t/12$ and after some manipulation,

$$K(t,x,u,\tilde{\lambda}(t))-\tilde{K}(t)=\frac{e^t}{12}[1-e^{-t}(3+e^t)x]^2+\frac{e^{-t}}{12}(3-e^t)^2xu\geqq 0$$

for all t, $u\in[0,1]$, $x\in[0,1]$.

When $t\in[\log 3,2]$, $\tilde{K}(t)=-\frac{1}{4}$ and

$$K(t,x,u,\tilde{\lambda}(t))-\tilde{K}(t)=x^2-x+\tfrac{1}{4}=(x-\tfrac{1}{2})^2\geqq 0.$$

When $t\in[2,4]$, $\tilde{K}(t)=-(4-t)/(4t)$ and after some manipulation,

$$K(t,x,u,\tilde{\lambda}(t))-\tilde{K}(t)=\tfrac{1}{4}(2-t)^2(1-u)x+\frac{4-t}{4t}(tx-1)^2\geqq 0$$

for all t, $u\in[0,1]$ and $x\in[0,1]$.

Note that inequality (ii) is satisfied trivially.

Thus inequality (i) of Corollary 2 holds strictly for $u\not\equiv u^*$ and it follows that (u^*,x^*) is a proper optimal pair. This agrees with the result of [2] where sufficiency was proved with the aid of a field-type theorem requiring the guessing of a function with special properties as well as lengthy calculations. Note that the optimal trajectory x^* has two nonsingular arcs (when $t\in[0,\log 3]$ and $t\in[2,4]$) separated by a singular arc.

4. Application of convexity and generalized convexity conditions. It can be shown [11] that inequalities (i) and (ii) in Theorem 2 and Corollaries 1 and 2 can be insured by convexity and generalized convexity conditions on the functions $K(t, x, u, \tilde{\lambda}(t))$ and $\tilde{G}(t_f, x_f)$. Such requirements lead to other sufficient conditions, less general than the conditions presented here, but which may be easier to verify (as shown in [11] with examples).

5. Conclusion. An inequality-type sufficient condition has been developed for a relatively general class of optimal control problems. The condition is applicable to problems involving free final time and/or multiple extremals, which is an improvement of the conditions reported by Leitmann and Stalford in [9].

REFERENCES

[1] G. A. BLISS, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1946.
[2] E. M. CLIFF AND T. L. VINCENT, *An optimal policy for a fish harvest*, J. Optimization Theory Appl., 12 (1973), pp. 485–496.
[3] G. M. EWING, *Sufficient conditions for global extrema in the calculus of variations*, J. Astronaut. Sci., 12 (1965), pp. 102–105.
[4] J. E. FUNK AND E. G. GILBERT, *Some sufficient conditions for optimality in control problems with state space constraints*, this Journal, 8 (1970), pp. 498–504.
[5] V. F. KROTOV, *Methods for solving variational problems on the basis of the sufficient conditions for an absolute minimum. I*, Automat. Remote Control, 23 (1962), no. 12, pp. 1473–1484.
[6]———, *Methods for solving variational problems. II. Sliding regimes*, Ibid., 24 (1963), no. 5, pp. 539–553.
[7] V. F. KROTOV, V. Z. BUKREEV AND V. I. GURMAN, *New Variational Methods in Flight Dynamics*, NASA Tech. Transl. TTF-657, 1971.
[8] E. B. LEE, *A sufficient condition in the theory of optimal control*, this Journal, 1 (1963), pp. 241–245.
[9] G. LEITMANN AND H. STALFORD, *A sufficiency theorem for optimal control*, J. Optimization Theory Appl., 8 (1971), pp. 169–174.
[10] O. L. MANGASARIAN, *Sufficient conditions for the optimal control of nonlinear systems*, this Journal, 4 (1966), pp. 139–152.
[11] P. MEREAU, *Global optimality conditions and the Darboux point*, Ph.D. thesis, University of Michigan, Ann Arbor, 1974.
[12] P. MEREAU AND J. G. PAQUET, *Sufficient conditions for optimal controls—Use of generalized convexity*, Internat. J. Control, 19 (1974), pp. 615–624.
[13] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
[14] W. F. POWERS AND J. P. MCDANELL, *Switching conditions and a synthesis technique for the singular Saturn guidance problem*, AIAA J. Spacecraft and Rockets, 8 (1971), pp. 1027–1032. (Example appears in AIAA Paper no. 70-965.)

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

THE REPRESENTATION OF MARTINGALES OF JUMP PROCESSES*

M. H. A. DAVIS†

Abstract. In this paper it is shown that all local martingales of the σ-fields generated by a jump process of very general type can be represented as stochastic integrals with respect to a fundamental family of martingales associated with the jump process.

1. Introduction. Recently a number of results have been obtained on filtering, detection and stochastic control problems involving discontinuous stochastic processes, using methods based on martingale theory. These developments stem from the realization [3] that such problems are mathematically analogous to the corresponding problems involving "signals in additive Gaussian white noise". The theory in the latter case has reached a certain degree of completeness, reflected in the appearance of a comprehensive account in [10], and is based on two fundamental results in the calculus of Brownian motion, namely, the Ito differential formula and the fact that all the martingales on the σ-fields generated by a Brownian motion can be represented as stochastic integrals.

Stochastic calculus having been developed to a high degree of generality in [7], an analogous theory can be developed in a situation where suitable martingale representation results are available. It is the purpose of this paper to establish these in the case where the underlying process lies in a certain class of jump processes (i.e., processes with piecewise-constant paths). Such processes arise in operations research, optical communications and in many other areas of communication and control theory.

The basic jump process considered in this paper is defined in § 2 below. It takes values in a measurable space $(X, \mathcal{S})$ and the jump times have a single accumulation point at a random termination time T_∞ (which may be identically $+\infty$). The main result is Theorem 2 in § 3.2 which states that every local martingale on the σ-fields generated by such a process has an integral representation with respect to a certain fundamental family of martingales associated with the basic process, and identifies the necessary class of integrands.

The basic process is defined in terms of a family of conditional distributions. In § 4 the possibility of an alternative specification in terms of a "local description" is investigated. This is perhaps more natural from the applications point of view.

Related papers are those by Boel, Varaiya and Wong [2, Part I], Chou and Meyer [4], Elliott [8] and Jacod [9]. Roughly speaking, this paper derives the results of [2, Part I] by the methods of [4]. The advantage of the approach in

* Received by the editors November 12, 1974, and in revised form April 10, 1975.

† Department of Computing and Control, Imperial College, London SW7 2BZ, England. This work was supported in part by the U.S. Office of Naval Research under the Joint Services Electronics Program by Contract N00014-67-A-0298-006 at Harvard University and by the U.S. Air Force Offices of Scientific Research, Air Force Systems Command, under Contract AF 44-620-69-C-0101 at Stanford University.

[4]—where the results below are proved for the special case of point processes, i.e., real-valued jump processes with all jumps having magnitude +1—is that it is virtually self-contained and no general theorems on stochastic integration or martingale decomposition are invoked. By contrast the argument in [2, Part I] is based on a theory of stochastic integration for square-integrable martingales and this entails the assumption that the jump times of the basic process are totally inaccessible stopping times, since otherwise their Lemma 3.1, which in our notation says that $\langle q(\cdot, A)\rangle_t = \tilde{p}(t, A)$, is not true. It was also assumed that $T_\infty \equiv \infty$. By generalizing the methods of [4] and introducing some sequences of stopping times to show that certain integrands are "locally L_1" we are able to show that both of these assumptions are unnecessary. Further generalizations will be found in [8] where the formula for $\langle q(\cdot, A)\rangle_t$ is given in case the basic jump times have an accessible part, and the process can continue beyond T_∞ (where, in our framework, it terminates).

The representation result is also given in a form close to ours in [9] where the proof proceeds via an exponentiation formula for positive martingales and the Radon–Nikodym theorem. This route is less direct than ours (representations are not the primary concern of [9]), and, although there is no stochastic integration, results on predictable projections, etc. from the "general theory of processes" [6] are freely used.

For applications of the results the reader is referred to [14] (filtering), [2, Part II] (filtering and detection) and [1] (stochastic control). Reference [13] deals (not from the martingale standpoint) with control of jump processes defined very much as below.

2. Definition. The basic jump process (x_t) is to take values in a measurable space $(X, \mathcal{S})$. We assume that $(X, \mathcal{S})$ is a Blackwell space [11, III D15], not a restrictive assumption since, for example, complete separable metric spaces are included. The salient feature of these spaces is that two separable sub-σ-fields of $\mathcal{S}$ are then identical if and only if they have the same atoms; this property is used in [12] to obtain the characterization of stopped σ-fields which we use in Lemma 0 below. Let z_0, z_∞ be fixed elements of X and for $i = 1, 2, \cdots$, let $(Y^i, \mathcal{Y}^i)$ denote a copy of the measurable space[1]

$$(Y, \mathcal{Y}) = ((R^+ \times X) \cup \{(\infty, z_\infty)\}, \quad \sigma\{\mathcal{B}(R^+) * \mathcal{S}, \quad \{(\infty, z_\infty)\}\})$$

Now define

$$\Omega = \prod_{i=1}^{\infty} Y^i,$$

$$\mathcal{F}^0 = \sigma\left\{\prod_{i=1}^{\infty} \mathcal{Y}^i\right\}.$$

An argument similar to that of [2, Part I, § 2] shows that $(\Omega, \mathcal{F}^0)$ is then also a Blackwell space.

[1] $\mathcal{B}(R^+)$ is the Borel σ-field of R^+; $\sigma\{\cdot\}$ denotes the σ-field generated by the sets or random variables (r.v.) in the braces.

Let $(S_i, Z_i) : \Omega \to Y^1$ be the coordinate mapping and $\omega_k : \Omega \to \Omega_k = \prod_{i=1}^{k} Y^i$ be the restriction of the identity function, i.e.,

$$\omega_k(\omega) = (S_1(\omega), Z_1(\omega) \cdots S_k(\omega), Z_k(\omega)).$$

We now set

$$T_k(\omega) = \sum_{i=1}^{k} S_i(\omega),$$

$$T_\infty(\omega) = \lim_{k\to\infty} T_k(\omega),$$

and define the path $(x_t(\omega))_{t\geqq 0}$ by

$$x_t(\omega) = \begin{cases} z_0 & \text{if } t < T_1(\omega), \\ Z_i(\omega) & \text{if } t \in [T_i(\omega), T_{i+1}(\omega)[, \\ z_\infty & \text{if } t \geqq T_\infty(\omega). \end{cases}$$

The random function (x_t) generates the increasing family of σ-fields $(\mathscr{F}_t^0)$, i.e.,

$$\mathscr{F}_t^0 = \sigma\{x_s, s \leqq t\}.$$

Notice that T_i is not a stopping time of $(\mathscr{F}_t^0)$ because we have not excluded the possibility that $Z_{i-1} = Z_i$. A probability measure P on $(\Omega, \mathscr{F}^0)$ is defined by the following family of conditional distribution functions: μ^1 is a probability measure on $(Y, \mathscr{Y})$ such that $\mu^1((\{0\}\times X)\cup(R^+\times\{z_0\})) = 0$, and for $i = 2, 3, \cdots$, $\mu^i : \Omega_{i-1}\times\mathscr{Y} \to [0, 1]$ is a function such that

(i) $\mu^i(\cdot\,; \Gamma)$ is measurable for each fixed Γ,

(ii) $\mu^i(\omega_{i-1}(\omega); \cdot\,)$ is a probability measure on $(Y, \mathscr{Y})$ for each fixed $\omega \in \Omega$,

(iii) $\mu^i(\omega_{i-1}(\omega); (\{0\}\times X)\cup(R^+\times\{Z_{i-1}(\omega)\})) = 0$ for all ω,

(iv) $\mu^i(\omega_{i-1}(\omega); \{(\infty, z_\infty)\}) = 1$ if $S_{i-1}(\omega) = \infty$.

The family (μ^i) defines a probability measure on $(\Omega, \mathscr{F}^0)$ as follows: for $\Gamma \in \mathscr{Y}$ and $\eta \in \Omega_{i-1}$,

$$P[(T_1, Z_1) \in \Gamma] = \mu^1(\Gamma),$$

$$P[(S_i, Z_i) \in \Gamma | \omega_{i-1} = \eta] = \mu^i(\eta; \Gamma).$$

The purpose of (iii) is to ensure that two "jump times" T_{i-1}, T_i do not occur at once and that the process x_t does effectively jump at its jump times; i.e.,(iii) implies

$$P[T_{i-1} = T_i] = P[Z_{i-1} = Z_i] = 0.$$

Part (iv) ensures that $P[Z_k = z_\infty | T_k = \infty] = 1$.

Now let $\mathscr{F}_t(\mathscr{F})$ be the σ-field obtained by augmenting $\mathscr{F}_t^0(\mathscr{F}^0)$ with all subsets of P-null sets of $\mathscr{F}^0$.

LEMMA 0. (a) *T_i is a stopping time of $(\mathscr{F}_t)$.*

(b) *T_∞ is a predictable stopping time of $(\mathscr{F}_t)$.*

(c) *$\mathscr{F}_\infty = \mathscr{F}$, where $\mathscr{F}_\infty = \bigvee_{t\geqq 0} \mathscr{F}_t$.*

(d) $\bigvee_n \mathscr{F}_{T_n} = \mathscr{F}_{T_\infty -} = \mathscr{F}_{T_\infty} = \mathscr{F}$.

Proof. (a) Let $p_t = \sum_{s \leqq t} I_{(x_s \neq x_{s-})}$ and $p(t, X) = \sum_i I_{(t \geqq T_i)}$. Then from (iii) above, $p_t = p(t, X)$ a.s. and $(T_i \leqq t) = (p(t, X) \geqq i)$. The result follows.

(b) $\tau_n = T_n \wedge n$ foretells T_∞ in the sense of [6, III D 36].

(c) By definition $\mathscr{F}^0 = \sigma\{T_i, Z_i, i = 1, 2, \cdots\}$ so that $\mathscr{F}_\infty \subset \mathscr{F}$. Since T_i is a stopping time of $\mathscr{F}_t$ it is $\mathscr{F}_\infty$-measurable, so to show $\mathscr{F} \subset \mathscr{F}_\infty$ it remains to show that $(Z_i \in A) \in \mathscr{F}_\infty$ for $A \in \mathscr{S}$. Now

$$(Z_i \in A) = ((Z_i \in A) \cap (T_i < \infty)) \cup ((Z_i \in A) \cap (T_i = \infty)).$$

The first set on the right is in $\mathscr{F}_\infty$ as is the set $(T_i = \infty)$. If $z_\infty \notin A$, then $P[(Z_i \in A) \cap (T_i = \infty)] = 0$ whereas if $z_\infty \in A$, $P[(T_i = \infty) \setminus (T_i = \infty) \cap (Z_i \in A)] = 0$. Hence $(Z_i \in A) \cap (T_i = \infty) \in \mathscr{F}_\infty$. This completes the proof.

(d) The first equality is given by [6, III T 35(b)]. To get the second and third we apply the characterization of stopped σ-fields given by Courrège and Priouret [5] and by Meyer [12]. However since this refers to the unaugmented σ-fields we first define

$$\Omega'' = \{\omega \in \Omega : Z_{i-1}(\omega) = Z_i(\omega) \text{ for some } i = 1, 2, \cdots\}$$

which by definition is a P-null set. Now let

$$\Omega' = \Omega \setminus \Omega'',$$

$$\mathscr{F}'_t = \{F \cap \Omega', F \in \mathscr{F}^0_t\}.$$

This is a Blackwell space since $(\Omega, \mathscr{F}^0)$ is.

T_i restricted to Ω' is a stopping time of $\mathscr{F}'_t$ since $\{\omega \in \Omega' : T_i(\omega) \leqq t\} = \{\omega \in \Omega' : p_t(\omega) \geqq i\}$, and hence $(\mathscr{F}'_t)$ is the family of σ-fields generated by (x_t) in Ω'. For U a stopping time of $(\mathscr{F}'_t)$ define the equivalence relation R_U by

$$\omega \sim \omega'(R_U) \Leftrightarrow U(\omega) = U(\omega') \quad \text{and}$$

$$x_t(\omega) = x_t(\omega') \quad \text{for } t \leqq U(\omega).$$

Then according to [12, Prop. 1], $A \in \mathscr{F}_U$ if and only if A is saturated for R_U (i.e., $\omega \in A$, $\omega' \sim \omega(R_U) \Rightarrow \omega' \in A$).

It follows that for each n, $\mathscr{F}'_{T_n} = \sigma\{T_i, Z_i, i = 1, 2, \cdots n\}$ since

$$\omega \sim \omega'(R_{T_n}) \Leftrightarrow T_i(\omega) = T_i(\omega'), \quad Z_i(\omega) = Z_i(\omega'), \qquad i = 1, 2, \cdots, n.$$

Also since $x_{T_\infty} \equiv z_\infty$,

$$\omega \sim \omega'(R_{T_\infty}) \Leftrightarrow T_i(\omega) = T_i(\omega'), \quad Z_i(\omega) = Z_i(\omega') \quad \text{for all } i,$$

and hence $\mathscr{F}'_\infty = \mathscr{F}'_{T_\infty} = \bigvee_n \mathscr{F}'_{T_n}$. Augmenting these σ-fields with the null sets of $\mathscr{F}^0$ (in particular, with Ω'') gives the result as stated.

For fixed k and $t \geqq 0$, $U = (T_{k-1} + t) \wedge T_k$ is a stopping time of $(\mathscr{F}_s)$. We shall need the following characterization of the σ-field $\mathscr{F}_U$.

LEMMA 1. $\mathscr{F}_U = \mathscr{F}_{T_{k-1}} \vee \sigma\{x_{(T_{k-1}+s) \wedge T_k}, s \in [0, t]\}$.

Proof. As in the proof of Lemma 0,

$$\mathscr{F}'_{T_{k-1}} = \sigma\{T_i, Z_i, i = 1, 2 \cdots k-1\}$$

$$= \sigma\{x_{s \wedge T_{k-1}}, s \in R^+\}.$$

Thus

$$\mathscr{F}'_{T_{k-1}} \vee \sigma\{x_{(T_{k-1}+s)\wedge T_k},\ s \in [0, t]\} = \sigma\{x_{s\wedge U},\ s \in R^+\}$$
$$= \mathscr{F}'_U.$$

Augmenting with null sets now gives the result since $\mathscr{F}_{T_{k-1}}$ already contains all P-null sets.

3. Representation results.

3.1. The single jump case. The process (x_t) has a single jump if

$$\mu^2(\eta; \{(\infty, z_\infty)\}) = 1 \quad \text{for all } \eta \in Y^1$$

so that $T_\infty \equiv \infty$ in this case. Such a process can be defined directly by taking $(\Omega, \mathscr{F}^0) = (Y, \mathscr{Y})$, $P = \mu^1$,

$$x_t(\omega) = \begin{cases} z_0 & \text{if } t < T_1, \\ Z_1(\omega) & \text{if } t \geqq T_1, \end{cases}$$

and $(\mathscr{F}_t)$ defined as before. (T_1, Z_1) is now the identity function on $(\Omega, \mathscr{F}^0)$ and it is not hard to see that $\mathscr{F}_t$ consists of $(\mathscr{B}[0, t])*\mathscr{S}$ together with the set $]t, \infty[\times X) \cup \{(\infty, z_\infty)\}$ and all μ^1-null sets of $\mathscr{F}^0$. For the remainder of this section, T_1, Z_1, μ^1 are denoted by T, Z, μ.

PROPOSITION 1. *Suppose τ is a stopping time of $(\mathscr{F}_t)$. Then there exists $t_0 \in R^+$ such that $t \wedge T = t_0 \wedge T$ a.s.*

Proof. Form $\Omega' = Y \setminus (R^+ \times \{z_0\})$ and $\mathscr{F}'_t$ as in the proof of Lemma 1. Then $]t, \infty[\times X) \cap \Omega'$ is an atom of $\mathscr{F}'_t$. Suppose τ takes on at least two values t_1 and t_2 on $(\tau \leqq T)$. Then for $t \in]t_1, t_2[$ we have

$$(\tau \leqq t) \cap (]t, \infty[\times X) \cap \Omega' \subsetneqq (]t, \infty[\times X) \cap \Omega'$$

so that $(\tau \leqq t) \notin \mathscr{F}'_t$. This shows that $(\tau \leqq T) \subset (t_0 \leqq T)$ for some $t_0 \in R^+$ and a similar argument gives the reverse inclusion. Augmenting $\mathscr{F}'_t$ the result follows.

For $A \in \mathscr{S}$ let us define

$$F_t^A = \mu(]t, \infty] \times A) \tag{1}$$

so that F^A is right-continuous for fixed A. In particular, the marginal distribution of T is given by

$$F_t = P[T > t] = F_t^X.$$

Now define

$$c = \inf\{t : F_t = 0\}.$$

Note that F_t is *decreasing* so that for integrable functions $f(T)$,

$$Ef(T) = \int f\, d\mu = -\int f\, dF.$$

Suppose $(M_t)_{t \geqq 0}$ is a local martingale of $(\mathscr{F}_t)$. The following result is proved as in [4], using Proposition 1.

PROPOSITION 2. (i) *If $c = \infty$ or, $c < \infty$ and $F_{c-} = 0$, then M_t is a martingale on* $[0, c[$.

(ii) *If* $c<\infty$ *and* $F_{c-}>0$, *then* (M_t) *is a uniformly integrable* (u.i.) *martingale.*

We now introduce the fundamental family of martingales associated with the process (x_t). For $A\in\mathcal{S}$ and $t\in R^+$ define

$$\tilde{p}(t,A)=I_{(t\geqq T)}I_{(Z\in A)},$$

$$\tilde{p}(t,A)=-\int_{]0,T\wedge t]}\frac{1}{F_{s-}}\,dF_s^A,$$

$$q(t,A)=p(t,A)-\tilde{p}(t,A),$$

PROPOSITION 3. $(q(t,A))_{t\in R^+}$ *is a martingale of* $(\mathcal{F}_t)$.

Proof. Direct computation. For $t>s$ we have

$$E[p(t,A)-p(s,A)|\mathcal{F}_s]=I_{(s<T)}\frac{F_s^A-F_t^A}{F_s},$$

$$E\left[\int_{]s\wedge T,t\wedge T]}\frac{1}{F_{u-}}\,dF_u^A|\mathcal{F}_s\right]=I_{(s<T)}\left\{\frac{F_t}{F_s}\int_{]s,t]}\frac{dF_u^A}{F_{u-}}-\frac{1}{F_s}\int_{]s,t]}\int_{]s,r]}\frac{dF_u^A}{F_{u-}}\,dF_r\right\}$$

and

$$\int_{]s,t]}\int_{]s,r]}\frac{dF_u^A}{F_{u-}}\,dF_r=\int_{]s,t]}\frac{1}{F_{u-}}\int_{[u,t]}dF_r\,dF_u^A=F_t\int_{]s,t]}\frac{dF_u^A}{F_{u-}}+F_s^A-F_t^A.$$

The martingale equality follows from these calculations.

Let $\mathcal{I}$ denote the set of measurable functions $g: Y\to R$ such that $g(\infty,z)=0$ for all $z\in X$. Since, for fixed (t,ω), the functions $p(t,A)$, $\tilde{p}(t,A)$ are countably additive in A, we can define Stieltjes integrals of the form $\int g(t,z)p(dt,dz)$, $\int g(t,z)\tilde{p}(dt,dz)$ for suitable integrands $g\in\mathcal{I}$. Explicitly,

$$\text{(2)}\qquad \begin{aligned}&\int_{R^+\times X}g(t,z)p(dt,dz)=g(T,Z),\\ &\int_{R^+\times X}g(t,z)\tilde{p}(dt,dz)=\int_{R^+\times X}I_{(s\leqq T)}g(s,z)\frac{1}{F_{s-}}\,d\mu(s,z).\end{aligned}$$

We introduce the following classes of integrands, which are similar to those defined in [2]:

$$L^1(p)=\left\{g\in\mathcal{I}: E\int_{R^+\times X}|g(t,z)|p(dt,dz)<\infty\right\},$$

$$L^1_{\text{loc}}(p)=\{g\in\mathcal{I}: gI_{(t<\sigma_k)}\in L^1(p),\ k=1,2,\cdots,\text{ for some sequence of stopping times } \sigma_k<T_\infty,\ \sigma_k\uparrow T_\infty \text{ a.s.}\}.^2$$

$L^1(p)$ and $L^1_{\text{loc}}(\tilde{p})$ are defined analogously.

PROPOSITION 4.

(i) $L^1(p)=L^1(\tilde{p})=\{g\in\mathcal{I}:\int_Y|g|\,d\mu<\infty\}$,

(ii) $L^1_{\text{loc}}(p)=L^1_{\text{loc}}(\tilde{p})=\mathcal{P}$, where $\mathcal{P}=\{g\in\mathcal{I}:\int_Y I_{(s\leqq t)}|g|d\mu<\infty \text{ for all } t<c\}$.

---

[2] Recall that in this section $T_\infty\equiv\infty$. T_∞ is written here for future reference.

Proof. (i) For $g \in \mathcal{I}$,

$$\int_{R^+ \times X} |g(t, z)| p(dt, dz) = |g(T, Z)|,$$

and hence,

$$E \int |g| \, dp = \int_Y |g| \, d\mu.$$

Now

$$\int_{R^+ \times X} |g| \, d\tilde{p} = \int_{]0,T] \times X} \frac{1}{F_{s-}} |g(s, z)| \, d\mu(s, z)$$

so that

$$\begin{aligned} E \int_{R^+ \times X} |g| \, d\tilde{p} &= - \int_{]0,\infty]} \int_{]0,t] \times X} \frac{1}{F_{s-}} |g(s, z)| \, d\mu(s, z) \, dF_t \\ &= \int_Y \left(- \int_{[s,\infty]} dF_t \right) \frac{1}{F_{s-}} |(g(s, z))| \, d\mu \\ &= \int_Y |g(s, t)| \, d\mu. \end{aligned}$$

Thus $L^1(p) = L^1(\tilde{p})$.

(ii) Suppose $g \in L^1_{\text{loc}}(p)$. Let (σ_k) be an associated sequence of stopping times, and (t_k) be numbers such that $\sigma_k \wedge T = t_k \wedge T$ (Proposition 1). Since $\sigma_k \uparrow \infty$ a.s. it is clear that $t_k \uparrow c$. Now

$$\int_Y I_{(s<\sigma_k)} |g| \, dp = g(T, Z) I_{(\sigma_k > T)}.$$

It is easy to see that $(\sigma_k > T) = [0, t_k[\times X$. Thus

$$E \int_Y I_{(s<\sigma_k)} |g| dp = \int_{]0,t_k[\times X} |g| \, d\mu.$$

Since $t_k \uparrow c$, it follows that $g \in \mathcal{P}$. Conversely, suppose $g \in \mathcal{P}$. Then $gI_{(s<\sigma_k)} \in L^1(p)$ with the following choice of stopping times (σ_k):

for $c = \infty$: $\sigma_k = k$,

for $c < \infty$, $F_{c-} > 0$: $\sigma_k = \infty$,

for $c < \infty$, $F_{c-} = 0$: take $t_k \uparrow c$ and define

$$\sigma_k = kI_{(T \leqq t_k)} + t_k I_{(T > t_k)}.$$

Evidently $\sigma_k \uparrow \infty$ a.s.; consequently $L^1_{\text{loc}}(p) = \mathcal{P}$. A calculation similar to that in the proof of (i) shows that for $\sigma_k \uparrow \infty$,

$$E \int_{]0,\infty] \times X} I_{(s<\sigma_k)} |g| \, d\tilde{p} = \int_{]0,t_k[\times X} |g| \, d\mu$$

and $\mathcal{P} = L^1_{\text{loc}}(\tilde{p})$ follows from this. This completes the proof.

PROPOSITION 5. *Suppose* (M_t) *is a uniformly integrable martingale of* $(\mathcal{F}_t)$ *such that* $M_0 = 0$ *a.s. Then there exists* $h \in \mathcal{I}$ *such that*

$$\int_Y |h|\, d\mu < \infty, \tag{3}$$

$$M_t = I_{(t \geqq T)} h(T, Z) - I_{(t<T)} \frac{1}{F_t} \int_{]0,t] \times X} h(s, z)\, d\mu(s, z). \tag{4}$$

Proof. Each u.i. martingale is of the form $M_t = E(\xi | \mathcal{F}_t)$ for some $\mathcal{F}$-measurable r.v. ξ, and from the definition of $\mathcal{F}$ each such r.v. is a.s. equal to $h(T, Z)$ for some measurable $h : Y \to R$. Expression (3) is satisfied since $E|M_t| < \infty$ and if $M_0 = 0$, then

$$\int_Y h\, d\mu = 0. \tag{5}$$

Now

$$E[h(T, Z) | \mathcal{F}_t] = I_{(t \geqq T)} h(T, Z) + I_{(t<T)} \frac{1}{F_t} \int_{]t,\infty] \times X} h(s, z)\, d\mu(s, z) \tag{6}$$

and (4) follows from (5) and (6).

For $g \in \mathcal{P}$, the stochastic integral

$$M_t^g = \int_{]0,t] \times X} g(s, z) q(ds, dz)$$

is defined as the obvious difference of Stieltjes integrals, i.e.,

$$M_t^g = \int_{R^+ \times X} I_{(s \leqq t)} g(s, z) p(ds, dz) - \int_{R^+ \times X} I_{(s \leqq t)} g(s, z) \tilde{p}(ds, dz). \tag{7}$$

Equation (2) gives explicit formulas for the integrals on the right.

It is shown below that M^g is a local martingale, and the question is whether all local martingales are of this form for suitable $g \in \mathcal{I}$. As a guide to the answer it is instructive to consider the special case where $(X, \mathcal{S}) = (R, \mathcal{B}(R))$ and μ is absolutely continuous with respect to Lebesgue measure, with density $\psi(t, z)$. Then, from (7) one sees that

$$\begin{aligned} M_t^g = {} & I_{(t \geqq T)} \left\{ g(T, Z) - \int_0^T \int_R \frac{1}{F_s} g(s, z) \psi(s, z)\, dz\, ds \right\} \\ & - I_{(t<T)} \left\{ \int_0^t \int_R \frac{1}{F_s} g(s, z) \psi(s, z)\, dz\, ds \right\}. \end{aligned} \tag{8}$$

If (M_t) is a u.i. martingale with associated function h as in (4), then, comparing the coefficients of $I_{(t \geqq T)}$ in (4) and (8), in order that $M_t = M_t^g$ it is necessary that

$$h(t, z) = g(t, z) - \int_0^t \int_R \frac{1}{F_s} g(s\, z) \psi(s, z)\, dz\, ds.$$

Thus $\eta_t \triangleq g(t, z) - h(t, z)$ must satisfy

$$\eta_t = \int_0^t \int_R \frac{1}{F_s}(\eta_s + h(s, z))\psi(s, z)\, dz\, ds.$$

Putting $\gamma_t = \int_R \psi h\, dz$ and $f_t = \int_R \psi\, dz = dF_t/dt$, we see that this becomes

$$\eta = \frac{f_t}{F_t}\eta_t + \frac{1}{F_t}\gamma_t,$$

$$\eta_0 = 0,$$

which has the unique solution $\eta_t = (1/F_t)\int_0^t \gamma(s)\, ds$, i.e.,

$$g(t, z) = h(t, z) + \eta_t = h(t, z) + \frac{1}{F_t}\int_0^t \int_R h(s, z)\psi(s, z)\, dz\, ds.$$

It is now easily checked that with this choice of g the coefficients of $I_{(t<T)}$ in (4) and (8) agree as well, so that $M_t = M_t^g$ as required. The general result is as follows.

THEOREM 1. *(M_t) is a local martingale of $(\mathcal{F}_t)$ if and only if $M_t = M_t^g$ for some $g \in L^1_{\text{loc}}(p)$.*

Proof. Suppose $g \in L^1_{\text{loc}}$. Calculations similar to those of Proposition 4 show that $M^g_{t\wedge\sigma_k}$ is a u.i. martingale, where (σ_k) is the sequence of stopping times introduced in the proof of Proposition 4. Thus M^g is a local martingale of $(\mathcal{F}_t)$.

Now suppose M is a local martingale of $\mathcal{F}_t$. The situation breaks up into two cases.

Case 1: $c < \infty$, $F_{c-} > 0$. From proposition 2(ii), M_t is u.i. and is therefore given by (4) for some h satisfying (3). We are going to show that $M_t = M_t^g$, where

$$(9)\qquad \begin{aligned} &g(t, z) = h(t, z) + I_{(t<c)}\frac{1}{F_t}\int_{]0,t]\times X} h(s, z)\, d\mu(s, z), && t < \infty,\\ &g(\infty, z_\infty) = 0. \end{aligned}$$

By using (2), (7) can be written as

$$(10)\quad M_t^g = I_{(t\geqq T)}\left\{g(T, Z) - \int_{]0,T]\times X}\frac{1}{F_{s-}}g(s, z)\, d\mu\right\} - I_{(t<T)}\int_{]0,t]\times X}\frac{1}{F_{s-}}g\, d\mu.$$

From (4) and (10), in order that $M_t = M_t^g$ we must have

$$(11)\qquad h(t, z) = g(t, z) - \int_{]0,t]\times X}\frac{1}{F_{s-}}g(s, z)\, d\mu.$$

Now for $t < c$, with g given by (9),

$$\begin{aligned}\int_{]0,t]\times X}\frac{1}{F_{s-}}g(s, z)\, d\mu &= \int_{]0,t]\times X}\frac{1}{F_{s-}}h(s, z)\, d\mu\\ &\quad - \int_{]0,t]}\frac{1}{F_sF_{s-}}\int_{]0,s]\times X} h(u, z)\, d\mu(u, z)\, dF_s\\ &= \int_{]0,t]\times X}\frac{1}{F_{s-}}h\, d\mu + \int_{]0,t]\times X}\left(\int_{[u,t]}\frac{-1}{F_sF_{s-}}dF_s\right)h(u, z)\, d\mu\end{aligned}$$

$$= \int_{]0,t]\times X} \frac{1}{F_{s-}} h \, d\mu + \int_{]0,t]\times X} \left(\frac{1}{F_t} - \frac{1}{F_{u-}}\right) h(u, z) \, d\mu$$

$$= \frac{1}{F_t} \int_{]0,t]\times X} h \, d\mu.$$

Thus (9) is satisfied for $t < c$. A similar calculation shows that the coefficients of $I_{(t<T)}$ agree in (4), (10), and hence that $M_t^g = M_t$ as long as $t < c$. Since M and M^g are stopped at T and $P[T > c] = 0$, it only remains to check that $M_c = M_c^g$ in case $T(\omega) = c$ and this is established, again, by a similar calculation to the above.

We now show that $g \in L^1(p)$. Since M_t is u.i.,

$$\int_Y |h| \, d\mu < \infty.$$

Now using (9),

$$\int |g| \, d\mu \leqq \int |h| \, d\mu - \int_{]0,c[} \frac{1}{F_t} \int_{]0,t]\times X} |h| \, d\mu \, dF_t$$

$$\leqq \int |h| \, d\mu - \frac{1}{F_{c-}} \int_{]0,c[} \int_{]0,t]\times X} |h| \, d\mu \, dF_t$$

$$= \int |h| \, d\mu + \frac{1}{F_{c-}} \int_{]0,c[\times X} (F_t - F_{c-}) |h| \, d\mu$$

$$\leqq \left(1 + \frac{1}{F_{c-}}\right) \int |h| \, d\mu < \infty.$$

Thus $g \in L(p)$ and a fortiori $g \in L^1_{\text{loc}}(p)$.

Case 2: $c = \infty$, *or* $c < \infty$, $F_{c-} = 0$. M_t is a martingale on $[0, c[$ according to Proposition 2. It is therefore u.i. on $[0, t]$ for any $t < c$ and hence of the form (2) for some h satisfying

$$\int_{]0,t]\times X} |h(s, z)| \, d\mu(s, z) < \infty \quad \text{for all } t < c.$$

Calculations as in Case 1 above show that $M_t = M_t^g$ a.s. for g given by (6). Now

$$\int_{]0,t]\times X} |g| \, d\mu \leqq \int_{]0,t]\times X} |h| \, d\mu - \int_{]0,t]} \frac{1}{F_s} \int_{]0,s]\times X} |h| \, d\mu \, dF_s$$

$$\leqq \int_{]0,t]\times X} |h| \, d\mu \left(1 - \int_{]0,t]} \frac{1}{F_s} \, dF_s\right)$$

$$< \infty \quad \text{for } t < c.$$

Thus $g \in L^1_{\text{loc}}(p)$, from Proposition 4. This completes the proof.

3.2. The general case. We now revert to the situation described in § 2. Define

$$p(t, A)(\omega) = \sum_i I_{(t \geqq T_i)} I_{(Z_i \in A)}.$$

Recall that $\Omega_k = \prod_{i=1}^k Y_i$ and $\omega_k : \Omega \to \Omega_k$ is the natural restriction. For $A \in \mathcal{S}$, $k = 2, 3, \cdots$, define $\phi_1^A : R^+ \to R^+$ and $\phi_k^A : \Omega_{k-1} \times R^+ \to R^+$ by

$$\phi_1^A(s) = -\int_{]0,s]} \frac{1}{F_{u-}}\, dF_u^A,$$

$$\phi_k^A(\omega_{k-1}(\omega); s) = -\int_{]0,s]} \frac{1}{F_{u-}^k}\, dF_u^{kA},$$

where F_u^A, F_u are as in § 3.1 and, for $k \geqq 2$,

$$F_u^{Ak} = \mu^k(\omega_{k-1};]u, \infty] \times A),$$
$$F_u^k = F_u^{Xk}.$$

Now define

$$\tilde{p}(t, A)(\omega) = \phi_1^A(T_1) + \phi_2^A(\omega_1; S_2) + \cdots \phi_j^A(\omega_{j-1}; t - T_{j-1}(\omega)) \qquad \text{for } t \in]T_{j-1}, T_j]$$

and

$$q(t, A) = p(t, A) - \tilde{p}(t, A).$$

Calculations as in the proof of Proposition 3 give the following result.

PROPOSITION 6. *For fixed k, and $A \in \mathcal{S}$, $(q(t \wedge T_k, A))_{t \geqq 0}$ is an $\mathcal{F}_t$-martingale.*

The class of integrands $\mathcal{J}$ for stochastic integration is defined as follows. A function $g : \Omega \times Y \to R$ belongs to $\mathcal{J}$ if there exist $g^1 : Y \to R$ and for $k = 2, 3, \cdots$ measurable functions $g^k : \Omega^{k-1} \times Y \to R$ such that

(i)
$$g(t, z, \omega) = \begin{cases} g^1(t, z), & t \leqq T_1(\omega), \\ g^k(\omega_{k-1}(\omega); t, z), & t \in]T_{k-1}(\omega), T_k(\omega)], \\ 0, & t \geqq T_\infty(\omega), \end{cases}$$

(ii)
$$g^1(\infty, z) = g^k(\omega_k; \infty, z) \equiv 0.$$

Now the definitions of $L^1(p)$, etc., read exactly as in § 3.1.

PROPOSITION 7. *Suppose $g \in L^1_{\text{loc}}(p)$ and define*

$$M_t^g = \int_{]0,t] \times X} g(s, z) q(ds, dz).$$

Then there exists a sequence of stopping times $\tau_n < T_\infty$ such that $\tau_n \uparrow T_\infty$ and $M_{t \wedge \tau_n}^g$ is a u.i. martingale for each n.

Proof. Let $\sigma_n \uparrow T_\infty$ be a sequence of stopping times such that $g(s, z) I_{(s \leqq \sigma_n)} \in L^1(p)$, and $\tau_n = T_n \wedge \sigma_n$. Then $\tau_n < T_\infty$ a.s. and calculations similar to those in § 2.2 show that $M_{t \wedge \tau_n}^g$ is a martingale. Thus $M_{t \wedge \tau_n}^g = E[M_{\tau_n} | \mathcal{F}_{t \wedge \tau_n}]$ so that $M_{t \wedge \tau_n}^g$ is uniformly integrable.

Now let $(M_t)_{t \geqq 0}$ be a uniformly integrable martingale of $(\mathcal{F}_t)$. It follows from the martingale convergence theorem (see [6, V T8, T10] that

$$E[M_{T_\infty} | \mathcal{F}_{T_\infty -}] = \lim_k E[M_{T_\infty} | \mathcal{F}_{T_k}] = \lim_k M_{T_k} = M_{T_\infty -}.$$

On the other hand, from Lemma 2, $\mathcal{F}_{T_\infty -} = \mathcal{F}$ so that $E[M_{T_\infty} | \mathcal{F}_{T_\infty -}] = M_{T_\infty}$

a.s. Thus $M_{T_\infty} = M_{T_\infty -}$ a.s., i.e., (M_t) is left-continuous at T_∞. Similarly, (M_t) is stopped at T_∞. This shows that the following formula is true a.s. for all t:

$$M_t = M_{t\wedge T_1} + \sum_{k=2}^{\infty} (M_{t\wedge T_k} - M_{T_{k-1}}) I_{(t \geqq T_{k-1})} \tag{12}$$

because this is an identity if $t < T_\infty$ and the right-hand side is equal to $\lim_k M_{T_k}$ if $t \geqq T_\infty$.

We are now in a position to state the main result.

THEOREM 2. *Let (M_t) be a local martingale of $(\mathscr{F}_t)$. Then there exists $g \in L^1_{\mathrm{loc}}(p)$ such that*

$$M_t - M_0 = \int_{]0,t]\times X} g(s, z) q(ds, dz). \tag{13}$$

Proof. Suppose, to start with, that (M_t) is a u.i. martingale. Define

$$X_t^1 = M_{t\wedge T_1},$$
$$X_t^k = M_{(t+T_{k-1})\wedge T_k} - M_{T_{k-1}}, \qquad k = 2, 3, \cdots.$$

Then in view of (12),

$$M_t = \sum_{k=1}^{\infty} X^k_{(t-T_{k-1})\vee 0}.$$

We can now use the result of Theorem 1 to represent each X^k. Fix k and define for $t \geqq 0$,

$$\mathscr{H}_t = \mathscr{F}_{(t+T_{k-1})\wedge T_k}$$

Now $X_0^k = 0$ and using the optional sampling theorem [6, V T8] we see that (X_t^k) is a martingale of $(\mathscr{H}_t)$; also, from Lemma 1, $\mathscr{H}_t$ is generated by $\mathscr{F}_{T_{k-1}}$ and the sample path of x_s for $s \in]T_{k-1}, T_k]$. Thus there exists a measurable function h^k such that

$$X_t^k = E(h^k(\omega_{k-1}; S_k, Z_k) | \mathscr{H}_t).$$

Since $E|X_t^k| < \infty$ we have

$$\int_{\Omega_{k-1}} \int_Y |h^k(\eta; s, z)| \mu^k(\eta; ds, dz) \nu^{k-1}(d\eta) < \infty,$$

where $\eta \in \Omega_{k-1}$ and ν^{k-1} is the marginal distribution of ω_{k-1}. The argument of Theorem 1 goes through unchanged if μ is a conditional measure given some σ-field. Thus, using Theorem 1 together with Lemma 1, there exists $g^k(\omega_{k-1}; s, z)$ such that

$$X_t^k = \int_{]0,t]\times X} g^k(\omega_{k-1}; s, z) q^k(ds, dz),$$

where $q^k(t, A) = q((t + T_{k-1}) \wedge T_k, A)$ and where g^k satisfies

$$\int_{]0,t]\times X} |g^k| \, d\mu^k \leqq \int_{]0,t]\times X} |h^k| \, d\mu^k \left(1 - \int_{]0,t]} \frac{1}{F_s^k} dF_s^k\right)$$
$$\text{for all } t < c^k(\omega_{k-1}) \triangleq \inf\{t : F^k(\omega_{k-1}; t) = 0\}.$$

The collection $\{g^k, k=1,2,\cdots\}$ defines an integrand $g \in \mathcal{J}$ such that (13) holds a.s. for each t; it remains to show that $g \in L^1_{\text{loc}}(p)$. For $n=1,2,\cdots$, define $s_n^k(\omega_{k-1})$ as follows: if $c^k(\omega_{k-1})$ or $c^k(\omega_{k-1}) < \infty$ and $F^k_{c^k-} < 1/n^3$, set

$$s_n^k(\omega_{k-1}) = \inf\left\{t : F^k(\omega_{k-1}; t) \geqq \frac{1}{n^3}\right\}.$$

If $c^k(\omega_{k-1}) < \infty$ and $F^k_{c^k-} \geqq 1/n^3$, set

$$s_n^k(\omega_{k-1}) = c^k(\omega_{k-1}).$$

Then

$$\int_{]0, s_n^k[} \frac{1}{F_s^k}\, dF_s^k \leqq n^3$$

so that

$$(14) \qquad \int_{\Omega_{k-1}} \int_Y I_{(s<s_n^k)} |g^k|\, d\mu^k\, d\nu^{k-1} \leqq (1+n^3) \int_{\Omega_{k-1}} \int_Y |h^k|\, d\mu^k\, d\nu^{k-1} < \infty.$$

Now define

$$\tau_n = T_j + s_n^j,$$

where

$$j = \min\{k : T_k + s_n^k \leqq T_{k+1}\}.$$

Then τ_n is a stopping time of $\mathcal{F}_t$ and

$$P[\tau_n < T_n] \leqq P \bigcup_{j=1}^{n} (s_n^j < S_j) \leqq n \frac{1}{n^3} = \frac{1}{n^2}.$$

Thus $\sum P[\tau_n < T_n] < \infty$ and hence

$$P[\liminf (\tau_n > T_n)] = 1.$$

It follows that $\tau_n \to T_\infty$ a.s. Now

$$\int_Y g I_{(t \leqq T_n)} p(dt, d\tau) = g^1(T_1, Z_1) + \cdots + g^n(T_1, Z_1, \cdots; S_n, Z_n).$$

Thus, using (14) we see that

$$E \int_Y |g| I_{(t<\tau_n \wedge T_n)} p(dt, dz) \leqq \sum_{k=1}^{n} \int_{\Omega_{k-1}} \int_Y I_{(s<s_n^k)} |g^k|\, d\mu^k\, d\nu^{k-1}$$
$$< \infty.$$

Since $T_n \wedge \tau_n \uparrow T_\infty$ a.s., this shows that $g \in L^1_{\text{loc}}(p)$, as claimed.

If (M_t) is a local martingale with associated stopping time sequence $u_n \uparrow \infty$ such that $M_{t \wedge u_n}$ is a u.i. martingale for each n, then the above argument goes through using $T_n \wedge \tau_n \wedge u_n$ as the stopping time sequence associated with g. This completes the proof.

Remark 1. Suppose $T_\infty \equiv \infty$. Then Proposition 7 and Theorem 2 combine to assert that M is a local martingale if and only if $M = M^g$ for some $g \in L^1_{\text{loc}}(p)$. This

is the result obtained (when the T_i's are totally inaccessible) by Boel, Varaiya and Wong [2]. When T_∞ is possibly finite our results are a slight generalization of Theorem 5.4 of Jacod [9] who, however, does not consider the integrability properties of his integrands g.

Remark 2. In our framework the basic process (x_t) stops at T_∞ and this implies the martingales stop at T_∞ and are left-continuous there. It is possible to generalize the framework so that (x_t) continues beyond T_∞, and to obtain representations for square-integrable martingales. The reader is referred to [8] for details.

4. The local description. In § 2 the probability measure P was specified through the family $\{\mu^k\}$ of conditional distributions. It is also possible to specify it through the "local description" [2], which gives the conditional probabilities of jump occurrence and the distribution of jump location given that one occurs. This is perhaps more natural from the applications point of view, and also has connections with the "Lévy system" for Hunt processes [15]. Let us return to the one-jump situation of § 2.2.

For fixed $A \in \mathcal{S}$ it is evident that the measure on $(R^+, \mathcal{B}(R^+))$ defined by $F_\cdot^A$ is absolutely continuous with respect to that defined by F. Thus there exists a positive real-valued measurable function $n(A, \cdot)$ such that

$$F_0^A - F_t^A = -\int_{]0,t]} n(A, s)\, dF_s. \tag{15}$$

It is easy to see that $n(A, s)$ is also equal to $P[A \times R^+ | T]$. In view of the Blackwell property a regular version of this conditional probability exists, which is assumed to be the one chosen. Then $n(\cdot, s)$ is a probability measure for each s. Now define

$$\tilde{\Lambda}(t) = \Lambda(t \wedge T),$$

where

$$\Lambda(t) = -\int_{]0,t]} \frac{1}{F_{s-}}\, dF_s. \tag{16}$$

The pair (n, Λ) is called the local description on account of the probabilistic interpretation

$$d\Lambda(t) = P[T \in\ ]t, t+dt] \,|\, T \geqq t],$$

$$n(A, s) = P[Z \in A \,|\, T = s].$$

In terms of the local description the compensator $\tilde{p}(t, A)$ can be written as

$$\tilde{p}(t, A) = \int_0^t n(A, s)\, d\tilde{\Lambda}(s).$$

This may be compared with equation (3.3) of [15].

The functions (n, Λ) have the following properties:

(i) $\Lambda(t)$ is defined for $t \in [0, c[$,

(ii) $\Lambda(t)$ is increasing and right-continuous; $\Lambda(0) = 0$,

(iii) $\Delta\Lambda(t) < 1$,

(iv) $n(A, s) \geqq 0$,

(v) $n(A, \cdot)$ is measurable for fixed A,

(vi) for all $s \in]0, c[$ except a set of $d\Lambda$-measure 0, $n(\cdot, s)$ is a probability measure on $(X, \mathcal{S})$, and $n(\cdot, c)$ is a probability measure if $c < \infty$, $\Lambda(c-) < \infty$.

PROPOSITION 8. *There is a bijective correspondence between the set of probability measures μ on $(Y, \mathcal{Y})$ and the set of local descriptions (n, Λ) satisfying* (i)–(vi) *above.*

Proof. There is a bijection between the set of functions Λ satisfying (i)–(iii) above and the set of probability distributions F on $]0, \infty]$ given by (15) and

$$
\begin{aligned}
F_t &= e^{-\Lambda(t)} \prod_{s \leqq t} (1 - \Delta\Lambda(s)) e^{\Delta\Lambda(s)}, \quad && t < c, \\
F_t &= 0, && t \geqq c.
\end{aligned} \tag{17}
$$

See Jacod [9, Lem. 3.5]. Thus given (n, Λ) satisfying (i)–(vi), the distribution F^A is specified for each $A \in \mathcal{S}$ by (17) and (15) and this uniquely defines the measure μ. Conversely μ defines (n, Λ) via (1), (15) and (16).

This result shows that there is a single-jump process corresponding to an arbitrary local description (n, Λ). Clearly the same is true of the jump process introduced in § 2.1, if by a local description we understand a family of functions $\{(n^k, \Lambda^k),\ k = 1, 2 \cdots\}$ corresponding to the family of conditional distributions $\{\mu^k\}$.

Acknowledgment. In conclusion, I would like to express my thanks to T. Kailath for providing the initial impetus for this work, and to J. M. C. Clark and J. Treuherz for helpful discussions.

REFERENCES

[1] R. BOEL AND P. VARAIYA, *Control of jump processes*, Electronics Research Lab. Memo., Univ. of California, Berkeley, 1974;-this Journal, to appear.

[2] R. BOEL, P. VARAIYA AND E. WONG, *Martingales on jump processes, Part I: Representation results*; *Part II: Applications*, this Journal, 13 (1975), pp. 999–1061.

[3] P. BRÉMAUD, *A martingale approach to point processes*, Electronics Research Lab. Memo M-345, Univ. of California, Berkeley, 1972.

[4] CHOU CHING-SUNG AND P. A. MEYER, *Sur la representation des martingales comme intégrales stochastiques dans les processus ponctuels*, Séminaire de Probabilités IX, Lecture Notes in Mathematics, vol. 465, Springer-Verlag, Berlin, 1975.

[5] P. COURRÈGE AND P. PRIOURET, *Temps d'arrêt d'une fonction aléatoire*, Publ. Inst. Statist. Univ. Paris, 14 (1965), pp. 245–274.

[6] C. DELLACHERIE, *Capacités et processus stochastiques*, Springer-Verlag, Heidelberg, 1972.

[7] C. DOLÉANS-DADE AND P. A. MEYER, *Intégrales stochastiques par rapport aux martingales locales*, Séminaire de Probabilités IV, Lecture Notes in Mathematics, vol. 124, Springer-Verlag, Berlin, 1970.

[8] R. J. ELLIOTT, *Martingales of a jump process with partially accessible jump times*, preprint, Dept. of Pure Mathematics, Univ. of Hull, Hull, England, January 1975.

[9] J. JACOD, *Multivariate point processes: Predictable projection, Radon–Nikodym derivatives, representation of martingales*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 31 (1975), pp. 235–253.

[10] R. S. LIPTSER AND A. N. ŠIRYAEV, *Statistics of Stochastic Processes*, Izdatyelstvo Nauka, Moscow, 1974.
[11] P. A. MEYER, *Probability and Potentials*, Blaisdell, Waltham, Mass., 1966.
[12] ———, *Temps d'arrêt algébriquement prévisibles*, Séminaire de Probabilités VI, Lecture Notes in Mathematics, vol. 258, Springer-Verlag, Berlin, 1972.
[13] R. W. RISHEL, *A minimum principle for controlled jump processes*, Control Theory, Numerical Methods and Computer Systems Modelling, Lecture Notes in Economics and Mathematical Systems, vol. 107, Springer-Verlag, Berlin, 1975.
[14] A. SEGALL, M. H. A. DAVIS AND T. KAILATH, *Nonlinear filtering with counting observations*, IEEE Transactions on Information Theory, IT-21 (1975), pp. 143–149.
[15] S. WATANABE, *On discontinuous additive functionals and Levy measure of a Markov process*, Japanese J. Math., 34 (1964), pp. 53–70.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

A SPECTRAL FACTORIZATION APPROACH TO THE DISTRIBUTED STABLE REGULAR PROBLEM; THE ALGEBRAIC RICCATI EQUATION*

J. WILLIAM HELTON†

Abstract. This paper is a study of the discrete-time infinite-dimensional "stable regulator problem" having a cost function which is not necessarily positive. We take a spectral factorization approach to the problem. Also there are results on the algebraic Riccatic equation which are equivalent to results about fixed points for a broad class of symplectic maps.

Introduction. This paper is a study of the infinite-dimensional "stable regulator problems" having a cost functional which is not necessarily positive. The control problem will have a solution or approximate solution in feedback form provided that one "completes" a certain square in a way familiar in control theory. In this paper, we use a spectral factorization method to obtain necessary and sufficient conditions for this to be possible (§ 2). Section 3 describes the stability of the feedback system resulting from the optimal control problem. Section 4 treats the infinite-dimensional algebraic Riccati equation associated with the control problem. This can also be described as a study of the fixed-point problem for certain infinite-dimensional "symplectic" maps (see Appendix plus § 4).

This paper follows in the footsteps of a paper by Willems [25] in which he gives necessary and sufficient conditions for solving a broad class of finite-dimensional continuous time algebraic Riccati equations. In addition to giving a discrete-time and an infinite-dimensional version of these results, our article gives proofs which in finite dimensions are rather simple. As this paper was being written, an elegant spectral factorization approach to Willems results was given by Molinari [15][16] and then applied to the stable regulator in [17]. His proof involves some basically finite-dimensional methods such as determinants and dimension counting while the key step in the proof here is subspace inclusion. The article [14] is a good reference for infinite-dimensional discrete-time systems having "positive cost operators". Our article gives a new approach to the time-invariant regulator results in that paper and extends them in several directions.

The results in this article apply to most least squares problems associated with the discretization of systems governed by a heavily damped variable coefficient wave equation (including the heat equation). A thorough list of applications of the finite-dimensional theory appears in [25]. Since we do not require our "cost operators" to be positive, the systems studied are capable of storing energy, that is, "cost". The basic principal which emerges in [25] and which is true to a large extent in infinite dimensions is that one can use the standard feedback approach to a control problem provided the zero state stores no energy. Roughly speaking,

---

* Received by the editors May 3, 1974, and in revised form September 25, 1975.

† Department of Mathematics, University of California, San Diego, La Jolla, California 92037.
This work was supported in part by the National Science Foundation.

conventional approaches suffice even when the system can store energy, but not when it can spontaneously produce energy.

1. Definitions and setting. We shall consider a system

$$x_{i+1} = Ax_i + Bu_i, \tag{1.1}$$

where the x_j are vectors in a Hilbert space $\mathcal{H}$ and the u_j are vectors in a Hilbert space $\mathcal{U}$. The cost of running the system from initial state x_0 for N time units is

$$J_N(x_0, u) = \sum_{i=0}^{N} [(x_i, Qx_i) + (u_i, Ru_i)]. \tag{1.2}$$

Here $Q : \mathcal{H} \to \mathcal{H}$ and $R : \mathcal{U} \to \mathcal{U}$ are bounded self-adjoint operators. The basic infinite time interval problem is: given a state x_0, determine exact or approximate a control sequence u which minimizes $J_\infty(x_0, \ )$. The admissible class of control sequences $u = \langle u_i \rangle_{i=1}^{\infty}$ in this paper will be those in $l^2(0, \infty, \mathcal{U})$, the set of all sequences from $\mathcal{U}$ whose norms are square summable. Also one usually requires that $x_n \to 0$ in some sense as $n \to \infty$. We shall always assume that A is stable, i.e., $\|A^n\| < M$ for all n, and that B is bounded.

Frequently in what follows it will be convenient to look at our system as one having an output. The natural choice for the output operator is $|Q|^{1/2}$. Thus the problem is equivalent to minimizing the cost

$$\sum (y_i, [\operatorname{sgn} Q] y_i) + (u_i, Ru_i) \tag{1.3}$$

of running the system

$$x_{i+1} = Ax_i + Bu_i, \qquad y_i = |Q|^{1/2} x_i. \tag{1.4}$$

Here sgn Q is the operator $P_+ - P_-$ on $\mathcal{H}$, where $P_+(P_-)$ is the projection onto the positive (negative) spectral subspace of Q. The frequency response function for the system $[A, B, |Q|^{1/2}]$ is

$$W(z) = z|Q|^{1/2}(I - zA)^{-1}B. \tag{1.5}$$

Since A is stable, the spectrum of A is contained inside the disk, and so $W(z)$ is well-defined and analytic inside the disk. It will be assumed that all systems we study have a uniformly bounded frequency response function. Let $l^2(0, \infty, \mathcal{U})$ denote all sequences $\langle u_i \rangle_{i=0}^{\infty}$ from $\mathcal{U}$ with square summable norm.

If $Q \geqq 0$, then it can be shown (see equations (2.2) and (2.3)) that the cost $J_\infty(0, u)$ of running the system initially at state 0 with l^2 input u is finite if and only if the frequency response function W is uniformly bounded on the disk. In dealing with the signed Q problem, we shall *always* assume that the cost $J_\infty(0, u)$ with $|Q|$ replacing Q is finite. This is also equivalent to the statement $W(z)$ is uniformly bounded on the unit disk, and we shall say that any (1.1) and (1.2) with this property have *absolutely finite cost.* This assumption will obviously be satisfied when A is very stable, for example, if (1.1) arises from discretizing a variable coefficient heat equation or heavily damped wave equation. This assumption can certainly be relaxed but the author suspects that the basic structure and proofs will change little while the technical complication will greatly increase. Thus it seems unwise to do so without first making a systematic list of compelling examples.

Henceforth, assume that (1.1) and (1.2) have absolutely finite cost. It is well known (see, [18, Chap. V]) that radial limits of such functions exist almost everywhere onto the unit circle, so we may consider W as being a function $W(e^{i\theta})$ defined for almost all θ.

The mathematical notation will be as follows. The unit circle will be denoted by $\mathbf{T}$, the set of all complex numbers by $\mathbb{C}$. If $\mathcal{H}$ is a separable complex Hilbert space, then $L^2(\mathcal{H})$ denotes the Hilbert space of norm square integrable Lebesgue-measurable $\mathcal{H}$-valued functions. We let $H^2(\mathcal{H})$ [resp., $\bar{H}^2(\mathcal{H})$] denote the closed subspace of functions in $L^2(\mathcal{H})$ with zero nonpositive (positive) Fourier coefficients. The operator P_{H^2} [resp., $P_{\bar{H}^2}$] is the orthogonal projection of L^2 onto this subspace. If $\mathcal{H}_1$ and $\mathcal{H}_2$ are separable complex Hilbert spaces, then $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ denotes the Banach space of bounded linear transformations from $\mathcal{H}_1$ to $\mathcal{H}_2$. We abbreviate $\mathcal{L}(\mathcal{H}, \mathcal{H})$ as $\mathcal{L}(\mathcal{H})$. Moreover $L^\infty(\mathcal{H}_1, \mathcal{H}_2)$ denotes the Banach space of essentially bounded weakly-measurable $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$-valued functions on $\mathbf{T}$, while $H^\infty(\mathcal{H}_1, \mathcal{H}_2)$ (resp., $\bar{H}^\infty(\mathcal{H}_1, \mathcal{H}_2)$) denotes the subspace of functions with negative (resp., nonnegative) Fourier coefficients equal to zero. When the context prevents ambiguities, we will only write H^∞ and H^2.

Functions in the Hardy spaces $H^2(\mathcal{H})$ [resp., $\bar{H}^2(\mathcal{H})$ and $H^\infty(\mathcal{H}_1, \mathcal{H}_2)$] can be identified with boundary values of functions analytic inside [resp., outside] the unit disk; see [18, Chap. V]. If φ is a function in $L^\infty(\mathcal{H}_1, \mathcal{H}_2)$, then $\mathcal{M}_\varphi$ is the operator from $L^2(\mathcal{H}_1)$ to $L^2(\mathcal{H}_2)$ defined by $(\mathcal{M}_\varphi f)(e^{i\theta}) = \varphi(e^{i\theta}) f(e^{i\theta})$ for f in $L^2(\mathcal{H}_1)$. A function φ in $H^\infty(\mathcal{H}_1, \mathcal{H}_2)$ is called *outer* provided that $\mathcal{M}_\varphi$ restricted to $H^2(\mathcal{H}_1)$ has dense range in $H^2(\mathcal{H}_2)$. A function φ in $\bar{H}^\infty(\mathcal{H}_1, \mathcal{H}_2)$ is called *conjugate* outer provided that it has the analogous properties on $\bar{H}^2(\mathcal{H}_1)$ and $\bar{H}^2(\mathcal{H}_2)$. Such a function is called *invertible* outer if its pointwise inverse is uniformly bounded. The invertible outer functions are precisely those in $H^\infty(\mathcal{H}_1, \mathcal{H}_2)$ whose inverse is in $H^\infty(\mathcal{H}_2, \mathcal{H}_1)$.

2. Optimality. In an infinite-dimensional problem, it is reasonable to expect the frequent occurrence of unbounded operators. This is so here. In fact, more unwieldy objects are necessary. For example, the cost of optimally driving a state x_0 to zero may not be finite on all states x_0 of the system, while it will frequently be finite for all states which actually occur in the running of the system. Thus it is only reasonable to expect the optimal cost functional to be a densely defined (quadratic) functional on the state space, and indeed that is what will be obtained. A good reference on such objects is [20, Chap. VIII]. The controllability map of a system $[A, B]$ is the densely defined map $\mathcal{C} : l^2(0, \infty, \mathcal{U}) \to \mathcal{H}$ given by

$$\mathcal{C}(u_0, u_1, \cdots) = \sum_{k=0}^{\infty} A^k B u_k.$$

We set $\mathcal{R} = \text{range } \mathcal{C}$ and $\tilde{\mathcal{R}} = \mathcal{C}$ (all sequences with only finite number of nonzero terms). These are domains which will be commonly used.

A trajectory of the system initially at 0 is a sequence of states $\{x_n\}_{n=0}^\infty$ which results from feeding some input sequences $\{u_i\}$ into the system. Unless otherwise specified, *trajectory* will refer to something arising from a $l^2(0, \infty, \mathcal{U})$ input. The finite cost assumption implies that $J_\infty(0, u)$ is well-defined and finite for all u in $l^2(\mathcal{U})$. In this section, we shall deal only with J_N for which $J_\infty(0, u) > 0$, all u in

$l^2(\mathcal{U})$. In this case, we may think of the quadratic functional $J_\infty(0, u)$ as giving a second degenerate norm on $l^2(\mathcal{U})$. Let σ be the completion of the orthogonal complement of $\mathcal{N}$ the nullspace of J_∞ in the J_∞-norm. This describes a space of input strings having finite cost. The elements of σ are not necessarily sequences; they are equivalence classes of sequences. However, under many circumstances they may be identified directly as sequences. This will be the case if $\mathcal{U}$ is finite-dimensional or, more generally, if $E(e^{i\theta})$ has closed range for almost all θ.

We shall approach the problem in the usual way (see [1], [2]), namely, by completing the square to put the cost functional J_N in a reduced form from which the optimal control law is apparent. This section is devoted to determining when this procedure is possible.

DEFINITION. The cost functional J_N for system (1.1) and (1.2) is in *reduced form* when there exists an auxiliary Hilbert space $\mathcal{H}_1$ with inner product $\langle \cdot, \cdot \rangle$, a continuous operator $G : \mathcal{U} \to \mathcal{H}_1$, an (possibly unbounded) operator $F : \mathcal{R} \to \mathcal{H}_1$, and a symmetric bilinear form $K(,)$ whose domain contains $\tilde{\mathcal{R}}$ such that for all N,

$$(2.1) \qquad J_N(x_0, u) = \sum_{i=0}^{N} \langle Gu_i + Fx_i, Gu_i + Fx_i \rangle + K(x_0, x_0) - K(x_{N+1}, x_{N+1})$$

for $u = \langle u_i \rangle$, an input to system (1.1), and x_i the corresponding states of the system.

If $K(x_N, x_N) \to 0$ along trajectories, then $J_\infty(0, u)$ is always $\geqq 0$, and the finite cost assumption implies that the map μ defined on $l^2(\mathcal{U})$ sequences by $\mu\{u_i\} = \{Gu_i + Fx_i\}$ satisfies $c\|u\|^2 \geqq J_\infty(0, u) = \sum_{i=0}^{\infty} \|(\mu u)_i\|^2_{\mathcal{H}_1}$; thus $\mu : l^2(\mathcal{U}) \to l^2(\mathcal{H}_1)$. A reduced form is called *outer* if μ has range dense in $l^2(\mathcal{H}_1)$. One could view the map μ as taking a dense subspace of σ isometrically to a dense subspace of $l^2(\mathcal{H}_1)$, and so we may extend μ to a unitary map $\tilde{\mu} : \sigma \xrightarrow{\text{onto}} l^2(\mathcal{H}_1)$.

Proceed formally for a moment. Once the reduced form is obtained, then to minimize J_∞ over inputs u with trajectories x_N on which $K(x_N, x_N) \to 0$, one would solve $Gu_i = -Fx_i$ to obtain a control sequence $\{u_i\}$. If $K(x_N, x_N) \to 0$ on the resulting trajectory, then clearly $\{u_i\}$ is the optimal control and the optimal cost is $K(x_0, x_0)$. To make this argument rigorous, let x_0 be the initial state to be controlled and let v be an input sequence of finite length n whose associated state sequence has $x_{n+1} = x_0$. The approach just described consists of extending v to an element $w = (v_0, \cdots, v_n, u_0, u_1 \cdots)$ abbreviated (v, u) in σ with the property that each entry of the sequence $\tilde{\mu}u$ beyond the nth is zero. Such a u can be obtained by solving $\tilde{\mu}u = -(y_{n+1}, y_{n+2}, \cdots)$ where $(y_0, y_1, \cdots) = \mu v$. This is because $\tilde{\mu}(v, u) = \mu(v, 0) + \tilde{\mu}(0, u) = (y_0, y_1, \cdots) - (0, \cdots, 0, y_{n+1}, \cdots)$. The element u of σ is an optimal control (also the unique one in σ) provided that $K(x_N, x_N) \to 0$ on the trajectory arising from u. It turns out that about the best we can expect is $K(x_N, x_N) \to 0$ along trajectories coming from $l^2(\mathcal{U})$ input strings. Under this circumstance, given $\varepsilon > 0$, any $u(\varepsilon)$ in $l^2(\mathcal{U})$ with $J_\infty(0, u - u(\varepsilon)) < \varepsilon$ is a control sequence which runs the system at within ε of $K(x_0, x_0)$, the optimal cost. Thus we have a reasonable sense in which to think of u as the optimal control. The approximate control is an approximate solution to $Gu_i = -Fx_i$ in the rather strong sense that $\sum_{i=0}^{\infty} \|Gu_i(\varepsilon) + Fx_i(\varepsilon)\|^2 < \varepsilon$. Conversely, any $l^2(\mathcal{U})$ control u' within ε of being optimal satisfies $\varepsilon > J_\infty(0, u') - K(x_0, x_0) \geqq \sum_{i=0}^{\infty} \|Gu_i' + Fx_i\|^2$. Thus *the control problem is solved provided that J_N can be put in outer reduced form with $K(x_N, x_N) \to 0$ along any trajectory coming from an $l^2(\mathcal{U})$ input.* The goal of this

section is to show when this can be done. After that is finished, we give some conditions under which the actual controlling sequence $u(\varepsilon)$ can be expressed concretely in terms of u.

The results of this section are given in terms of the power spectrum operator:

$$E(e^{i\theta}) = R + W(e^{i\theta})^*[\operatorname{sgn} Q]W(e^{i\theta}) \tag{2.2}$$

which is defined for almost all $e^{i\theta}$ on $\mathbf{T}$. In particular, we shall be concerned with spectral factorizations of it. Recall that an $\mathscr{L}(\mathscr{U})$ function $P(e^{i\theta})$ has a spectral factorization if it can be written in the form

$$P(e^{i\theta}) = M(e^{i\theta})^*M(e^{i\theta}),$$

where M is in $H^\infty(\mathscr{U}, \mathscr{H}_1)$ for some auxiliary Hilbert space $\mathscr{H}_1$. Given a nonnegative operator or matrix-valued function, a spectral factorization may or may not exist. This is a classical question, and the answer is that a factorization usually exists. In recent engineering literature, the results of Gohberg–Krein [10] are usually cited; however, necessary and sufficient conditions are available (see [18, Chap. V, § 4], [23]). The more applicable sufficient conditions for nonnegative P are

(I) $P(e^{i\theta}) \geqq \delta(e^{i\theta})I$ with δ a log integrable function [18. Chapt. V, § 7].

(II) For P matrix-valued log det $P(e^{i\theta})$ is integrable [11, Thm. 18].

(III) P has a (pseudo) meromorphic continuation to $\mathbb{C}$ [23, Thm. 3.1]. This includes the case where P is a rational function.

Thus there are quite a few ways to check if a function has a spectral factorization and so the hypotheses of the theorems appearing in this section are hopefully easy to apply.

Next we shall observe that placing J_N in reduced form is related to spectral factorization of E. Let $u(e^{i\theta})$ denote the Fourier transform of $\{u_n\}$ in l^2; it is a function in $H^2_{\mathbf{T}}(\mathscr{U})$. Fourier transforming (1.1) and the definition (1.2) of the cost function J_∞ gives

$$J_\infty(0, u) = \int_{-\pi}^{\pi} (u(e^{i\theta}), E(e^{i\theta})u(e^{i\theta}))\, d\theta. \tag{2.3}$$

Thus E is closely related to $J_\infty(0, u)$. Suppose that J_N is in reduced form with $K(x_N, x_N) \to 0$ on each trajectory. Then (2.1) can be Fourier transformed to give

$$J_\infty(0, u) = \int_{-\pi}^{\pi} (u(e^{i\theta}), M(e^{i\theta})^*M(e^{i\theta})u(e^{i\theta}))\, d\theta,$$

where M is the uniformly bounded function

$$M(z) = G + zF(I - zA)^{-1}B. \tag{2.4}$$

Comparing this with expression (2.3) for $J_\infty(0, u)$, we find that the Toeplitz operator generated by $E - M^*M$ is identically zero; thus (see [19]) we get that

$$E = M^*M.$$

That is, E has a spectral factorization. The main theorem of this section says that not only this but its converse is true.

THEOREM 2.1. *Suppose that the reachable states $\tilde{\mathcal{R}}$ for the system $[A, B]$ are dense in its state space and that the cost functional J_N is absolutely finite. Then the power density function $E(e^{i\theta})$ has a spectral factorization if and only if the cost J_N can be put in outer reduced form with an optimal cost functional K satisfying $K(x_N, x_N) \to 0$ on trajectories $\{x_N\}$ of the system arising from $l^2(\mathcal{U})$ inputs.*

Proof. One side of the theorem has already been proved. The converse requires the rest of this section. By hypothesis, E has a spectral factorization. Spectral factorizations are not unique and not all of these factorizations have the required form (2.4). However, since E has a factorization, it has [18, Chap. V, § 4] an outer factorization $M \in H^\infty(\mathcal{U}, \mathcal{H}_1)$ which is unique up to a constant multiple and which we will now prove has the form (2.4). The proof relies on realizability theory, in particular, on that developed in [8] in the one-dimensional case, in greater generality in [12], and surveyed in [13].

We begin with a quick sketch of realizability theory. The system $[A, B]$ is called *approximately* (*exactly*) controllable if the range of $\mathcal{C}$ is dense in $\mathcal{H}$ (is all of $\mathcal{H}$). It is *continuously controllable* if $\mathcal{C}$ is a continuous map. Exact controllability is equivalent to the standard pseudoinverse $\mathcal{C}^{-1}$ of $\mathcal{C}$ being a continuous operator. This follows immediately from the open mapping theorem. Similar considerations with adjoint systems give the obvious notions of *approximate* (*exact*) *observability*. A slight modification of Theorem 3C.1 of [12] is the

REALIZABILITY THEOREM. *Any $\mathcal{L}(\mathcal{U}, \mathcal{H}_1)$-valued function $F(z)$ analytic and bounded on the unit disk is the frequency response function of some exactly observable and approximately controllable system $[A, B, C, D]$.*

The operators A, B, C, D in the theorem are given explicitly: A is the restriction of $P_{H^2(\mathcal{H}_1)}\mathcal{M}_{e^{-i\theta}}$ to the subspace $X = \text{cl } P_{H^2(\mathcal{H}_1)}\mathcal{M}_F\bar{H}^2(\mathcal{U})$, $B : \mathcal{U} \to X$ is given by $Bu_0 = P_{H^2(\mathcal{H}_1)}\mathcal{M}_{F(e^{i\theta})}u_0e^{-i\theta}$, C is the projection of X onto the subspace of constant functions in $H^2(\mathcal{H}_1)$ and D is $F(0)$. The space X is the state space for the system. This particular realization of F is called the restricted shift realization by Fuhrmann. A fact critical to our control problem can be read off from this construction.

LEMMA 2.2. *If two functions $T_1(z)$ and $T_2(z)$ with $\mathcal{L}(\mathcal{S}_1, \mathcal{S}_2)$ and $\mathcal{L}(\mathcal{S}_3, \mathcal{S}_2)$ values, respectively, satisfy the hypotheses of the Realizability Theorem, if in the above representation, $T_2(z) = D + zC(I - zA)^{-1}B$, and if the state space X_1 for T_1 is contained in the state space X_2 for T_2, then $T_1(z)$ can be written in the form $T_1(z) = D_1 + zC(I - zA)^{-1}B_1$.*

It is now easy to show that M has the realization (2.4). Since M is in $H^\infty(\mathcal{U}, \mathcal{H}_1)$, the function $\tilde{M}$ defined by $\tilde{M}(e^{i\theta}) = M(e^{-i\theta})^*$ is in $H^\infty(\mathcal{H}_1, \mathcal{U})$. We now compare $\tilde{M}$ to the function $\tilde{W}$ defined by (1.5) using Lemma 2.3. Since M is outer,

$$X_{\tilde{M}} = \text{cl } (P_{H^2(\mathcal{U})}\mathcal{M}_{\tilde{M}}\bar{H}^2(\mathcal{H}_1)) = \text{cl } (P_{H^2(\mathcal{U})}\mathcal{M}_{\tilde{E}}\bar{H}^2(\mathcal{U}))$$

which, in turn, by the definition of E, is contained in

$$X_{\tilde{W}} = \text{cl } (P_{H^2(\mathcal{U})}\mathcal{M}_{\tilde{W}}\bar{H}^2(\mathcal{H})).$$

If $\tilde{W}$ has restricted shift realization $[\psi, \Delta, \Lambda]$, then $\tilde{M}$ has a realization $[\psi, \alpha, \Lambda, \rho]$.

The function W has two realizations $[\psi^*, \Lambda^*, \Delta^*]$ and $[A, B, |Q|^{1/2}]$. A straightforward infinite-dimensional version of the state space isomorphism theorem [12, Thm. 3b.1] says that if $[A, B, |Q|^{1/2}]$ is approximately observable, there exists a 1–1 densely defined operator $\beta : \mathcal{H} \to X_{\tilde{W}}$ such that

$$\psi^*\beta = \beta A, \quad \Lambda^* = \beta B, \quad |Q|^{1/2} = \Delta^*\beta.$$

This implies that M has the realization $[A, B, \alpha^*\beta, \rho^*]$, that is, M has the representation (2.4) with $F = \alpha^*\beta$ and $G = \rho^*$.

To finish the theorem, we require some fine structure from Theorem 3b.1 of [12]. It is shown there under the assumption of continuous controllability and approximate observability that β is $\dot{\mathcal{C}}\mathcal{C}^{-1}$, where $\mathcal{C}$ (resp., $\dot{\mathcal{C}}$) is the controllability operator of $[A, B, |Q|^{1/2}]$ (resp., $[\psi^*, \Lambda^*, \Delta^*]$) and $\mathcal{C}^{-1}$ is a pseudoinverse of $\mathcal{C}$. These results extend immediately to the case at hand and validate this definition of β provided that it is interpreted as follows. If $y \in \mathcal{R}$, there is u such that $\mathcal{C}u = y$ define $\beta y = \dot{\mathcal{C}}u$. To check that this is not ambiguous, note that by the lemma in [12] null $\mathcal{C} \subset$ null $\mathcal{Q}^*\mathcal{C} =$ null Hankel$_W =$ null $\dot{\mathcal{Q}}^*\dot{\mathcal{C}}$ and since $\dot{\mathcal{Q}}^*$ is 1–1 this equals null $\dot{\mathcal{C}}$; thus if $\mathcal{C}u = 0$, then $\dot{\mathcal{C}}u = 0$. The construction in the theorem can be completed by setting $F = \alpha^*\beta$. Note that when one does not have approximate controllability, β will not be 1–1; in finite dimensions, for example, null $\beta =$ (range $\mathcal{Q})^{\perp}$. Also observe that $F\mathcal{C} = \alpha^*\dot{\mathcal{C}}$ which is a continuous operator.

Now we must show that having the appropriate factorization for E implies that J_N can be put in reduced form. To see this we first observe

LEMMA 2.3. *The cost functional J_N can be written in reduced form* (2.1) *if and only if there exist appropriately defined F, G and $K(\cdot, \cdot)$ which satisfy*

$$\langle Gu_0, Gu_0\rangle = (u_0, Ru_0)_{\mathcal{U}} + K(Bu_0, Bu_0), \tag{2.5a}$$

$$\langle Fx, Gu_0\rangle = K(Ax, Bu_0), \tag{2.5b}$$

$$\langle Fx, Fy\rangle = (x, Qy)_{\mathcal{H}} + K(Ax, Ay) - K(x, y) \tag{2.5c}$$

for x, y in $\tilde{\mathcal{R}}$.

Proof. One simply substitutes (2.5) into the right side of (2.1) and observes, after using (1.1), that (1.2) the definition of J_N has been obtained.

As one might expect, the operators F, G, A, B and the space $\mathcal{H}_1$, appearing in the representation (2.4) for M, will turn out to be the operators required in the lemma. Here we let $\langle \cdot, \cdot \rangle$ denote the inner product on $\mathcal{H}_1$. The optimal cost form $K(\cdot, \cdot)$ is yet to be constructed. Formally, it is, for $x, y \in \tilde{\mathcal{R}}$,

$$K(x, y) = \sum_{j=0}^{\infty} (A^j x, [Q - F^*F]A^j y),$$

and it is not too difficult to check that this formally satisfies (2.5). For example, if this were a finite-dimensional problem, a very simple manipulation would finish the proof. However, our task is a bit tiresome.

Now we give the precise definition of $K(\cdot, \cdot)$. Set

$$\begin{aligned} L(e^{i\theta}) &= E(e^{i\theta}) - [M(e^{i\theta}) - G]^*[M(e^{i\theta}) - G] - R \\ &= G^*M(e^{i\theta}) + M(e^{i\theta})^*G - G^*G - R. \end{aligned} \tag{2.6}$$

Formally one should think of this as

$$L(e^{i\theta}) = B^*(e^{-i\theta} - A^*)^{-1}[Q - F^*F](e^{i\theta} - A)^{-1}B.$$

Define

$$K(A^l Bu_0, Bv_0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} (u_0, L(e^{i\theta})v_0)_{\mathcal{U}}\, e^{+il\theta}\, d\theta \tag{2.7a}$$

and

$$K(Bu_0, A^l Bv_0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} (u_0, L(e^{i\theta})v_0)_{\mathcal{U}}\, e^{-il\theta}\, d\theta \tag{2.7b}$$

for $l \geqq 0$. Next we use (2.5c) to define $K(,)$ inductively; it is

$$\begin{aligned} K(A^{l+1}Bu_0, A^{j+1}Bv_0) &= \langle FA^l Bu_0, FA^j Bv_0\rangle \\ &\quad -(A^l Bu_0, QA^j Bv_0)_{\mathcal{H}} - K(A^l Bu_0, A^j Bv_0) \end{aligned} \tag{2.7c}$$

for u_0, v_0 in $\mathcal{U}$. Note that the term involving F is well-defined on $\tilde{\mathcal{R}}$, and so we have defined a function. After some tedious work, which we leave as an exercise, one can show that K is a consistently defined bilinear functional on $\tilde{\mathcal{R}}$.

By construction, $K(,)$ satisfies (2.5c). The identity (2.5a) follows by setting $l = 0$ and performing the integration on the right side of (2.7a) while observing that $1/2\pi \int_{-\pi}^{\pi} M(e^{i\theta})\, d\theta = G$. The identity (2.5b) follows from (2.7) and the fact that

$$\langle FA^l Bv_0, Gu_0\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} (v_0, L(e^{i\theta})u_0)_{\mathcal{U}}\, e^{(l+1)\theta}\, d\theta.$$

Only one property of K remains unverified; that is, $K(x_n, x_n) \to 0$. This follows because $J_\infty(0, u) = (1/(2\pi)) \int_{-\pi}^{\pi} (u, M^*Mu) = \sum_{n=0}^{\infty} \|Gu_i + Fx_i\|^2$, and so (2.1) implies that $K(x_n, x_n) \to 0$. The proof of Theorem 2.1 is finished.

The theorem just completed shows that an approximate control sequence always exists. During the remainder of the section, we describe ways for identifying approximate control sequences explicitly. By the discussion preceding Theorem 2.1, we are confronted with the problem: Given the outer factorization M of E and $y(e^{i\theta})$ in $H^2(X)$ (actually, we may take y to be a polynomial in $e^{i\theta}$), for each $\varepsilon > 0$, find $u_\varepsilon(e^{i\theta})$ in L^2 such that $\int_0^{2\pi} \|y - Mu_\varepsilon\|_X^2 < \varepsilon$. The Fourier transform $u(\varepsilon)$ of u_ε is an l^2-sequence which yields an ε-approximate optimizing control. We know that such u_ε must exist, but the problem is to give a method for finding them explicitly.

We begin by treating the case where E is scalar-valued. The function $u(z) \triangleq M(z)^{-1} y(z)$ is analytic on the disk, but can have fearsome boundary values $u(e^{i\theta})$. Standard ways to approximate $u(e^{i\theta})$ with L^2-functions are

(i) $[A_r u](e^{i\theta}) = u(re^{i\theta})$, Abel approximation,

(ii) $[P_N u](e^{i\theta}) = \sum_{j=0}^{N} u_j\, e^{ij\theta}$, Fourier approximation,

(iii) $[F_N u](e^{i\theta}) = (1/(N+1)) \sum_{K=0}^{N} P_K$, Cesaso approximation,

and the question before us is: Does $\int_0^{2\pi} \|y - MA_r u\|^2$,etc., go to zero? Since $y = Mu$ and $E = M^*M$, an alternative phrasing of this question is: Does $A_r u \to u$; $P_N u \to u$, *or* $F_N u \to u$ in $L^2(E\,d\theta)$? Fortunately, these are standard questions in harmonic analysis, and the answers are known.

Necessary and sufficient conditions on E for $P_N f \to f$ in $L^2(E\,d\theta)$ are

(a) (Helson and Szego [26]) E has the form $E = e^{g+\tilde{h}}$ where g and h are bounded functions with $\sup |h| < \pi/2$ and $\tilde{h}$ equals the harmonic conjugate of h;

or equivalently

(b) (Hunt, Muckenhaupt and Wheeden [27]) there is a constant C, independent of I such that for every interval I,

$$\frac{1}{|I|}\int_I E(e^{i\theta})\,d\theta \frac{1}{|I|}\int_I \frac{1}{E(e^{i\theta})}\,d\theta \leqq C.$$

Here $|I|$ is the length of I.

This settles approximation theory questions surrounding P_N. The first thing to note is that these conditions are extremely restrictive. They allow E to be singular or to vanish like $E(e^{i\theta}) = \theta^v$ only if $-1 < v < 1$; thus rational E with zeros or poles on the unit circle are eliminated. Traditionally, Cesaro or Abel summation is much more likely to converge than simple Fourier approximation, and our rather negative conclusion suggests that we turn to them as being more practical.

The first thing we mention is a theorem of Rosenblum [22, Thm. 2] which says that Cesaro summation converges on $L^2(E\,d\theta)$ if and only if Abel summation converges on $L^2(E\,d\theta)$. Thus we restrict attention to Abel summation. A necessary and sufficient condition [22, Thm. 1] for $A_r f$ to always converge in $L^2(E\,d\theta)$ is

$$\int_0^{2\pi} P_r(e^{i(\theta-\psi)}) \left|\frac{M(e^{i\theta})}{M(re^{i\theta})}\right|^2 d\theta < K$$

for all $0 < r < 1$ and ψ. Here $P_r(e^{i\theta})$ is the Poisson kernel. Thus a sufficient condition for the Abel approximation to always work is for

$$\sup_{r,\theta} \left|\frac{M(e^{i\theta})}{M(re^{i\theta})}\right| \leqq K^{1/2};$$

that is, M belongs to a class of functions discussed in Chap. 3, § 1.3 of [18]. This class does include the rational functions.

In the multi-input case where E is an operator, similar structure holds for Abel convergence, and we now derive the sufficiency condition, just used, directly.

$$\left\{\int_0^{2\pi} \|y - MA_r u\|^2\right\}^{1/2} \leqq \left\{\int_0^{2\pi} \|M(e^{i\theta})u(e^{i\theta}) - M(re^{i\theta})u(re^{i\theta})\|^2\right\}^{1/2}$$
$$+ \left\{\int_0^{2\pi} \|[M(re^{i\theta}) - M(e^{i\theta})]u(re^{i\theta})\|^2\right\}^{1/2}.$$

The first majorizing term is $\int_0^{2\pi} \|y(e^{i\theta}) - y(re^{i\theta})\|^2$ which goes to zero since $y \in H^2(X)$. The second term is

$$\int_0^{2\pi} \|[I - M(e^{i\theta})M(re^{i\theta})^{-1}]y(re^{i\theta})\|^2$$

which goes to zero for y in $H^2(X)$ if

$$\operatorname*{ess\,sup}_{\theta} \|I - M(e^{i\theta})M(re^{i\theta})^{-1}\|$$

goes to zero or, for y in H^∞, if $\int_0^{2\pi} \|I - M(e^{i\theta})M(re^{i\theta})^{-1}\|^2 \to 0$. Since $M(re^{i\theta}) \to M(e^{i\theta})$ pointwise, the dominated convergence and uniform boundedness theorem imply that the first condition is equivalent to $M(e^{i\theta})M(re^{i\theta})^{-1}$ being uniformly bounded.

Now we turn to a formal question. Recall that the coefficients of the power series expansion for $u(z) = M(z)^{-1}y(z)$ give *formally* our control sequence. The following proposition gives a reasonable condition on the power spectrum E which guarantees that $M(z)^{-1}$ exists for $|z| < 1$ and consequently that this formal sequence exists.

PROPOSITION 2.4. *If M in $H^\infty(\mathcal{U}, \mathcal{H}_1)$ is outer, and if $M(e^{i\theta})^* M(e^{i\theta}) \geqq N(e^{i\theta})^* N(e^{i\theta})$, where N is also outer but with* Range $N(z) = \mathcal{H}_1$ *for some $|z| < 1$, then* Range $M(z) = \mathcal{H}_1$. *In particular, if $M(e^{i\theta})^* M(e^{i\theta}) \geqq \delta(e^{i\theta})I + T(e^{i\theta})$ where $\delta \geqq 0$, $\log \delta(e^{i\theta})$ is integrable and $T(e^{i\theta})$ is a trace class operator with* $\log\det[I + T(e^{i\theta})/\delta(e^{i\theta})]$ *integrable, then* Range $M(z) = \mathcal{H}_1$ *for any $|z| < 1$.*

Proof. The first statement follows immediately from the fact that $M(z)^* M(z) \geqq N(z)^* N(z)$ (see [18, Chap. V, Prop. 4.1]). The log integrability conditions imply that δ and $I + T(e^{i\theta})/\delta(e^{i\theta})$ have outer spectral factorizations $\varphi \in H^\infty(\mathbb{C})$ and φ analytic with a lenient growth condition (see [23, Thm. 3.8]). Since ψ is outer, $\det \psi$ is outer or identically zero. If it is identically zero, then we can write ψ as an infinite matrix with respect to a basis, one subset of which spans cl (Range ψ). The determinant of the minor derived from this basis is outer and so its value at the origin is not zero. Since the pseudoinverse $\psi(z)^{-1}$ can be constructed by Cramer's rule; this says that it is in fact bounded, and consequently Range $\psi(z)$ is closed. The function $N = \psi\varphi$ has closed range and satisfies the majorization hypothesis of the first part of this theorem. Consequently Range $M(z) = \mathcal{H}_1$.

We now give examples to show that Theorem 2.1 is in several senses the best possible. Theorem 2.1 says that J_N has reduced form if and only if E has a spectral factorization. Since E has the special form $R + W^*$ sgn QW, it is conceivable that a weak assumption such as $E \geqq 0$ actually forces E to have a spectral factorization. The following example shows that this is not the case. Take $\mathcal{U}$ to be one-dimensional $R = 1$, sgn $Q = -1$ and set $W^* W = n \leqq 1$. By the realization theorem, any function W in $H(\mathbb{C})$ with $W(0) = 0$ comes from a system and so can arise in this context. By Theorem 18 [11], $\int_{-\pi}^{\pi} \log n(e^{i\theta})\, d\theta > -\infty$ if and only if n has a factorization $n = W^* W$ with W in $H^\infty(\mathbb{C})$. However, if $E = 1 - n \geqq 0$ has a spectral factorization, then $\int_{-\pi}^{\pi} \log(1 - n(e^{i\theta}))\, d\theta > -\infty$, and this is simply not guaranteed by the fact that $\log n$ is integrable.

The second example is of a system for which no exact optimal control law exists. Let $\mathcal{H} = \mathcal{U} = l^2(0, \infty, \mathbb{C})$, $R = 0$ and $Q = 1$. We shall take $W(e^{i\theta})$, and consequently $E(e^{i\theta})$, to be diagonal in the natural basis for $l^2(0, \infty, \mathbb{C})$ and denote the diagonal entries of $E(e^{i\theta})$ by $e_j(e^{i\theta})$. The outer factor M of E is diagonal with entries $m_j(e^{i\theta})$ each of which is an outer factor of e_j, i.e., $\bar{m}_j m_j = e_j$. The function M has the representation $G + zF(I - zA)^{-1}B$, and the system has an exact optimal control law only if Range G contains $F\tilde{\mathcal{R}}$. To see this, suppose that for each x_0' in $\tilde{\mathcal{R}}$ there is an input $v_0, v_1, \cdots$ which gives the optimal performance of the system. Let $u_0, u_1, \cdots, u_n$ be a control which drives the system from 0 to x_0' and set $u = (u_0, u_1, \cdots, u_n, v_0, v_1, \cdots)$. By optimality, $J(0, u)$ is finite. Thus $K(x_N, x_N) \to 0$ and $J(0, u) = \sum_{i=0}^{n} \|Gu_i + Fx_i\|^2 + \sum_{j=0}^{\infty} \|Gv_j + Fx_{j+n+1}\|^2$. However, for each $\varepsilon > 0$ we can find a control sequence so that the resulting cost is within ε of the $\sum^n$ term. Thus the $\sum^\infty$ term is 0, and so we can actually solve $Gv_0 = Fx_{n+1} = Fx_0^*$, that is, $Fx_0' \in$ Range G.

Range G contains $F\tilde{\mathcal{R}}$ if and only if for $\{x_N\}$, any trajectory of the vector $\int_{-\pi}^{\pi} Fx(e^{i\theta})\, e^{-il\theta}\, d\theta$ for each $l = 1, 2, 3, \cdots$ belongs to Range G. This is equivalent to the statement $\int_{-\pi}^{\pi} M(e^{i\theta}) u(e^{i\theta})\, e^{-il\theta}\, d\theta$ belongs to Range $M(0)$ for each $u \in H^2(\mathcal{U})$ and $l > 0$, and this in turn is equivalent to the statement that

$$\text{Range } M_n \subset \text{Range } M_0,$$

where $M(z) = \sum_{k=0}^{\infty} M_k z^k$. The operator M_n is multiplication by the sequence $\{(1/n_n')(d^n/dz) n_j|_{z=0}\}_{j=0}^{\infty}$ on $l^2(0, \infty, \mathbb{C})$. Now Range $M_1 \subset$ Range M_0 if and only if $M_1 = M_0 Y$ for Y a bounded operator (see [5]). Thus $(d/dz) m_j(0)/m_j(0) \equiv \delta_j$ must be a bounded sequence. The functions m_j are outer and consequently can be written

$$m_j(z) = \exp \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{e^{it} + z}{e^{it} - z} \log e_j(e^{it})\, dt.$$

Thus

$$m_j(0) = \exp \frac{1}{4\pi} \int_{-\pi}^{\pi} \log e_j$$

and

$$\frac{d}{dz} m_j(0) = \frac{-1}{2\pi} \int_{-\pi}^{\pi} e^{-it} \log e_j(e^{it})\, dt\, m_j(0).$$

To obtain the example, choose a sequence l_j of functions with $l_j(e^{it}) \leqq 0$ with $\int_{-\pi}^{\pi} l_j > -\infty$ and $\int_{-\pi}^{\pi} e^{-it} l_j(e^{it})\, dt \to -\infty$. Set $e_j = \exp l_j$, let w_j be a spectral factorization of e_j which vanishes at $z = 0$, and use the realizability theorem to determine a system which gives rise to W. By construction $\delta_j \to \infty$ and so Range $G \not\supset F\tilde{\mathcal{R}}$.

Remark 2.1. The case where J can be reduced with $K(x_N, x_N) \not\to 0$ is analyzed in § 4.

Remark 2.2. If Q and R are both nonnegative, then (1.2) and (2.1) together imply that $K(\cdot, \cdot)$ is a nonnegative bilinear form.

Remark 2.3. The feedback law we have obtained can frequently be expressed in terms of the optimal cost functional K. If Range $G = \mathcal{H}_1$ and G^{-1} denotes the standard pseudoinverse of G, then formally

$$\begin{aligned} u_i &= -G^{-1}Fx_i = (G^*G)^{-1}G^*Fx_i, \\ u_i &= (R+B^*KB)^{-1}B^*KAx_i. \end{aligned} \tag{2.8}$$

This expression has a reasonable interpretation even when K is a bilinear functional. As a further aside, we note that the feedback law can be expressed in terms of L. Namely, if $x_i = \sum_{j=0}^{i} A^j B_j v_j$, then

$$u_i = \frac{1}{2\pi}\left[R + \frac{1}{2\pi}\int_{-\pi}^{\pi} L(e^{i\theta})\,d\theta\right]^{-1} 1 \sum_{j=0}^{N} \int_{-\pi}^{\pi} L(e^{i\theta}) v_j\, e^{-i(j+1)\theta}\,d\theta.$$

3. Stability of feedback systems. In this section, we give some stability theorems which are suitable for analyzing the behavior of the control systems found in § 1. We shall not belabor this, since our results are near to existing results (see [4], [3]). Consider the system

$$x_{i+1} = Ax_i + Bu_i, \qquad y_i = Cx_i, \tag{3.1}$$

with feedback law $u_i = \Omega y_i$. The frequency response function is $R(z) = zC(I - zA)^{-1}B$. Define a function $\mathcal{J}(z) = I - \Omega R(z)$ and note that if M and $R(z)$ are scalars, then the classical Nyquist stability criterion (which we shall presently extend) is expressed in terms of the set $\{\mathcal{J}(e^{i\theta}) : \text{all } \theta\}$. Set $C = I$ and note that the formal feedback law $\Omega = -G^{-1}F$, obtained in § 2 from the spectral factor M, satisfies

$$G\mathcal{J}(z) = M(z). \tag{3.2}$$

If Range $G = \mathcal{H}_1$, then $\mathcal{J}$ is outer. Let G^{-1} denote the standard pseudoinverse for G.

The crux of this business is an easily verified identity

$$(I - z[A - B\Omega C])^{-1}B\mathcal{J}(z) = (I - zA)^{-1}B. \tag{3.3}$$

If u is an admissible input in $l^2[0, \infty, \mathcal{U}]$ with Fourier transform u in $H^2(\mathcal{U})$, then the Fourier transform of the trajectory associated with u is $x(z) = z(I - zA)^{-1}Bu(z)$. Thus $\mathcal{J}$ relates trajectories of the original system to those of the feedback system.

THEOREM 3.1. *Suppose that the power density function E for the system $[A, B]$ with absolutely finite J_N satisfies $E(e^{i\theta}) \geqq \delta I > 0$. Then the outer factorization M of E gives rise to an optimal feedback law Γ via § 2 with the property that the feedback system $[A, A + B\Gamma]$ has the same trajectories as the original system. Furthermore, the ranges of the controllability operators for the two systems are equal.*

Thus if all trajectories of the original system tend to zero, then all trajectories of the controlled system tend to zero. This will also guarantee a weak form of asymptotic stability; namely, if x is a state of the feedback system which is reachable in a finite amount of time, then $(A + B\Gamma)^n x \to 0$ as $n \to \infty$.

The first part of this theorem is an immediate consequence of the fact that G is invertible when $E \geqq \delta I > 0$ and of the following

PROPOSITION 3.2. *The function $\mathcal{J}(e^{i\theta})$ (resp., $\mathcal{J}(e^{i\theta})^{-1}$) is in $H^\infty(\mathcal{U}, \mathcal{U})$ if and only if the trajectories of $[A+B\Omega C, B]$ are contained in (resp., contain) the trajectories of $[A, B]$.*

Proof. One side is obvious. To do the other side, suppose that the trajectories of the feedback system contain those of the original system. If u is a l^2-input sequence, let $x(u)$ be the corresponding trajectory of the original system. By assumption, there is an input Lu to the feedback system with trajectory $x(u)$, and clearly this determines Lu uniquely. So L is a map of $l^2(0, \infty, \mathcal{U})$ into itself. It is trivial to check that the graph of L is closed. Consequently, L is a bounded operator. However, (3.3) implies that L is just the operator "multiplication by $\mathcal{J}(e^{i\theta})$", and so $\mathcal{J}$ is in $H^\infty(\mathcal{U})$. The same type of argument applies to $\mathcal{J}^{-1}$.

Next we look at (3.3) in terms of controllability operators. Let $\mathcal{C}$ and $\mathcal{C}_f$ denote the controllability operators for the original and the feedback system. Let $P_{\bar{H}^2}$ and P_{H^2} denote the orthogonal projection of L^2 onto H^2 and $\bar{H}^2$. If $L \in L^\infty$, we define $\mathcal{T}_L : H^2 \to H^2$ by

$$\mathcal{T}_L f = P_{H^2} M_L f.$$

It is called the Toeplitz operator with generating function L. The best reference for scalar Toeplitz operators is [6]; for Hilbert Toeplitz operators see [19]. Let T_L denote the operator induced on l^2 by Fourier transforming $\mathcal{T}_L$ on H^2. Let $\mathcal{J}^+(e^{i\theta}) = \mathcal{J}(e^{-i\theta})$. If $\mathcal{J} \in H^\infty$, then $\mathcal{J}^+ \in \bar{H}^\infty$. The second part of Theorem 3.1 follows from

PROPOSITION 3.3. *If $\mathcal{J}(z)$ is in $H^\infty(\mathcal{U}, \mathcal{U})$, then*

$$\mathcal{C}_f T_{\mathcal{J}^+} = \mathcal{C}$$

If $\mathcal{J}^{-1}(z)$ is in $H^\infty(\mathcal{U}, \mathcal{U})$, then

$$\mathcal{O}_f = \mathcal{C} T_{(\mathcal{J}^+)^{-1}}.$$

The operator $T_{\mathcal{J}^+}$ is invertible if both $\mathcal{J}^+$ and $(\mathcal{J}^+)^{-1}$ are in $H^\infty(\mathcal{U}, \mathcal{U})$.

Proof. We do the second relationship first. Observe that

$$\mathcal{C}\{u_j\}_{j=0}^\infty = \sum_{j=0}^\infty A^j B u_j = \lim_{r \uparrow 1} \frac{1}{2\pi} \int_{-\pi}^{\pi} (I - re^{-i\theta}A)^{-1} B u(e^{i\theta})\, d\theta.$$

Equation (3.2) implies

$$\mathcal{C}_f\{u_j\}_0^\infty = \lim_{r \uparrow 1} \frac{1}{2\pi} \int_{-\pi}^{\pi} (I - re^{-i\theta}A)^{-1} B \mathcal{J}(re^{-i\theta})^{-1} u(e^{i\theta})\, d\theta.$$

Since $\mathcal{J}(e^{-i\theta})u(e^{i\theta})$ is in $L^2(\mathcal{U}, \mathcal{U})$, it can be written as the sum of its projection V onto $\bar{H}^2$ and its projection, $T_{\mathcal{J}^+}u$ on H^2. Since $(I - \bar{z}A)^{-1}Bv(z)$ is in $\bar{H}^2$, its integral over $\mathbf{T}$ is zero. This gives the desired result. The first part of the theorem follows similarly. The last statement in the theorem is a standard fact about Toeplitz operators.

Now that Theorem 3.1 is proved, we make a few remarks. The identity

$$\mathcal{J}_0(z)C(I - z[A - B\Omega C])^{-1} = C(I - zA)^{-1}, \tag{3.4}$$

where $\mathcal{J}_0(z) = I - R(z)\Omega$ is the analogue to (3.3) which allows one to connect all statements about trajectories and controllability in Propositions 3.1 and 3.2 to statements about observability. Another remark is that one can lift the hypothesis $E \geqq \delta I > 0$ and get a palatable theorem. Namely, if E has a spectral factorization and the feedback law Γ comes from an outer factor M, then there is a set of inputs u to the feedback system dense in $l^2(0, \infty, \mathcal{U})$ whose trajectories are precisely the trajectories of the original system. This is obtained by strengthening Proposition 3.1 in the obvious way.

Also note that continuous exact controllability and observability imply stability (see [8, Appendix] or [12, § 4, Remark]). The structure is

PROPOSITION 3.4. *If a system $[F, \varphi]$ is continuously exactly observable, then φ is asymptotically stable. If a system $[\varphi, D]$ is continuously exactly controllable, then φ^* is asymptotically stable.*

Remark. Suppose that $[A, B]$ is a finite-dimensional controllable system and that the eigenvalues of A lie inside $|z| < 1$. Then Theorem 3.1 can be strengthened because of these additional assumptions. One obtains that the state operator $A + B\Gamma$ for the feedback system has no eigenvalues on $|z| = 1$ if and only if $E(e^{i\theta}) \geqq \delta I > 0$. The eigenvalues of $A + B\Gamma$ always lie in $|z| \leqq 1$.

The last statement follows trivially from (3.2) since $M(z)^{-1}$ exists for $|z| < 1$. The absence of eigenvalues on $|z| = 1$ follows from Theorem 3.1. Conversely, if E has a zero on $|z| = 1$, then M^{-1} has a pole there. Since $(I - zA)^{-1}Bv_0 \neq 0$ for any v_0 or $|z| \leqq 1$, equation (3.3) implies that $(I - z[A + B\Gamma])^{-1} B$ has a pole on the circle, and so $A + B\Gamma$ has an eigenvalue there.

4. The algebraic Riccati equation. With the control problem we have studied (when R is invertible), one associates the formal linear fractional map

$$\mathcal{F}(P) = A^*P(I + CP)^{-1}A + Q, \tag{4.1}$$

where $C = BR^{-1}B^*$ and expects that the optimal cost "operator" K will be a fixed point $\mathcal{F}(K) = K$ of it. In this section, we give a fairly thorough study of when a fixed point exists. Although everything done is intimately linked with the original control problem, we try to present the forthcoming results as a study of the fixed-point problem for its own sake.

Throughout this section, we shall work with a slightly more general class of $\mathcal{F}$ than those given by (4.1). Any self-adjoint operator C can be written in the form $C = B^*R^{-1}B$, where R is an invertible self-adjoint operator. Provided that the appropriate inverses exist, a simple manipulation converts (4.1) to

$$\mathcal{F}(P) = A^*PA - A^*PB[R + B^*PB]^{-1}B^*PA + Q. \tag{4.2}$$

This formula is more symmetric than (4.1) and consequently easier to use. Also R need not be invertible in (4.2), so it is more general than (4.1). Henceforth, we work with $\mathcal{F}$ of (4.2). The self-adjoint operator $R + B^*PB$ plays an important role in the study of $\mathcal{F}$; we denote it by Λ_P and call it the *indicator* of P. When, for example, $\mathcal{H}$ is finite-dimensional, the natural domain of definition for $\mathcal{F}$ is precisely the set $\mathcal{P}_0$ of those matrices P satisfying Range $\Lambda_P \supset$ Range B^*PA since these are the matrices for which the second term of (4.2) is well-defined.

Not too surprisingly, spectral factorizations play as big a role in this section as they have previously. In fact, we shall require a type of signed factorization. A *signature operator* J is a self-adjoint operator with the property that $J^2 = I$. We say that the self-adjoint $(\mathcal{T}, \mathcal{H}, \mathcal{H})$-valued function E has a (outer) *signed spectral factorization* if and only if there is a signature operator $\mathcal{S}$ on a Hilbert space $\mathfrak{H}_1$ such that for each u in $l^2(\mathcal{U})$, the limit as $r\uparrow 1$ of $\int_0^{2\pi}(u(e^{i\theta}), M^*(re^{i\theta})M(e^{i\theta})u(e^{i\theta}))\,d\theta$ exists and is $\int_0^{2\pi}(u(e^{i\theta}), E(e^{i\theta})u(e^{i\theta}))\,d\theta$; here M is a (bounded outer) $\mathcal{L}(\mathcal{H}, \mathcal{H}_1)$-valued function analytic in the unit disk. The question of which functions have such factorizations was studied by Symeninco (cf. [10]), and he obtained that in many situations E has a signed spectral factorization if and only if $E = AB$ for some outer functions A in $\bar{H}^\infty$ and B in H^∞. This is consistent with the fact privately observed by A. Devinetz and R. G. Douglas that a uniformly invertible E has a signed spectral factorization if and only if the Toeplitz operator generated by E is invertible. Neither of these conditions are practical to apply, and it is fortunate for control theory purposes that only positive factorizations are interesting. Although the main theorems of this section concern the infinite-dimensional situation, the following corollary (of Theorem 4.7) is new in finite dimensions and describes the behavior there.

THEOREM 4.1. *Suppose that A, B, R, Q are finite-dimensional matrices with R, Q self-adjoint and all eigenvalues of A less than 1. Then the map $\mathcal{F}$ has a self-adjoint fixed point K in $\mathcal{P}_0$ with nonnegative indicator if and only if the function*

$$E(e^{i\theta}) = R + B^*(I - e^{i\theta}A)^{*-1}Q(I - e^{i\theta}A)^{-1}B$$

is nonnegative. The map $\mathcal{F}$ has a fixed point in $\mathcal{P}_0$ (if) and only if E has an (outer) signed spectral factorization.

By (2.5a) the optimal cost functionals from § 2 have positive indicator. These are the important ones and the author suspects without an improved theory of signed factorizations that the first part of Theorem 4.1 is the only part of real interest. It is analogous to the condition of Willems [25] for the continuous-time Riccati equation although here no controllability assumption is required.

4.1. Decomposition of a map into linear and quadratic parts. The fixed-point problem for $\mathcal{F}$ is, to a superficial glance, a quadratic problem, but it can also contain affine linear fixed-point problems of the form $K = NKD + Q$, one example being when $B = 0$. These problems have been studied [21] and can be treated by quite a different approach than a purely quadratic problem. Fortunately, the fixed-point problem decomposes neatly into what we may think of as purely quadratic and purely linear parts. This we now demonstrate.

Let $\mathcal{H}$ and $\mathcal{U}$ be Hilbert spaces and suppose that A, Q acting on $\mathcal{H}$, R acting on $\mathcal{U}$, and $B : \mathcal{H} \to \mathcal{H}$ are bounded operators with R and Q self-adjoint. Let $l_F^2(0, \infty, \mathcal{H})$ denote the $\mathcal{H}$-valued sequences of finite length and define $\mathcal{C} : l_F^2 \to \mathcal{H}$ by $\mathcal{C}\{x_j\} = \sum_{j=0}^{\infty} A^j B x_j$. Set $\tilde{\mathcal{R}} = \text{Range } \mathcal{C}$; denote its closure by $\mathcal{R}_1$ and its orthogonal complement by $\mathcal{R}_2$. If P is any self-adjoint operator on $\mathcal{H}$, then in the $\mathcal{R}_1, \mathcal{R}_2$ basis it can be written as a matrix $\begin{bmatrix} P_1 & P_2 \\ P_2^* & P_3 \end{bmatrix}$ with P_1 and P_3 self-adjoint. We would like to see how $\mathcal{F}$ acts on such 2×2 matrices. If Γ_1 and Γ_2 are the orthogonal projections of $\mathfrak{H}$ onto $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively, then $A\Gamma_1 = \Gamma_1 A\Gamma_1$ and

$A^*\Gamma_2 = \Gamma_2 A^*\Gamma_2$. We write $A = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix}$ and $Q = \begin{bmatrix} Q_1 & Q_2 \\ Q_2^* & Q_3 \end{bmatrix}$ and begin computing $\Gamma_i \mathscr{F}(P)\Gamma_j$:

$$\Gamma_1 \mathscr{F}(P)\Gamma_1 = A_1^* P_1 A_1 - A_1^* P_1 B(R + B^* P_1 B)^{-1} B^* P_1 A_1 + Q_1; \tag{4.3}$$

that is, $\mathscr{F}_1(P_1) = \Gamma_1 \mathscr{F}(P)\Gamma_1$, where $\mathscr{F}_1$ is the linear fractional map defined by (4.3),

$$\begin{aligned} \Gamma_1 \mathscr{F}(P)\Gamma_2 = {} & [A_1^* - A_1^* P_1 B(R + B^* P_1 B)^{-1} B^*] P_2 A_3 \\ & + Q_2 + [A_1^* - A_1^* P_1 B(R + B^* P_1 B)^{-1} B^*] P_1 A_2, \end{aligned} \tag{4.4}$$

that is, $\Gamma_1 \mathscr{F}(P)\Gamma_2$ is an affine linear function of P_2,

$$\begin{aligned} \Gamma_2 \mathscr{F}(P)\Gamma_2 = {} & A_3^* P_3 A_3 + [A_2^* - A_2^* P_1 B(R + B^* P_1 B)^{-1} B^*] \\ & \cdot [P_1 + P_2] A \Gamma_2 \\ & + A_3^* P_2^* [B(R + B^* P_1 B)^{-1} B^* P_1 A_2 - A_2] \\ & - A_3^* P_2^* B[R + B^* P_1 B]^{-1} B^* P_2 A_3, \end{aligned} \tag{4.5}$$

which is an affine linear function of P_3. Thus we see that the only truly quadratic part of the fixed-point problem $\mathscr{F}(P) = P$ is the equation $\mathscr{F}_1(P_1) = P_1$. Also this is the only part of the problem which is interesting from the control theory point of view. We shall call the map $\mathscr{F}$ of (4.2) *purely quadratic* if and only if $\tilde{\mathscr{R}}$ is dense in $\mathscr{H}$. Such maps will take our main attention, and treatment of the linear maps is postponed to the end of this section.

4.2. Purely quadratic maps. Throughout this section, we assume that $\mathscr{F}$ is purely quadratic. The map $\mathscr{F}$ is clearly defined on all bounded operators with invertible indicator. It also will extend continuously to many unbounded operators, and so it is not clear offhand just what should be the natural domain of definition. However, the control problem strongly suggests that the natural space on which $\mathscr{F}$ should act is the space of all possible cost functionals. We formalize this: Let $\mathscr{P}$ denote the space of all symmetric bilinear forms P on $\tilde{\mathscr{R}}$ with the property that

$$P(\mathscr{C}\{u_j\}_{j=0}^N, v) = P(v, \mathscr{C}\{u_j\}_{j=0}^N) \tag{4.6}$$

is for fixed N continuous in u and v belonging to $l^2(\mathscr{U})$. If $P \in \mathscr{P}$, then bilinear form $\Lambda_P(x, y) = (x, Ry) + P(Bx, By)$ for x, y in $\mathscr{U}$ is actually continuous on $\mathscr{U}$, and so by the Riesz representation theorem, there is a bounded operator Λ_P such that $\Lambda_P(x, y) = (x, \Lambda_P y)$. Naturally, Λ_P will be called the *indicator of the bilinear form P*. Given P in $\mathscr{P}$, the bilinear form $P(Bx, y)$ for y in $\tilde{\mathscr{R}}$ is continuous in x, and, consequently, there is an operator E_P defined on $\mathscr{R}$ so that $(x, E_P y) = P(Bx, y)$. We want to have $\mathscr{F}$ defined on as big a subset of $\mathscr{P}$ as is reasonably possible. With this in mind define

$$\begin{aligned} \mathscr{P}_0 = \{P : {} & \text{there exists a decomposition } \Lambda_P = NSN \text{ with } S \text{ a signature} \\ & \text{operator and } N \text{ a nonnegative self-adjoint operator satisfying} \\ & \text{Range } N \supset \text{Range } E_P A\}. \end{aligned}$$

The map $\mathscr{F}$ is defined on $\mathscr{P}_0$ by

$$\mathscr{F}(P)(x, y) = P(Ax, Ay) - (N^{-1} E_P Ax, SN^{-1} E_P Ay) + (x, Qy) \tag{4.7}$$

for x, y in $\tilde{\mathcal{R}}$. Here N^{-1} denotes the standard pseudoinverse of N. This is clearly consistent with (4.2) when P is a bounded operator, and it is straightforward to check that the definition of $\mathcal{F}$ depends only on $\Lambda_P{}^1$ and, consequently, is independent of which factorization NSN is used.

All results on fixed points will be given in terms of the function $W(z) = z|Q|^{1/2}(I - zA)^{-1}B$, which we henceforth assume to be in $H^\infty(\mathcal{H}, \mathcal{U})$, and, in particular, they will involve

$$E(e^{i\theta}) = R + W(e^{i\theta})^* \operatorname{sgn} QW(e^{i\theta}). \tag{4.8}$$

An operator A will be called *asymptotically stable* if $A^n x \to 0$ for each x. If B is an operator with one-dimensional range, then E is a real-valued function on the circle and we shall prove

THEOREM 4.2. *Consider a purely quadratic map $\mathcal{F}$ as in* (4.2) *with B a rank one operator and A asymptotically stable.*

(i) *If $\mathcal{H}$ is finite-dimensional, then $\mathcal{F}$ has a fixed-point in $\mathcal{P}_0$ if and only if E has one sign. The indicator for the fixed point has the same sign as E.*

(ii) *If $\mathcal{H}$ is not finite-dimensional, then $\mathcal{F}$ has many fixed points in $\mathcal{P}_0$. Some fixed points will have positive and some will have negative indicators.*

This theorem sets down the basic behavior of the purely quadratic fixed-point problem. The problem of higher-dimensional B is simply a mixture of these cases. In Theorem 4.6, we sort out this mixture to a large extent, and Theorem 4.2 will be an easy consequence of it.

It turns out that fixed points of $\mathcal{F}$ in $\mathcal{P}_0$ fall into two categories, those for which $P(x_N, y_N) \to 0$ along trajectories x_N, y_N of the system $[A, B]$, called *standard points*, and those which are not standard. Clearly, $P(A^n x, A^n y) \to 0$ for $x, y \in \tilde{\mathcal{R}}$ if P is standard, and one can show that up to terrible pathologies, fixed points for $\mathcal{F}$ of this type are standard. If P has positive indicator, this is always equivalent to being standard. Standard fixed points are the only ones of obvious control theoretic interest, and they correspond to signed spectral factorizations as the following theorem states.

THEOREM 4.3. *The map $\mathcal{F}$ has a standard fixed point (if and) only if E has a (outer) signed spectral factorization.*

In finite dimensions for asymptotically stable A, all points are standard, and so this theorem describes that situation completely. Before stating our most complete theorem on fixed points, we give the proof of this theorem since it is instructive.

Proof. We begin with the observation that the bilinear functional $\langle \cdot, \cdot \rangle$ defined on the space $\mathcal{H}_1$ which was used throughout § 2 (see (2.1) and (2.5), in particular) need not be nonnegative. In fact, had we assumed that $\mathcal{H}_1$ has inner product $[\cdot, \cdot]$ and that $\langle x, y \rangle = [\mathcal{S}, x, y]$ for some signature operator $\mathcal{S}$; then the proofs in § 2 would have gone through with the modification that $E = M^* \mathcal{S} M$. Then Lemma 2.3 and Theorem 2.1 combine to give

PROPOSITION 4.4. *There exist G, F and K satisfying* (2.5) *with signed $\langle \cdot, \cdot \rangle$ and having $K(A^N x, A^N y) \to 0$ if E has an outer signed factorization. Conversely, if such G, F, K exist, then E has a signed factorization. In the above statement, K satisfies $\Lambda_K \geqq 0$ if and only if "signed" is removed from the statements about factorizations.*

One thing which requires clarification is that the existence of G, F, K implies a signed factorization for E. The function $M(z) = G + zF(I - zA)^{-1}B$ is analytic inside the disk. If $u \in l^2(\mathcal{U})$, then set $y(re^{i\theta}) = M(re^{i\theta})u(re^{i\theta})$. We wish to show that $\int_0^{2\pi} (\mathcal{S}y(re^{i\theta}), y(re^{i\theta}))\, d\theta$ converges to $\int_0^{2\pi} (u(e^{i\theta}), E(e^{i\theta})u(e^{i\theta}))\, d\theta$. If the power series for $y(z)$ is $\sum_{n=0}^{\infty} y_n z^n$, then the first integral is $\sum_{n=0}^{\infty} r^{2n}(\mathcal{S}y_n, y_n)$. Finiteness of the second integral forces the sequence $(x_i, Qx_i) + (u_i Ru_i)$ to be summable, and since $K(x_N, x_N) \to 0$, this implies that $(\mathcal{S}Gu_i + Fx_i, Gu_i + Fx_i)$ is summable. However, $y_i = Gu_i + Fx_i$ and so $(\mathcal{S}y_i, y_i)$ is summable; its sum is $\int_0^{2\pi} (u(e^{i\theta}), E(e^{i\theta})u(e^{i\theta}))\, d\theta$. An Able summation argument (cf. [24, § 1.22]) gives $\sum_{n=0}^{\infty} r^{2n}(\mathcal{S}y_n, y_n) \to \sum_{n=0}^{\infty} (\mathcal{S}y_n, y_n)$ as $r \uparrow 1$.

The next thing to prove is that families of three objects G, F, K satisfying (2.5) correspond precisely to fixed points of $\mathcal{F}$.

PROPOSITION 4.5. *The bilinear functional K in $\mathcal{P}_0$ is a fixed point of $\mathcal{F}$ if and only if there exist G and F and possibly signed $\langle \cdot, \cdot \rangle$ so that (2.5) holds.*

Proof. Suppose that (2.5) holds. From (2.5a) you see $G^*\mathcal{S}G = \Lambda_K$. By (2.5b) the operator $E_K A$ is $G^*\mathcal{S}F$. If $G = UN$ denotes the polar decomposition of G, then since Range G is dense, U^* is an isometry with Range $U^* \subset$ cl Range N and $u^*\mathcal{S}U - (I - U^*U) = \mathcal{S}'$ is a signature operator. Now $\Lambda_K = N\mathcal{S}'N$, but $E_K A = NU^*\mathcal{S}F$, and so Λ_K is in $\mathcal{P}_0$ and we have

$$\begin{aligned}\mathcal{F}(K)(x, y) &= (N^{-1}NU^*\mathcal{S}Fx, \mathcal{S}N^{-1}NU^*\mathcal{S}Fy) + K(Ax, Ay) + (x, Qy)\\ &= (Fx, Fy) + K(Ax, Ay) + (x, Qy).\end{aligned}$$

By (2.5c) this is just $K(x, y)$.

If K in $\mathcal{P}_0$ is a fixed point of $\mathcal{F}$, then $\Lambda_K = N\mathcal{S}N$. Set $G = N$ and take N^{-1} to be the standard pseudoinverse of N. The bilinear functional $K(BN^{-1}\mathcal{S}x, Ay)$ for fixed y in $\tilde{\mathcal{R}}$ and a dense space of x's equals $(N^{-1}\mathcal{S}x, E_K Ay) = (\mathcal{S}x, N^{-1}E_K Ay)$ and so is continuous in x. Thus there is an operator F on $\mathcal{R}$ for which this equals (x, Fy). One can reverse the brief computations above and get that $G, F, \mathcal{S}$ and K satisfy (2.5).

The general situation is described by

THEOREM 4.6. *Suppose P in $\mathcal{P}_0$ is a fixed point of $\mathcal{F}$ for which $\lim_{N\to 0} P(x_N, x_N)$ exists for each trajectory of the system $[A, B]$.* (Note that for any fixed point with positive indicator this limit either exists or is infinite.) *Then there is a bilinear form $\lambda(,)$ defined on $\tilde{\mathcal{R}}$ so that $\lim_{N\to\infty} P(A^N x_1, A^N x_2) = \lambda(x_1, x_2)$, and there is a self-adjoint function $\lambda \in L^\infty(\mathcal{U}, \mathcal{U})$ so that*

$$\lambda(\mathcal{C}\{u_j\}, \mathcal{C}\{v_j\}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} (u(e^{i\theta}), \lambda(e^{i\theta})v(e^{i\theta}))_{\mathcal{U}}\, d\theta. \tag{4.9}$$

The function λ has the properties

(4.10a) *if both $u(e^{i\theta})$ and $(I - e^{i\theta}A)^{-1}Bu(e^{i\theta})$ are vector-valued polynomials in $e^{i\theta}$, then $\lambda(e^{i\theta})u(e^{i\theta}) = 0$;*

(4.10b) *$E + \lambda$ has a signed spectral factorization;*

(4.10c) *if $u \in l^2_F(0, \infty)$ and $\mathcal{C}u = 0$, then $\lambda(e^{-i\theta})u(e^{i\theta}) \in H^2(\mathcal{U})$.*

Conversely, if λ *is an* $L^\infty(\mathcal{U}, \mathcal{U})$ *function for which*

(4.11a) *property* (4.10a) *holds,*

(4.11b) $E+\lambda$ *has an outer signed factorization,*

(4.11c) *if* $u(e^{i\theta}) \in H^2(\mathcal{U})$ *satisfies* $W(e^{-i\theta})u(e^{i\theta}) \in H^2(\mathcal{H})$, *then* $\lambda(e^{-i\theta})u(e^{i\theta}) \in H^2(\mathcal{U})$,

then $\mathcal{F}$ *has a fixed point* P *in* $\mathcal{P}_0$ *which satisfies*

$$P(A^N x_0, A^N y_0) \to \lambda(x_0, y_0),$$

where λ *is given by* (4.9) *and* $x_0, y_0 \in \tilde{\mathcal{R}}$.

Proof. Suppose $\lambda \in L^\infty(\mathcal{U}, \mathcal{U})$ is a function which satisfies (4.11). We shall give a construction for associating a fixed point of $\mathcal{F}$ with λ. The set of $u \in H^2$ such that $W(e^{-i\theta})u(e^{i\theta}) \in H^2$ is invariant under multiplication by $e^{i\theta}$, and so by the Lax–Beurling theorem, there is a $\varphi \in H^\infty(\mathcal{U}_1, \mathcal{U})$ for which $W(e^{-i\theta})\varphi(e^{i\theta}) \in H^\infty(\mathcal{U}_1, \mathcal{U})$ and $\varphi(e^{i\theta})^* W(e^{-i\theta})^* \in \bar{H}^\infty(\mathcal{H}, \mathcal{U}_1)$ is outer. Since (4.11c) is equivalent to $\varphi(e^{i\theta})^* \lambda(e^{-i\theta}) \in \bar{H}^\infty(\mathcal{U}, \mathcal{U}_1)$, we have $\operatorname{cl} P_{H^2(\mathcal{U})}(\mathcal{M}_{\lambda(e^{-i\theta})}\bar{H}^2(\mathcal{U})) \subset \operatorname{cl} P_{H^2(\mathcal{U})}(\mathcal{M}_{\tilde{W}}\bar{H}^2(\mathcal{H}))$. This along with (4.11b) is the crucial fact in the proof of Theorem 2.1 which yields that there are operators G and F so that the signed outer factor M of $E+\lambda$ has the representation $M(z) = G + zF(I - zA)^{-1}B$. Now we can follow the construction in Theorem 2.1 to obtain a bilinear functional K so that G, F, K reduces J_N. Consequently, K is a fixed point of $\mathcal{F}$.

To see this, we began by associating a bilinear functional $\lambda(\cdot, \cdot)$ on $\tilde{\mathcal{R}}$ with the function $\lambda(e^{i\theta})$ by equation (4.9). To see that this is well-defined we only need $\lambda(x_0, x_0) = 0$ whenever x_0 or $y_0 = 0$. That is, if $x_l = \int_{-\pi}^{\pi} (I - e^{i\theta}A)^{-1} B u(e^{i\theta}) e^{-il\theta}\, d\theta = 0$ for $u(e^{i\theta})$ some polynomial of order $\leqq l$ in $e^{i\theta}$, then $\int_{-\pi}^{\pi} (V(e^{i\theta}), \lambda(e^{i\theta})u(e^{i\theta}))\, d\theta = 0$. This is equivalent to (4.11a). It is immediate from the definition that $\lambda(Ax, Ay) = \lambda(x, y)$ for $x, t \in \tilde{\mathcal{R}}$. Formally, if we set $K_1(x, y) = \sum_{n=0}^{\infty} (x, A^{*n}(Q - F^*F)A^n y) + \lambda(x, y)$, then the fact that K_1, etc., satisfies (2.5c) is a straightforward consequence of $E + \lambda = M^* \mathcal{S} M$. It is, however, unclear that such a K_1 can be rigorously defined. To see that K actually does exist define L by (2.6) and use $L + \lambda$ in (2.7) to define a function. With a bit of work one can check that this function actually has the properties (2.5) required of K.

Now we do the converse direction. Suppose K is a fixed point of $\mathcal{F}$. Associated with K we have operators $G, F, \mathcal{S}$ as in (2.5) and the function $M(z)$. From (2.2) we see that if $\Lambda_P \geqq 0$,

$$\lim_{N\to\infty} K(x_N, x_N) = -\frac{1}{2\pi} \int_{-p}^{\pi} (u(e^{i\theta}), E(e^{i\theta})u(e^{i\theta}))\, d\theta + \lim_{N\to\infty} \sum_{j=0}^{N-1} \|Gu_i + Fx_i\|^2,$$

where $\{x_N\}$ is the $[A, B]$ trajectory corresponding to input u, and consequently $\lim_{N\to\infty} K(x_N, x_N)$ exists or is $+\infty$, In general,

$$\lim_{N\to\infty} K(x_N, x_N) = \frac{1}{2\pi}\int_{-\pi}^{\pi} (u(e^{i\theta}), [-E(e^{i\theta}) + M(e^{i\theta})^* \mathcal{S}M(e^{i\theta})]u(e^{i\theta}))\, d\theta$$

is finite for all trajectories if and only if the function $\lambda = M^*\mathcal{S}M - E$ is in $L^\infty(\mathcal{U}, \mathcal{U})$, thus establishing (4.10b). We also get $\lim_{N\to\infty} K(x_N, y_N)$ exists. From (2.1) one can also see that

$$\sum_{n=0}^{\infty} (A^n x_0, QA^n y_0) = \sum_{n=0}^{\infty} (FA^n x_0, FA^n y_0) + K(x_0, y_0) - \lim_{N\to\infty} K(A^N x_0, A^N y_0),$$

and so the last limit exists and serves to define $\lambda(\cdot, \cdot)$ *on* $\tilde{\mathcal{R}}$. Now $\lambda(\cdot, \cdot)$ will clearly be given by (4.9), and the fact that $\lambda(0, x_0) = \lambda(y_0, 0) = 0$ for x_0 in $\tilde{\mathcal{R}}$ is equivalent to (4.10a).

Finally we verify that λ satisfies (4.10c) and in the process show that (4.10c) and (4.11c) are closely related. Suppose that $u \in l_F^2(0, \infty)$ and $\mathcal{C}u = 0$. Then

$$\mathcal{C}u = \frac{1}{2\pi}\lim_{r\uparrow 1}\int_{-\pi}^{\pi} (I - re^{-i\theta}A)^{-1} Bu(e^{i\theta})\, d\theta = 0$$

and

$$A\mathcal{C}u = \frac{1}{2\pi}\lim_{r\uparrow 1}\int_{-\pi}^{\pi} e^{i\theta}[(I - re^{-i\theta}A)^{-1} - I]Bu(e^{i\theta})\, d\theta = 0.$$

Since $\int_{-\pi}^{\pi} e^{i\theta} Bu(e^{i\theta})\, d\theta = 0$, one has $\lim_{r\uparrow 1}\int_{-\pi}^{\pi} e^{i\theta}(I - re^{-i\theta}A)^{-1} Bu(e^{i\theta})\, d\theta = 0$, and by a similar manipulation of $A^n\mathcal{C}u = 0$, one obtains

$$\lim_{r\uparrow 1}\int_{-\pi}^{\pi} e^{in\theta}(I - re^{-i\theta}A)^{-1} Bu(e^{i\theta})\, d\theta = 0$$

for $n = 0, 1, 2, \cdots$. Thus if N is any operator defined on $\tilde{\mathcal{R}}$ for which $N(I - e^{-i\theta}A)^{-1}B \in \tilde{H}^\infty$, then $N(I - e^{-i\theta}A)^{-1}Bu(e^{i\theta})$ is in H^2 and has the 0th Fourier coefficient equal to zero. In particular, by taking $N = |Q|^{1/2}$ or $N = F$ we get that $W(e^{-i\theta})u(e^{i\theta})$ or $M(e^{-i\theta})u(e^{i\theta})$ is in $H^2(\mathcal{U})$. Since $\lambda(e^{-i\theta}) = M(e^{-i\theta})^*\mathcal{S}M(e^{-i\theta}) - W(e^{-i\theta})^*$ sgn $QW(e^{-i\theta}) - R$, the function $\lambda(e^{-i\theta})u(e^{i\theta})$ is in $H^2(\mathcal{U})$; thus (4.10c) holds. Also one sees that statement (4.10c) implies statement (4.11c) for $u \in l_F^2(0, \infty)$. The converse will be true under strong observability assumptions on the system $[|Q|^{1/2}, A]$ provided that $\mathcal{C}$ acts continuously on $l^2(0, \infty)$.

Proof of Theorem 4.2. If $\mathcal{H}$ is finite-dimensional, then E is rational and so has a signed spectral factorization if and only if E has one sign. Thus Theorem 4.3 applies. The statement about the indicator is clear from the proof of Theorem 4.3. If $\mathcal{F}$ has a fixed point P in $\mathcal{P}_0$, then since $\mathcal{H}$ is finite-dimensional, P is a continuous functional and the stability of A implies that $K(A^n x_0, A^n y_0) \to 0$. Thus Theorem 4.4 implies thut λ is 0; that is, E has a signed spectral factorization and consequently λ has one sign.

If $\mathcal{G}$ is not finite-dimensional, then $(I-zA)^{-1}B$ cannot be rational. Thus $(I-zA)^{-1}Bp(z)$ can never be a polynomial in z when $p(z)$ is a polynomial in z. Thus property (4.11a) holds for any real-valued $\lambda \in L^{\infty}(\mathbb{C})$. Set $\lambda(e^{i\theta}) \equiv \sup_{\theta} |E(e^{i\theta})| + 1 = \lambda$. The function $E+\lambda \geqq \delta > 0$ certainly satisfies (4.11b). Since λ is a constant, (4.11c) is vacuously satisfied by λ. The resulting fixed point has a positive indicator. The function $E-\lambda$ gives a fixed point with negative indicator.

4.3. The linear part. Finally we shall consider maps which are not purely quadratic. In a formal sense, the linear and quadratic parts of the map $\mathcal{F}$ have a very nice relationship as we shall see. Technically speaking, the problems might be incompatible because the space $\tilde{\mathcal{R}}$ is crucial to the quadratic problem, and the operator $\Gamma_1 A \Gamma_2 = A_2$, which links the linear and quadratic parts of the equation, might have range very much disjoint from $\tilde{\mathcal{R}}$. Note that if P_1 satisfies (4.3) and Range A_2 does not strongly intersect the domain of definition of P_1, then the last term in (4.4) is not well-defined. Throughout most of this we shall assume that Range $A_2 \subset \tilde{\mathcal{R}}$ and call the linear and quadratic parts of $\mathcal{F}$ *compatible* whenever this happens. This assumption is certainly satisfied when $\mathcal{H}$ is finite-dimensional. Although weaker assumptions will do, we shall assume that B has finite-dimensional range in order to avoid annoying details. If $\|A^n\| \leqq Ka^n$ for some $a < 1$, then A is called *exponentially stable.*

Suppose that A is exponentially stable, that E has an outer signed factorization $M\mathcal{S}^*M$ and that P_1 is the fixed point of $\mathcal{F}_1$ which corresponds to it. Clearly, P_1 satisfies (4.3). Next we seek a solution to (4.4). A formal solution is

$$P_2 = \sum_{k=2}^{\infty} N^{*k} P_1 A_2 A_3^{k-1} + \sum_{k=1}^{\infty} N^{*k} Q_2 A_3^k, \tag{4.12}$$

where $N^* = A_1^* - A_1^* P_1 B(R + B^* P_1 B)^{-1} B^*$. Note by (2.7) that N miraculously is $A_1^* + (G^{-1}F^*)B$ and so it is the adjoint of the operator which propagates the states of the feedback system. We have, in § 3, a stability analysis for this operator. If M were invertible outer, then by Theorem 3.1 $N^k x \to 0$ for $x \in \tilde{\mathcal{R}}$. Thus $|P_1(N^k x, A_3^k y)| \leqq C\|A_3^k\| \leqq C' a^n$, and so (4.12) makes good sense. If M is not invertible outer, then given $u(z)$ in H^2 there is a function $g(z)$ such that $M(z)g(z) = u(z)$ for $|z| < 1$.

The identity (3.3) which underlies Theorem 3.1 implies that any trajectory $\{x_n\}$ of the feedback system satisfies $\|x_n\| \leqq Cr^n$ for any $r > 1$. Thus $\|Nx^k\| \leqq C(1/(a+\varepsilon))^k$, and this is clearly enough to guarantee that (4.12) defines a bilinear functional on $\tilde{\mathcal{R}} \times \mathcal{R}_2$.

The final step is to obtain a solution for (4.5). The final solution is

$$P_3 = \sum_{k=1}^{\infty} A_3^{*k} T A_3^k,$$

where

$$T = \Gamma_2 N^* P_1 A_2 + \Gamma_2 N^* P_2 A_3 - A_3^* P_2^* N \Gamma_2 - A_3^* P_2^* B[R + B^* P_1 B]^{-1} B^* P_2 A_3.$$

The first three operators are, in fact, bounded primarily because range $N\Gamma_2 \subset \tilde{\mathcal{R}}$. Also $B^* P_2 A_3$ is bounded because of its construction and the $\tilde{\mathcal{R}}$ continuity of P_1. Thus P_3 is a well-defined bounded operator and we have

THEOREM 4.7. *Suppose the map $\mathcal{F}$ of* (4.2) *with finite rank B and exponentially stable A has compatible linear and quadratic parts. Then E has an* (*outer*) *signed spectral factorization* (*if and*) *only if $\mathcal{F}$ has a fixed point.*

The fixed point in this theorem is a bilinear functional on the obvious subspace of $\mathcal{H}$, namely, $(\tilde{\mathcal{R}} \oplus \tilde{\mathcal{R}}^{\perp}) \times (\tilde{\mathcal{R}} \oplus \tilde{\mathcal{R}}^{\perp})$. It is continuous on $\tilde{\mathcal{R}}^{\perp}$ and has the continuity properties of $\mathcal{P}_0$ on $\tilde{\mathcal{R}}$.

Appendix. Symplectic maps. The linear fractional maps (4.1) we studied are close to the class of symplectic maps of C. L. Siegel [28] except we work with complex rather than real matrices. Complex symplectic maps have the form

$$\mathfrak{S}(K) = (\mathcal{B} + \mathcal{U}K)(\mathcal{D} + \mathcal{C}K)^{-1}, \tag{A.1}$$

where the coefficient matrix $\mathcal{M} = \begin{bmatrix} \mathcal{U} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{bmatrix}$ satisfies $\mathcal{M}^* \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \mathcal{M} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ or equivalently $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \mathcal{M} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \mathcal{M}^*$. These intertwining conditions are equivalent to $\mathcal{U}\mathcal{B}^* = \mathcal{B}\mathcal{U}^*$, $\mathcal{C}\mathcal{D}^* = \mathcal{D}\mathcal{C}^*$, $\mathcal{U}\mathcal{D}^* - \mathcal{B}\mathcal{C}^* = I$ which, in turn, are equivalent to $\mathcal{D}^*\mathcal{B} = \mathcal{B}^*\mathcal{D}$, $\mathcal{C}^*\mathcal{U} = \mathcal{U}\mathcal{C}$, $\mathcal{D}^*\mathcal{U} - \mathcal{B}^*\mathcal{C} = I$. If $\mathcal{D}$ is invertible, a straightforward computation shows that A is invertible; then $\mathcal{F}$ in (4.7) equals $\mathfrak{S}$ if and only if

$$A = \mathcal{D}, \quad Q = \mathcal{B}\mathcal{D}^{-1}, \quad C = \mathcal{D}^{-1}\mathcal{C}, \tag{A.2}$$

or equivalently, if

$$\mathcal{U} = QA^{-1}C + A^*, \quad \mathcal{B} = QA^{-1}, \quad \mathcal{C} = A^{-1}\mathcal{C}, \quad \mathcal{D} = A^{-1}. \tag{A.3}$$

This computation in fact shows

PROPOSITION A.1. *The map $\mathcal{F}$ of* (4.2) *is symplectic if and only if A is invertible. Any symplectic map with $\mathcal{D}$ invertible can be written in the form* (4.2) *with A invertible.*

The function E which determines the fixed-point behavior of $\mathfrak{S}$ is

$$E(e^{i\theta}) = \operatorname{sgn} \mathcal{D}^{-1}\mathcal{C} + |\mathcal{D}^{-1}\mathcal{C}|^{1/2}(\mathcal{D} - e^{i\theta})^{-1*}\mathcal{D}^*\mathcal{B}(\mathcal{D} - e^{i\theta})^{-1}|\mathcal{D}^{-1}\mathcal{C}|^{1/2};$$

the *indicator* for a point K is $\Lambda_K = \operatorname{sgn} \mathcal{D}^{-1}\mathcal{C} + |\mathcal{D}^{-1}\mathcal{C}|^{1/2} K |\mathcal{D}^{-1}\mathcal{C}|^{1/2}$ and Theorem 4.1 translates to

THEOREM A.2. *If $\mathfrak{S}$ is the symplectic map* (A.1) *and the eigenvalues of $\mathcal{D}$ lie outside of $|z| = 1$, then $\mathfrak{S}$ has a self-adjoint fixed point K in $\mathcal{P}_0$ with nonnegative indicator if and only if $E \geqq 0$. It has a self-adjoint fixed point* (*if and*) *only if E has an* (*outer*) *signed factorization.*

Even though the class of maps given by (4.2) is not the same as the symplectic maps, these maps do take the set of matrices K with $\operatorname{Im} K > 0$ into those with $\operatorname{Im} K \geqq 0$. This is true because, formally,

$$\mathcal{F}(K) - \mathcal{F}(K)^* = A^*(1 + CK)^{-1*}(K - K^*)(1 + CK)^{-1}A,$$

and a glance at (4.2) reveals that $\mathcal{F}$ is well-defined when $\operatorname{sgn} C + |C|^{1/2} K |C|^{1/2}$ is invertible; $\operatorname{Im} K > 0$ implies such invertibility.

REFERENCES

[1] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[2] R. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[3] F. M. CALLIER AND C. A. DESOER, *Necessary and sufficient conditions for stability for n-input, n-output convolution feedback systems with a finite number of unstable poles*, IEEE Trans. Automatic Control, AC-18 (1973), no. 3.

[4] J. H. DAVIS, *Stability conditions derived from spectral theory: Discrete systems with periodic feedback*, this Journal, 10 (1972), pp. 1–13.

[5] R. G. DOUGLAS, *On majorization, factorization and range inclusion of operators on Hilbert space*, Proc. Amer. Math. Soc., 17 (1966), pp. 413–415.

[6] ———, *Banach algebra techniques in the theory of Toeplitz operators*, Regional Conf. Series, vol. 15, American Mathematical Society, Providence, R.I., 1972.

[7] R. G. DOUGLAS AND J. W. HELTON, *Inner dilations of analytic matrix functions and Darlington synthesis*, Acta Sci. Math. (Szeged), 34 (1973), pp. 61–67.

[8] P. FUHRMANN, *On realization of linear systems and applications to some questions of stability*, to appear.

[9] ———, *Exact controllability and observability and realization theory in a Hilbert space*, to appear.

[10] I. C. GOHBERG AND M. G. KREIN, *Systems of integral equations on a half-line with kernels depending on the difference of the arguments*, Amer. Math. Soc. Transl., 14 (1960), pp. 217–287.

[11] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.

[12] J. W. HELTON, *Discrete time systems, operator models and scattering theory*, J. Functional Analysis, to appear.

[13] ———, *Operator techniques for distributed systems*, Proc. 11th Allerton Conf., October, 1973.

[14] K. Y. LEE, S. CHOW AND R. O. BARR, *On the control of discrete-time distributed parameter systems*, this Journal, 10 (1972), pp. 361–376.

[15] B. P. MOLLINARI, *The stabilizing solution of the algebraic Riccati equation*, this Journal, 11 (1973), pp. 262–271.

[16] ———, *Equivalence relations for the algebraic Riccati equation*, this Journal, 11 (1973), pp. 272–285.

[17] ———, *The stable regulator problem and its inverse*, IEEE Trans. Automatic Control, AC-18 (1973), no. 5.

[18] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on a Hilbert Space*, North-Holland, Amsterdam, 1970.

[19] M. RABINDRANATHAN, *On the inversion of Toeplitz operators*, J. Math. Mech., 19 (1969-70), pp. 195–206.

[20] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics*, Academic Press, New York, 1972.

[21] M. ROSENBLUM, *On the operator equation* $BX - XA = Q$, Duke Math., J., 23 (1956), pp. 263–269.

[22] ———, *Summability of Fourier series in* $L^p(d)$, Trans. Amer. Math. Soc., 105 (1962), pp. 32–92.

[23] M. ROSENBLUM AND J. ROVNYAK, *The factorization problem for nonnegative operator valued functions*. Bull. Amer. Math. Soc., 77 (1971), pp. 287–318.

[24] E. L. TITCHMARSH, *The Theory of Functions*, 2nd ed., Oxford University Press, London, 1939.

[25] J. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), no. 6.

[26] H. HELSON AND G. SZEGO, *A problem in production theory*, Ann. Mat. Pure Appl., 51 (1960), no. 4.

[27] R. HUNT, B. MUCKENHOUPT AND R. WHEEDEN, *Weighted norm inequalities for the conjugate function and Hilbert transform*, Trans. Amer. Math. Soc., 176 (1973), pp. 227–251.

[28] C. L. SIEGEL, *Symplectic Geometry*, Academic Press, New York, 1968.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

BOUNDARY CONTROL OF PARABOLIC DIFFERENTIAL EQUATIONS IN ARBITRARY DIMENSIONS: SUPREMUM-NORM PROBLEMS*

KLAUS GLASHOFF AND NORBERT WECK†

Abstract. The temperature distribution of a given body Ω in $\mathbb{R}^N$, $N \geqq 1$, is controlled via the temperature $\mathbf{u}$ of the medium surrounding Ω. The task is to choose $\mathbf{u}$ (subject to certain restrictions) in such a way that the temperature distribution at time $t = T$ of Ω comes as close as possible (with respect to the supremum-norm, i.e., the norm in $C(\bar{\Omega})$) to a fixed given function. We study the question of existence of optimal controls and prove the bang-bang principle for two different control problems (in the first case, the control $\mathbf{u}$ is time- and space-dependent and in the other one $\mathbf{u}$ is only time-dependent). In the last part of this paper, we formulate an abstract minimum-norm problem and derive general theorems on the existence and characterization of optimal controls.

Introduction. The temperature distribution of a heated body Ω can be described by a well-known parabolic initial-boundary value problem. In this paper, we consider some control-theoretic questions arising in connection with such a heating process.

We give a short explanation in technological terms. Suppose that we can vary (subject to certain restrictions) the temperature $u(t, \xi)$ $(t \in [0, T], \xi \in \partial\Omega)$ of the medium which surrounds the body Ω. Here $[0, T]$ is a fixed time interval. The task is to choose u under the given restrictions in such a manner that the temperature distribution $y(T, x)$ $(x \in \Omega)$ of the body at the time-level T comes as close as possible to some desired temperature $z(x)$, $x \in \Omega$.

Problems of this type have been considered by Yegorov [1], Plotnikov [2], Butkovskiy [3]. Related questions were studied by many authors. Let us mention only Fattorini [4], [5], Lions [6]. The main distinction between these publications and ours is that as a measure for the deviation of $y(T, x)$ from $z(x)$, we take the supremum-norm (the norm in $C(\bar{\Omega})$ of $y(T, \cdot) - z(\cdot)$). This leads to some interesting questions both in analysis and optimization theory.

The paper is organized as follows: In § 1 we formulate the parabolic initial-boundary value problem and collect some facts about its solutions which can be developed in a series with respect to the eigenfunctions v_k of the corresponding elliptic eigenvalue problem. (The proofs of some auxiliary theorems are given in Appendix A). In § 2 we formulate the minimum-norm control problem (P1), where u is restricted to the unit ball U of $L_\infty(\Gamma)$, $\Gamma = (0, T) \times \partial\Omega$. We prove an existence theorem by showing that the 'reachable set' $S(U)$ is compact in $C(\bar{\Omega})$. Then we characterize the solutions of (P1) (bang-bang principle) which immediately implies the uniqueness of the optimal control. The bang-bang principle does *not* hold for arbitrary $z \in L_\infty(\Omega)$ as we show by a simple counterexample.

In § 3 we assume that the boundary control function is of the form $g(\xi) \cdot u(t)$ $(\xi \in \partial\Omega, t \in [0, T])$, where g is given and fixed and u can be chosen in the unit ball of $L_\infty(0, T)$. This model is seemingly better realizable in practical applications.

* Received by the editors May 6, 1975.

† Fachbereich Mathematik, Technische Hochschule Darmstadt, Darmstadt, West Germany.

Existence of an optimal control of the resulting problem (P2) is proven as in § 2. Uniqueness and the bang-bang property hold for $z \in C(\bar{\Omega})$ provided z and g have the same symmetry properties which we formulate in terms of the eigenfunctions v_k.

In § 4 we put our control problem into a more general framework. After some comments on the existence of optimal solutions of the abstractly defined control problem, we treat two different approaches to the bang-bang principle which correspond to the methods used for (P1) and (P2) in § 2 and § 3, respectively. The first method is based on the controllability and normality properties of the linear operator S which appears in the definition of the general control problem. Controllability and normality of S are defined by means of certain properties of the range of S and the range of its adjoint S', respectively. We use the separating hyperplane theorem in order to prove a bang-bang principle for minimum-norm problems with controllable and normal operators. The other method which, for some minimum-time problems, was, for example, used by Fattorini [4] proceeds as follows: one has to prove "strong controllability" of the operator S mentioned above—this immediately implies the validity of the bang-bang principle by using elementary calculations (this corresponds to the proof of Theorem 3). The section is finished by showing the connection between the two methods of proof for the bang-bang principle; we prove that strong controllability of an operator is equivalent to controllability together with normality. We conclude the paper by giving an example for problem (P2) in § 3. We consider the boundary control for the heat equation in the unit ball of R^N, $N \geqq 2$.

1. The initial-boundary value problem.

1.1. Function spaces, norms and bilinear forms. The spaces $L_p(S)$ $(1 \leqq p \leqq \infty)$, $C(S)$, $C_k(S)$ and $C_\infty(S)$ are defined in the usual way as is the notion of "support of f (supp f)" if f belongs to one of these spaces (cf., e.g., [7]). Let us fix some notations:

$$(f, g)(S) := \int_S f(x)g(x)\,dx \qquad \text{if either } f, g \in L_2(S) \text{ or } f \in L_1(S) \text{ and } g \in L_\infty(S),$$

$$\|f\|_p(S) := \left[\int_S |f(x)|^p\,dx\right]^{1/p} \qquad \text{if } p \in L_p(S), \quad 1 \leqq p < \infty,$$

$$\|f\|_\infty(S) := \operatorname{ess\,sup}_{x \in S} |f(x)| \qquad \text{if } f \in L_\infty(S),$$

$$\|f\|_\infty(S) := \max_{x \in S} |f(x)| \qquad \text{if } f \in C(S), \quad S \text{ compact},$$

$$\langle \alpha, f \rangle(S) := \alpha(f) \qquad \text{if } \alpha \in C(S)', \quad f \in C(S),$$

where $C(S)'$ denotes the topological dual of $C(S)$. We often write (f, g) instead of $(f, g)\,(S)$, etc.

1.2. Formulating the initial-boundary value problem. Let L denote a symmetric and uniformly elliptic operator of second order in some bounded region $\Omega \subset R^N$:

$$Ly(x) := \sum_{i,j=1}^{N} \partial_i(a_{ij}(x)\partial_j y(x)) + a(x)y(x),$$

$$\partial_i := \frac{\partial}{\partial x_i}, \quad a,\ a_{ij} \in C_\infty(\bar{\Omega}), \quad a_{ij} = a_{ji},$$

$$\sum_{i,j} a_{ij}(x)p_i p_j \geqq c_0|p|^2, \qquad c_0 > 0.$$

Furthermore we assume that $\partial\Omega$ is a C_∞-manifold.

Remark. For part of what follows, a, a_{ij} and $\partial\Omega$ actually only need to belong to some class C_k, whereas, for some other considerations, they have to be analytic. In the latter case, we shall say "all data are analytic". Given β, $T \in R^+$ let us introduce

$$n(\xi) := \text{outer normal in } \xi \in \partial\Omega,$$

$$\partial y(\xi) := \sum_{i,j} n_i(\xi) \cdot a_{ij}(\xi)\partial_j y(\xi), \qquad \xi \in \partial\Omega,$$

$$By(\xi) := \beta\partial y(\xi) + y(\xi),$$

$$G := (0, T) \times \Omega,$$

$$\Gamma := (0, T) \times \partial\Omega,$$

$$\Gamma_t := (0, t) \times \partial\Omega, \qquad t \in (0, T).$$

We shall consider a parabolic initial-boundary value problem:

$$\partial_t y(t, x) - L_x y(t, x) = 0, \qquad (t, x) \in G, \tag{1.1}$$

$$y(0, x) = 0, \qquad x \in \Omega, \tag{1.2}$$

$$B_\xi y(t, \xi) = u(t, \xi), \qquad (t, \xi) \in \Gamma, \tag{1.3}$$

($\partial_t := \partial/\partial t$; the subscripts in L_x and B_ξ indicate that these differential operators are acting with respect to the space-variables only.)

1.3. The series solution of (1.1)–(1.3). Let λ_k and v_k denote the eigenvalues and eigenfunctions, respectively, of the following eigenvalue problem:

$$Lv(x) + \lambda v(x) = 0, \qquad x \in \Omega, \tag{1.4}$$

$$Bv(\xi) = 0, \qquad \xi \in \partial\Omega. \tag{1.5}$$

The following facts are well known (Agmon [7, Thm. 14.6, p. 103 ff.]).

$$\{v_k\} \text{ is a complete orthonormal system in } L_2(\Omega). \tag{1.6}$$

$$\lambda_k \to +\infty, \qquad \lambda_k \sim c \cdot k^{2/N}. \tag{1.7}$$

$$\|v_k\|_\infty(\bar{\Omega}) = O(k^m) \quad \text{for some } m \in \mathbb{N}. \tag{1.8}$$

(1.8′) $v_k \in C_\infty(\bar{\Omega})$. Equation (1.8) holds for any derivative of v_k, too.

Let us introduce

$$g(t, x; \tau, \xi) := \sum_{k=1}^{\infty} e^{-(t-\tau)\lambda_k} v_k(x) v_k(\xi),$$

$$0 \leqq \tau < t \leqq T, \quad x \in \bar{\Omega}, \quad \xi \in \partial\Omega.$$

We have

$$g(t, x; \tau, \xi) \geqq 0, \tag{1.9}$$

$$g(t, x; \cdot, \cdot) \in L_1(\Gamma), \qquad \|g(t, x; \cdot, \cdot)\|_1(\Gamma) \leqq 1. \tag{1.10}$$

These follow from the maximum principle [8] since

$$y(u; t, x) = \int_{\Gamma_t} g(t, x; \tau, \xi) u(\tau, \xi)\, d\tau\, d\xi \tag{1.11}$$

solves (1.1)–(1.3) for $u \in C_\infty(\bar{\Gamma})$. Expressions (1.9)–(1.11) are discussed in Appendix A.

2. Control time- and space-dependent.

2.1. Statement of the problem and an existence theorem. From (1.10) it is clear that (1.11) makes sense for $u \in L_\infty(\Gamma)$, too. So we can define the mapping

$$S : L_\infty(\Gamma) \to L_\infty(\Omega),$$

$$u \mapsto Su := y(u; T, \cdot).$$

Given $z \in C(\bar{\Omega})$, we want to solve the problem

$$\|Su - z\|_\infty(\bar{\Omega}) = \min, \qquad u \in U := \{u \in L_\infty(\Gamma) / \|u\|_\infty(\Gamma) \leqq 1\}. \tag{P1}$$

LEMMA 1. *Let χ_δ denote the characteristic function of $\Gamma_{T-\delta}$. Then*

$$\lim_{\delta \to 0} \sup_{u \in U} \|S(\chi_\delta u) - Su\|_\infty = 0.$$

Proof. From (1.9) we have

$$|S(\chi_\delta u)(x) - Su(x)| \leqq S(1 - \chi_\delta)(x) = y(1; \delta, x).$$

But $\|y(1; \delta, \cdot)\|_\infty \to 0$ by Lemma A.2 (see Appendix A).

COROLLARY. $S(L_\infty(\Gamma)) \subset C(\bar{\Omega})$.

Proof. For $u \in L_\infty(\Gamma)$ we have: $S(\chi_\delta u)$ is continuous by (1.7) and converges uniformly to Su by Lemma 1.

The corollary shows that we can consider S as a mapping into $C(\bar{\Omega})$ which we shall do in the sequel.

THEOREM 1. *$S(U)$ is compact. Problem (P1) is solvable.*

Proof. It is clear from (1.7) and (1.8) that

$$S_\delta : L_2(\Gamma) \to C(\bar{\Omega}),$$
$$u \mapsto S(\chi_\delta u),$$

is well-defined and compact for any $\delta > 0$. Since U is weakly compact in $L_2(\Gamma)$, Lemma 1 implies that $S(U)$ is compact. Therefore (P1) is solvable.

Remark. Compactness of $S(U)$—although it may be of some interest in itself—is *not* needed for proving (P1) to be solvable. Compare Lemma 3 and Theorem 7.

2.2. Uniqueness and the bang-bang property of optimal controls. The following is a special case ($K = \mathbb{N}$) of Theorem A.1 (see Appendix A).

LEMMA 2. *The finite linear combinations of the eigenfunctions v_k of* (14), (15) *are dense in $C(\bar{\Omega})$.*

Next we want to determine the adjoint S' of the mapping $S : L_\infty(\Gamma) \to C(\bar{\Omega})$.

LEMMA 3. *For $\alpha \in C(\bar{\Omega})'$ put*

$$w(t, x) := \sum_k \langle \alpha, v_k \rangle e^{-\lambda_k (T-t)} v_k(x). \tag{2.1}$$

Then

(i) $w \in C_\infty([0, T) \times \bar{\Omega})$,

(ii) $-\partial_t w(t, x) - L_x w(t, x) = 0$, $B_\xi w(t, \xi) = 0$,

(iii) $w|_\Gamma \in L_1(\Gamma)$,

(iv) $\langle \alpha, Su \rangle = (w|_\Gamma, u)$ *for* $u \in L_\infty(\Gamma)$, *i.e.*, $S'\alpha = w|_\Gamma$.

Proof. By (1.7) and (1.8′), the series (2.1) converges uniformly together with all its partial derivatives in any set $[0, T-\delta) \times \bar{\Omega}$. This implies (i) and (ii) (because of (1.8′) and (1.4), (1.5)). Let u be any element of $L_\infty(\Gamma)$. Then Lemma 1 shows

$$\begin{aligned} \langle \alpha, Su \rangle &= \lim_{\delta \to 0} \langle \alpha, S(\chi_\delta \cdot u) \rangle \\ &= \lim_{\delta \to 0} \langle \alpha, \sum_k v_k(\cdot) \int_\Gamma e^{-(T-\tau)\lambda_k} v_k(\xi) \chi_\delta(\tau) u(\tau, \xi)\, d\tau\, d\xi \rangle \\ &= \lim_{\delta \to 0} (\chi_\delta \cdot w|_\Gamma, u). \end{aligned}$$

This implies (iii) (if we put $u = \operatorname{sgn} w$) and (iv) (for arbitrary u). Let us state two results from the theory of parabolic equations.

LEMMA 4 (Mizohata [9]). *Let $w \in C_\infty([0, T) \times \bar{\Omega})$ be a solution of $-\partial_t w - L_x w = 0$. If $w|_\gamma = \partial_\xi w|_\gamma = 0$ for some nonvoid open subset $\gamma \subset \Gamma$, then $w = 0$.*

LEMMA 5 (Tanabe [10]). *Let $w \in C_\infty([0, T) \times \bar{\Omega})$ satisfy $-\partial_t w - L_x w = 0$ and $B_\xi w = 0$. If all data are analytic, then $w|_\Gamma$ is analytic, too.*

From these lemmas we can deduce the next theorem.

THEOREM 2. *Let all data be analytic and let $\gamma \subset \Gamma$ have positive measure. Then*

$$A(\gamma) := \{Su \mid u \in L_\infty(\Gamma), \operatorname{supp} u \subset \gamma\}$$

is dense in $C(\bar{\Omega})$.

Proof. If $A(\gamma)$ is not dense, we have an $\alpha \in C(\bar{\Omega})'$, $\alpha \neq 0$, such that $\langle \alpha, A(\gamma)\rangle = 0$. Let w be defined by (2.1). Then for any $u \in L_\infty(\Gamma)$ with supp $u \subset \gamma$, we have

$$0 = \langle \alpha, Su\rangle = (w|_\Gamma, u) = (w|_\gamma, u).$$

This implies $w|_\gamma = 0$. Since $w|_\Gamma$ is analytic, we find $w|_\Gamma = 0$. Lemma 3(ii) shows $\partial_\xi w|_\Gamma = 0$ and therefore $w = 0$ (by Lemma 4). Since the v_k are orthogonal, we find $\langle \alpha, v_k\rangle = 0$ for all k which contradicts Lemma 2 and $\alpha \neq 0$.

THEOREM 3. *If all data are analytic and if $z \notin S(U)$, then any solution $\hat{u}$ of* (P1) *satisfies $|\hat{u}(\tau, \xi)| = 1$ almost everywhere on Γ.*

Proof. If the theorem were false, we would have an optimal $\hat{u}$, some $\delta > 0$ and a subset $\gamma \subset \Gamma$ of positive measure such that $|u| \leqq 1 - \delta$ on γ. By Theorem 2 there exists $u_1 \in L_\infty(\Gamma)$ with supp $u_1 \subset \gamma$ such that

$$\|S(\hat{u} + u_1) - z\|_\infty \leqq \tfrac{1}{2}\|S\hat{u} - z\|_\infty.$$

Defining $\eta := \delta\|u_1\|_\infty^{-1}$ we find

$$\begin{aligned}\|S(\hat{u} + \eta u_1) - z\|_\infty &\leqq (1-\eta)\|S\hat{u} - z\|_\infty + \eta\|S(\hat{u} + u_1) - z\|_\infty \\ &< \|S\hat{u} - z\|.\end{aligned}$$

Since $\hat{u} + \eta u_1 \in U$ this is not compatible with $\hat{u}$ being optimal.

THEOREM 4. *If all data are analytic, then there exists a unique optimal solution of* (P1).

Proof. Only uniqueness remains to be shown. Let u_1 and u_2 be solutions of (P1). Then $u := \frac{1}{2}(u_1 + u_2)$ is optimal, too. But u can have the bang-bang property ($|u| = 1$ a.e.) only if $u_1 = u_2$.

Remarks.

1. One can also discuss the control problem (P1) in the case $z \in L_\infty(\Omega)$. Then existence is a direct consequence of Theorem 1. But, in general, the bang-bang property (Theorem 2) no longer holds as can be seen from the following simple counterexample: Let $\bar{\Omega} = \Omega_1 \cup \Omega_2$, $\Omega_1 \cap \Omega_2 = \varnothing$, Meas. $(\Omega_i) > 0$. If $z(x)$ equals $+1$ for $x \in \Omega_1$ and -1 for $x \in \Omega_2$, then z cannot be better approximated by an Su than by $S0 = 0$ because Su is always continuous.

2. All the preceding results are true for $z \in L_p(\Omega)$ $(1 \leqq p < \infty)$ if optimality is defined with respect to the corresponding L_p-norm.

3. In the case of nonanalytic data, Lemma 4 still gives some information on optimal solutions (cf. § 4.4).

3. Control only time-dependent. In this section, we consider a slightly modified control problem which seems to be important for technical applications [2], [3]. Here the control function u depends on the time variable t only.

3.1. Statement of the problem and existence of an optimal control. We use the same notation as in § 1. Let $g \in L_\infty(\partial\Omega)$ be a fixed given function, and consider

the following initial-boundary value problem:

$$\partial_t y(t,x) - L_x y(t,x) = 0, \qquad (t,x) \in G, \tag{3.1}$$

$$y(0,x) = 0, \qquad x \in \Omega, \tag{3.2}$$

$$B_\xi y(t,\xi) = g(\xi)u(t), \qquad (t,\xi) \in \Gamma. \tag{3.3}$$

For $u \in L_\infty(0,T)$, the generalized solution $y(u;t,x)$ of (3.1)–(3.3) in the sense of § 2 is given by

$$y(u;t,x) = \int_0^t \sum_{k=1}^{\infty} g_k u(\tau)\, e^{-(T-\tau)\lambda_k} v_k(x)\, d\tau, \tag{3.4}$$

where

$$g_k := \int_\Omega g(\xi) v_k(\xi)\, d\xi. \tag{3.5}$$

By the Corollary to Lemma 1, we know that for all $u \in L_\infty(0,T)$,

$$y(u;T,\cdot) \in C(\bar{\Omega}),$$

and we can define the linear mapping

$$R : L_\infty(0,T) \to C(\bar{\Omega})$$

by $(Ru)(x) = y(u;T,x)$, $x \in \bar{\Omega}$.

Given $z \in C(\bar{\Omega})$, we consider the problem

(P2) Minimize $\|Ru - z\|_\infty$ under the constraint $u \in L_\infty(0,T)$, $\|u\|_\infty \leqq 1$.

In this section, we denote by U the unit ball of $L_\infty(0,T)$.

THEOREM 5. *$R(U)$ is a compact set of $C(\bar{\Omega})$. Problem* (P2) *has a solution.*

The *proof* can be given by exactly the same arguments as that of Theorem 1, but we want to give another proof by using the general existence result (Theorem 7) which we prove in the following section. By this theorem, all we have to show is that for $\alpha \in C(\bar{\Omega})'$, the relation

$$R'\alpha \in L_1(0,T)$$

holds where R' is the adjoint of R. Now

$$R = S \circ G,$$

where S was defined in § 2 and

$$G : L_\infty(0,T) \to L_\infty(\Gamma)$$

is defined by $(Gu)(\xi,t) = g(\xi)u(t)$, $(\xi,t) \in \Gamma$.

As $R' = G' \circ S'$ and (according to Lemma 3) $S'(C(\bar{\Omega})') \subset L_1(\Gamma)$, it remains to prove that

$$G'(L_1(\Gamma)) \subset L_1(0,T).$$

But this follows at once by definition of G, because for all $\lambda \in L_1(\Gamma)$, $u \in L_\infty(0,T)$,

we have

$$(\lambda, Gu) = \int_\Gamma \lambda(\xi, t) g(\xi) u(t)\, dt\, d\xi = \int_0^T \left(\int_{\partial\Omega} \lambda(\xi, t) g(\xi)\, d\xi \right) u(t)\, dt$$

and

$$G'\lambda = \int_{\partial\Omega} \lambda(\xi, \cdot) g(\xi)\, dt \in L_1(0, T)$$

by Fubini's theorem. It is not difficult to prove an existence theorem if we replace the $C(\bar{\Omega})$-norm in (P2) by any $L_p(\Omega)$-norm $\| \cdot \|_p$, $1 \leqq p \leqq \infty$ and take $z \in L_p(\Omega)$. Compare the remark following Theorem 7.

3.2. Uniqueness and a bang-bang principle. As a main tool for the proof of a bang-bang principle for problem (P2), we use the following "maximum principle" which we show during the proof of Theorem 8 in § 4.

LEMMA 6. *Let $\hat{u} \in L_\infty(0, T)$ be an optimal solution of* (P2). *We assume* $\inf\{\|Ru - z\|_\infty / \|u\|_\infty \leqq 1\} = \hat{\rho} > 0$. *Then there is a nonzero functional $\hat{\alpha} \in C(\bar{\Omega})'$ (independent of $\hat{u}$) such that*

$$\sup_{u \in U} (R'\hat{\alpha}, u) = (R'\hat{\alpha}, \hat{u}) \tag{3.6}$$

and

$$\langle \hat{\alpha}, R\hat{u} - z \rangle \neq 0. \tag{3.7}$$

We define $\lambda \in L_1(0, T)$ by $\lambda = R'\hat{\alpha}$; by (3.6) we conclude that

$$\hat{u}(t) = \operatorname{sgn} \lambda(t) \tag{3.8}$$

for all $t \in (0, T) \backslash N$, where N is the set of zeros of λ. So if we can show that N has measure zero, we know by (3.8) that $\hat{u}$ is the unique solution of (P2) and that $\hat{u}$ is "bang-bang".

In contrast to the preceding problem (P1) in § 1, the bang-bang principle generally does not hold for (P2)—this depends on certain "symmetry-properties" of the functions z and g.

We define the set $K(g) \subset \mathbb{N}$ by

$$K(g) = \{k \in \mathbb{N} / g_k = \int_{\partial\Omega} g(\xi) v_k(\xi)\, d\xi \neq 0\}.$$

Let

$$\hat{L}_g = \{y \in C(\bar{\Omega}) / (y, v_k) = 0 \text{ if } k \notin K(g)\} \tag{3.9}$$

and

$$L_g = \operatorname{span}\{v_k / k \in K(g)\}. \tag{3.10}$$

LEMMA 7. *The closure $\bar{L}_g$ (in $C(\bar{\Omega})$) of L_g is equal to $\hat{L}_g$.*

We give a proof for this in the Appendix A, Theorem A.1. (The statement of Lemma 7 would be trivial if we replaced $C(\bar{\Omega})$ by $L_2(\Omega)$ in the definition of $\hat{L}_g$ and in Lemma 7!)

LEMMA 8. *We assume*

(i) $\lambda_i \neq \lambda_k$ *for* $i, k \in K(g)$,

(ii) $(z, v_k) = 0$ *if* $k \notin K(g)$ (i.e., $z \in \hat{L}_g$),

(iii) $\inf_{u \in U} \|Ru - z\|_\infty = \hat{\rho} > 0$.

Then the set of zeros of λ *has measure zero.*

Remark. We do *not* require that the elliptic operator L has finite multiplicity (compare the example in § 5).

Proof of Lemma 8. As in Lemma 3, one can show that $\lambda = R'\hat{\alpha}$ is represented by

$$\begin{aligned} \lambda(t) &= \sum_{k \in \mathbb{N}} g_k \langle \hat{\alpha}, v_k \rangle \, e^{-\lambda_k (T-t)} \\ &= \sum_{k \in K(g)} g_k \langle \hat{\alpha}, v_k \rangle \, e^{-\lambda_k (T-t)}, \qquad t \in [0, T). \end{aligned} \tag{3.11}$$

λ is analytic for $t < T$, so if the conclusion of the theorem were not true, we would have $\lambda(t) = 0$ for all $t < T$. By standard analyticity arguments (see, for instance, Tsujioka [11, Lemma 3]) this implies

$$g_k \langle \hat{\alpha}, v_k \rangle = 0 \quad \text{for } k \in K(g)$$

and by definition of $K(g)$,

$$\langle \hat{\alpha}, v_k \rangle = 0, \quad \text{for } k \in K(g).$$

Using Lemma 7, we get

$$\langle \hat{\alpha}, y \rangle = 0 \quad \text{for } \textit{all } y \in \hat{L}_g. \tag{3.12}$$

But this is impossible because from (ii) it is clear that

$$R\hat{u} - z \in \hat{L}_g$$

and by (3.7) (which is true because of (iii)), we get

$$\langle \hat{\alpha}, R\hat{u} - z \rangle \neq 0$$

in contradiction to (3.12). This proves Lemma 8.

We remark that (3.8) together with (3.11) gives us some information about the "jumps" of the optimal control $\hat{u}$ as the set of zeros of λ can have an accumulation point at most at the right endpoint of the interval $[0, T]$. So we have proven the following theorem.

THEOREM 6. *Under the assumptions of Lemma* 8, *there is a unique optimal solution* $\hat{u}$ *of* (P2) *which is piecewise constant equal to* $+1$ *or* -1 *with finitely many "jumps" on any interval* $[0, T-\varepsilon]$, $\varepsilon > 0$.

This theorem can be slightly generalized if we assume that the control on the boundary is performed by the equation

$$B_\xi y(t, \xi) = \sum_{i=1}^{M} g^i(\xi) u^i(t), \qquad (\xi, t) \in \Gamma, \tag{3.13}$$

which replaces (3.3). Here the g_i are fixed $L_\infty(\partial\Omega)$-functions, and the controls $u^i \in L_\infty(0, T)$ are subject to the conditions $\|u^i\|_\infty \leqq 1$. The definition of $K(g)$ and

the assumption (i) have to be altered appropriately. In this case, one can prove that *one* of the components of each optimal control $\hat{u} = (\hat{u}^1, \hat{u}^2, \cdots, \hat{u}^M)$ is bang-bang. For further details see Appendix B.

4. A general approach to a class of control problems. In this section, we treat an abstract version of the control problems (P1), (P2) given above. We prove an existence theorem and introduce the concepts of controllability, normality and strong controllability in order to prove an abstract bang-bang principle in two different ways (this corresponds to the different proofs of Theorem 3 and Theorem 6). In Theorem 10, we show the connection between these two methods.

4.1. Statement of the problem and an existence result. Let $(X, \mathcal{T}, \mu)$ be a finite measure space. Given a normed space F, an element $z \in F$ and a linear mapping

$$T : L_\infty(\mu) \to F,$$

we consider the following optimization problem

$$\text{(P)} \qquad \begin{aligned} &\text{Minimize } \|Tu - z\|_F \\ &\text{under the constraint } u \in B_\infty, \end{aligned}$$

where B_∞ is the unit ball in $L_\infty(\mu)$:

$$B_\infty = \{u \in L_\infty(\mu) / \|u\|_\infty \leqq 1\}.$$

Let F' be the topological dual of F. By T' we denote the adjoint of T (which maps the algebraic dual F^* of F into the algebraic dual of $L_\infty(\mu)$). We recall that $L_\infty(\mu) = L_1(\mu)'$ by the Riesz representation theorem.

THEOREM 7. *If $T'(F') \subset L_1(\mu)$, then there exists an optimal solution $\hat{u}$ of* (P).

Proof. We consider the weak topologies $\sigma(L_\infty, L_1)$ in $L_\infty(\mu)$ and $\sigma(F, f')$ in F. Now $T'(F') \subset L_1(\mu)$ is equivalent to the fact that T is continuous with respect to the weak topologies (cf., e.g., [12]). Because of $L_\infty(\mu) = L_1(\mu)'$, B_∞ is $\sigma(L_\infty, L_1)$-compact by the Alaoglu theorem. So the set $T(B_\infty)$ is $\sigma(F, F')$-compact in F. As the norm $\|\cdot\|_F$ is $\sigma(F, F')$-lower semicontinuous on F, the proof is complete.

We remark that the assumption of this existence theorem is met for the operators S and R of §§ 2 and 3 (see Lemma 3 (iii) and Theorem 5). The existence of an optimal solution of (P1), (P2) does not depend on the compactness of the "reachable sets" $S(U)$ and $R(U)$, respectively, but on only their weak compactness which is implied by $S'(C(\bar{\Omega})') \subset L_1(\Gamma)$ and $R'(C(\bar{\Omega})') \subset L_1(0, T)$, respectively.

If F is one of the spaces $L_p(\Omega)$, $1 \leqq p \leqq \infty$, then $C(\bar{\Omega})$ is a subspace of F the norm, $\|\cdot\|_\infty$, of which is stronger than the norm in F. This implies

$$F' \subset C(\bar{\Omega})'.$$

By this we see that we can replace both $z \in C(\bar{\Omega})$ and $\|\cdot\|_\infty$ in (P1), (P2) by $z \in L_p(\Omega)$ and $\|\cdot\|_p$ for $1 \leqq p \leqq \infty$ and still have an existence theorem for these problems.

4.2. Controllability, normality and the bang-bang principle. We consider the general problem (P) and assume for the remaining part of this section, as in

Theorem 7, that

$$T'(F') \subset L_1(\mu).$$

The operator T is called *controllable* if $T(L_\infty(\mu))$ is a norm-dense subspace of F; T is called *normal* if for any nonzero $\lambda \in T'(F')$ the set of zeros of λ has μ-measure zero. A function $u \in B_\infty$ is called *bang-bang* if $\mu(M) = 0$ for the set

$$M = \{t \in X / |u(t)| < 1\},$$

i.e., if u is an extreme point of B_∞.

THEOREM 8. *Under the assumptions*

(i) $\inf\{\|Tu - z\|_F / u \in B_\infty\} = \hat{\rho} > 0$,

(ii) *T is controllable,*

(iii) *T is normal,*

there exists a unique *solution $\hat{u}$ of* (P), *and $\hat{u}$ is bang-bang.*

Proof. First we prove the maximum principle (see Lemma 6). Part (i) implies that there is a nonzero $\hat{\alpha} \in F'$ separating $T(B_\infty)$ and the open ball Q with radius $\hat{\rho}$ around z([12]):

$$\begin{aligned} &\langle \hat{\alpha}, y \rangle \leqq A, && y \in T(B_\infty), \\ &\langle \hat{\alpha}, z \rangle > A \end{aligned} \tag{4.1}$$

for some real A. Let $\hat{u}$ be an arbitrary solution of (P) (which exists by Theorem 7). Then $\|T\hat{u} - z\| = \hat{\rho}$; i.e., $T\hat{u}$ is in the closure of Q. This implies

$$\langle \hat{\alpha}, T\hat{u} \rangle \geqq A, \tag{4.2}$$

and comparing (4.1) and (4.2), we get the *maximum principle*

$$\langle \hat{\alpha}, T\hat{u} \rangle = \sup_{u \in B_\infty} \langle \hat{\alpha}, Tu \rangle \quad (= A) \tag{4.3}$$

and

$$\langle \hat{\alpha}, T\hat{u} - z \rangle = \langle \hat{\alpha}, T\hat{u} \rangle - \langle \hat{\alpha}, z \rangle < 0.$$

We define $\lambda = T'\hat{\alpha}$ and rewrite (4.3) as

$$\int_X \lambda(t)\hat{u}(t)\, d\mu = \sup_{u \in B_\infty} \int_X \lambda(t)u(t)\, d\mu. \tag{4.4}$$

The controllability of T implies that λ is not the zero function on X (as the nullspace of T' is the orthogonal of the range of T and because $T(L_\infty(\mu))$ is dense in F). The normality of T then implies that the set of zeros of λ has μ-measure zero. Therefore $\hat{u}$ is uniquely defined by (4.4) as

$$\hat{u}(t) = \operatorname{sgn} \lambda(t), \qquad \mu \text{ a.e. on } X,$$

which proves the theorem.

We remark that the operator R in § 3 is controllable for $F = \hat{L}_g$ under the conditions of Lemma 8. Normality of R is implied by analyticity of $R'\alpha$, $\alpha \in C(\bar{\Omega})'$.

If an operator $T : L_\infty(\mu) \to C(\bar{\Omega})$ is controllable and normal, then T has the

same properties as a mapping into any of the spaces

$$L_p(\Omega), \qquad 1 \leqq p < \infty,$$

as $C(\bar{\Omega})$ is a dense subspace of these spaces. So the conclusions of Theorem 8 hold if in (P2) we replace $z \in C(\bar{\Omega})$ by $z \in L_p(\Omega)$ and $\|\cdot\|_\infty$ by $\|\cdot\|_p$, $1 \leqq p < \infty$, respectively. The same arguments apply to (P1); compare the following § 4.3. All this is *not* true if we choose $p = \infty$ and $z \in L_\infty(\Omega)$ as $C(\bar{\Omega})$ is not dense in $L_\infty(\Omega)$. Although we have an existence result in this case too (see the remark following Theorem 7) the bang-bang principle cannot be proved with Theorem 8. This is not surprising in view of the counterexample given in § 2.

4.3. Strong controllability. Let M be a μ-measurable subset of X. χ_M is the characteristic function of M, and we define the operator $T_M : L_\infty(\mu) \to F$ by

$$T_M u = T(\chi_M u).$$

T is called *strongly controllable* if T_M is controllable for *any* μ-measurable $M \subset X$ with $\mu(M) > 0$.

THEOREM 9. *Assume*

(i) $\inf\{\|Tu - z\|_F \mid u \in B_\infty\} = \hat{\rho} > 0$,

(ii) *T is strongly controllable.*

Then there is a unique solution $\hat{u}$ of (P), *and $\hat{u}$ is* bang-bang.

We omit the easy proof (see Theorems 3 and 4).

Comparing Theorem 8 with Theorem 9 one may ask for the connections between controllability, normality and strong controllability. This is explained by the following theorem.

THEOREM 10. *T is controllable and normal if and only if T is strongly controllable.*

Proof. (a) Let T be strongly controllable. Then T is controllable by definition, and we have to show normality. If T is *not* normal, there is a $\lambda \in T'(F')$, λ not identically zero such that

$$\lambda(t) = 0, \qquad \mu \text{ a.e. } on\ M$$

for a measurable subset M with $\mu(M) > 0$. Now there is a nonzero $\alpha \in F'$ such that $\lambda = T'\alpha$. We choose $z \in F$ satisfying

$$\tfrac{1}{2}\|\alpha\|_{F'} \leqq \langle \alpha, z \rangle \tag{4.5}$$

which is possible by definition of $\|\cdot\|_{F'}$. Then for each $u \in L_\infty(\mu)$,

$$\|z - T_M u\|_f \geqq \|\alpha\|_{F'}^{-1} \cdot \langle \alpha, z - T_M u \rangle \geqq \tfrac{1}{2}$$

because of (4.5) and

$$\langle \alpha, T_M u \rangle = (T'\alpha, \chi_M u) = \int_X \lambda(t)\chi_M(t)u(t)\, d\mu = 0$$

(as $\lambda(t)\chi_M(t) = 0$, μ a.e. on X). Thus T_M is not controllable in contradiction to the assumption of strong controllability of T.

(b) Now assume that T is controllable, but not strongly controllable. We have to show that T is not normal. By assumption there is a μ-measurable set

$M \subset X$, $\mu(M) > 0$, such that the subspace

$$\{T_M u / u \in L_\infty(\mu)\}$$

is not dense in F. By a well-known corollary to the Hahn–Banach theorem, there is a nonzero functional $\alpha \in F'$ such that

$$\langle \alpha, T_M u \rangle = 0 \quad \text{for all } u \in L_\infty(\mu). \tag{4.6}$$

We define $\lambda \in L_1(\mu)$ by $\lambda = T'\alpha$. λ is *not* the zero function on X as T is assumed to be controllable. We choose $u(t) = \operatorname{sgn} \lambda(t)$, $t \in X$, in (4.6) and obtain

$$0 = \langle \alpha, T_M u \rangle = (T'\alpha, x_M u) = \int_M |\lambda(t)| \, d\mu$$

which shows that T is not normal.

By Theorem 10, we see that it would have been possible to prove the bang-bang principle for (P1), § 2 with the arguments used in the proof of the bang-bang principle for (P2) in § 3 and vice versa.

4.4. $\mathcal{M}$-normality and $\mathcal{M}$-controllability. Normality of an operator T can often be verified by using analyticity arguments, but the controllability-normality approach of Theorem 8 is of some interest also in "nonanalytic" cases: Let us assume that the conditions (i), (ii) of Theorem 8 are met. We weaken assumption (iii) in the following manner. Let $\mathcal{M}$ be a set of measurable subsets of X. We call the operator T *$\mathcal{M}$-normal* if for any nonzero $\lambda \in T'(F')$ the set N_λ of zeros of λ does not contain a member of $\mathcal{M}$. If $\mathcal{M}$ is the set of *all* measurable subsets of X with $\mu(M) > 0$, then each $\mathcal{M}$-normal operator T is normal in the terminology introduced in § 4.2. If we replace the condition (iii) of Theorem 8 by

$$T \text{ is } \mathcal{M}\text{-normal}, \tag{iii'}$$

then the proof of this theorem shows that under the assumptions (i), (ii) and (iii'), we get for any solution $\hat{u}$ of (P),

$$|\hat{u}(t)| = 1 \quad \text{a.e. on} \quad X \setminus N,$$

where N is a subset of X which does not contain a member of $\mathcal{M}$.

Let us look at an *example*: If the data in (1.1)–(1.3) are not analytic, we know by Lemma 4 (in connection with Lemma 3) that S is $\mathcal{M}$-normal if we take $\mathcal{M}$ as the set of all nonvoid *open* subsets of Γ. In addition to that, we know that S is controllable, for if $(S(L_\infty(\Gamma))$ is *not* dense in $C(\bar{\Omega})$, there exists an $\alpha \in C(\bar{\Omega})'$, $\alpha \neq 0$, such that

$$S'\alpha = w|_\Gamma = 0,$$

where w is defined as in Lemma 3. Arguing as in the last part of Theorem 2, we get a contradiction to $\alpha \neq 0$.

Thus by the preceding remarks we get the following result for the *nonanalytic case of* (P1): any solution $\hat{u}$ of (P1) is *not* bang-bang (i.e., $|u(t)| < 1$) at most on the union of a set of measure zero and a nowhere dense set.

Remarks.

1. The last result does not imply uniqueness of an optimal control $\hat{u}$ because the set of points where $|\hat{u}(t, \xi)| = 1$ can still have arbitrary small measure. Therefore it would be interesting to sharpen Lemma 4 without analyticity arguments!

2. The same arguments as in the proof of Theorem 10 show that T is controllable and $\mathcal{M}$-normal iff T is "*$\mathcal{M}$-controllable*," i.e., iff T_M is controllable for any $M \subset \mathcal{M}$.

3. In some cases, it may be easier to prove a bang-bang-type result by means of controllability and $\mathcal{M}$-normality arguments than by proving it "directly" by the (equivalent) $\mathcal{M}$-controllability property as it was performed in the "strong" case in Theorem 9 (Theorem 3).

5. Example. Let Ω be the unit ball in R^N, $N \geqq 2$. We consider the Laplace operator $L = \Delta_N$ and suppose that Ω is heated uniformly on $\partial\Omega$; i.e., we choose

$$g(\xi) = 1, \qquad \xi \in \partial\Omega, \tag{5.1}$$

and consider problem (P2) in § 3. In this case, it is convenient to formulate the initial-boundary value control problem with respect to spherical coordinates

$$(\vartheta, r) = (\vartheta_1, \cdots, \vartheta_{N-1}, r).$$

We get the equations

$$\frac{\partial}{\partial t} y(t, \vartheta, r) - \frac{1}{r^{N-1}} \frac{\partial}{\partial r}\left(r^{N-1} \cdot \frac{\partial}{\partial r} y(t, \vartheta, r)\right) + \frac{1}{r} B y(t, \vartheta, r) = 0,$$

$$y(0, \vartheta, r) = 0, \qquad (\vartheta, r) \in \Omega,$$

$$\beta \frac{\partial}{\partial r} y(t, \vartheta, 1) + y(t, \vartheta, 1) = u(t), \qquad (\vartheta, 1) \in \partial\Omega, \quad t \in (0, T),$$

where B is the Beltrami-differential operator.

The eigenfunctions of the corresponding elliptic boundary value problem are given by

$$\{v_{k,l,m} / k = 1, 2, 3, \cdots;\ l = 0, 1, 2, \cdots;\ m = 1, 2, \cdots, V_N(l)\},$$

where

$$v_{k,l,m} = S_m^{(l)}(\vartheta) r^{-N/2+1} \cdot J_{l+N/2-1}(\kappa_k^{(l+N/2-1)} r);$$

here $V_N(l)$ denotes the number of linear independent spherical harmonics $S_m^{(l)}$ of order l. $J_{l+N/2-1}$ is the Bessel function of order $l + N/2 - 1$. The eigenvalues are given by

$$\lambda_{l,k} = (\kappa_k^{(l+N/2-1)})^2,$$

where κ_k^{ν} is the kth solution of the equation

$$\kappa J_\nu'(\kappa) + \left(\frac{1}{\alpha} - 1 + \frac{N}{2}\right) J_\nu(\kappa) = 0. \tag{5.2}$$

(We remark that for $N \geqq 3$ the Laplace operator Δ_N is of infinite multiplicity as each $\lambda_{l,k}$ has multiplicity

$$V_N(l) = \binom{l+N-1}{N-1} - \binom{l+N-3}{N-1}.$$

See [13, p. 424].)

Now we determine the subspace $\hat{L}_g$ of $C(\bar{\Omega})$ defined in § 3.1.

$$g(\xi) = S_1^{(0)}(\xi)$$

by definition. Using the orthogonality relations of the system $\{S_m^{(l)} / l \geqq 0,\ m = 1, \cdots, V_N(l)\}$ in $L_2(\partial\Omega)$ we get

$$\begin{aligned} g_{k,l,m} &= \int_{\partial\Omega} g(\xi) v_{k,l,m}(\xi)\, d\xi \\ &= J_{l+n/2-1}(\sqrt{\lambda_{l,k}}) \int_{\partial\Omega} S_l^{(0)}(\xi) S_m^{(l)}(\xi)\, d\xi = 0 \end{aligned} \tag{5.3}$$

for all $l > 0$ and all $m = 1, \cdots, V_N(l)$. So the case $l = 0$, $m = 1 = V_N(0)$ is left; we get by definition of g and $v_{k,0,1}$ for $k \geqq 1$,

$$g_{k,0,1} = \int_{\partial\Omega} g(\xi) v_{k,0,1}(\xi)\, d\xi = \omega_N \cdot J_{N/2-1}(\kappa_k^{(N/2-1)}), \tag{5.4}$$

where ω_N is the surface area of the unit sphere. We see that

$$g_{k,0,1} \neq 0$$

for all $k \geqq 1$, because $g_{k,0,1} = 0$ for some $k \geqq 1$ would imply

$$J_{N/2-1}(\kappa_k^{(N/2-1)}) = 0$$

and by (5.2) also

$$J'_{N/2-1}(\kappa_k^{(N/2-1)}) = 0$$

which is impossible as $J_{N/2-1}(r)$ is a nonzero solution of a linear differential equation of order 2.

Let us take a "desired end-temperature" z defined on Ω which is independent of ϑ (i.e., it depends only on the distance r from the origin). In this case,

$$(z, v_{k,l,m}) = 0 \quad \text{for } l > 0, \quad m = 1, \cdots, V_N(l),$$

which implies that $z \in \hat{L}_g$. Thus if, in addition, z is "not reachable" in time T (i.e., condition (iii) in Lemma 6 is met), then the conclusions of the bang-bang principle (Theorem 8) hold in this case.

Appendix A: On the initial-boundary value problem (1.1)–(1.3). For $f \in C(\bar{\Omega})$ put

$$H(t)f(x) := \sum_{k=1}^{\infty} e^{-\lambda_k t}(f, v_k) v_k(x). \tag{A.1}$$

For $t>0$ this defines a bounded operator $H(t)$ from $C(\bar{\Omega})$ into $C(\bar{\Omega})$. It is convenient to introduce

$$F_p := \{f \in C_\infty(\bar{\Omega}) / Bf = BLf = \cdots = BL^{p-1}f = 0\} \tag{A.2}$$

which is a dense subspace of $C(\bar{\Omega})$ for each $p \in \mathbb{N}$.

LEMMA A.1. *If p is large enough and $f \in F_p$, then*

$$w(t, x) := H(t)f(x)$$

satisfies

$$w \in C_2(\bar{G}), \tag{A.3}$$

$$(-\partial_t + L_x)w(t, x) = 0,$$

$$w(0, x) = f(x), \tag{A.4}$$

$$B_\xi w(t, \xi) = 0.$$

Proof. For $f \in F_p$ we have

$$(f, v_k) = -\lambda_k^{-1}(f, Lv_k) = -\lambda_k^{-1}(Lf, v_k) = \cdots = (-\lambda_k)^{-p}(L^p f, v_k).$$

Therefore using (1.4)–(1.8′), we see that the series (A.1) converges uniformly in $\bar{G}$ together with its derivatives up to the second order and solves (A.3).

Lemma A.1 and the maximum principle [8] show

$$\|H(t)g\|_\infty(\bar{\Omega}) \leqq \|g\|_\infty(\bar{\Omega}), \qquad g \in F_p, \tag{A.4′}$$

$$\|H(t)g - g\|_\infty(\bar{\Omega}) \to 0, \qquad g \in F_p. \tag{A.5}$$

By a well known argument (A.4′) and (A.5) imply

$$\|H(t)f - f\|_\infty(\bar{\Omega}) \to 0, \qquad f \in C(\bar{\Omega}). \tag{A.5′}$$

(So $H(t)f(x)$ is a solution of (A.3) belonging to $C(\bar{G}) \cap C_\infty((0, T] \times \bar{\Omega})$ for $f \in C(\bar{\Omega})$, too).

THEOREM A.1. *For some subset $K \subset \mathbb{N}$ define*

$$L_K := \text{span}\{v_k | k \in K\},$$

$$\hat{L}_K := \{u \in C(\bar{\Omega}) | (u, v_k) = 0 \text{ if } k \notin K\}.$$

Then the closure (in $C(\bar{\Omega})$) of L_K is $\hat{L}_k$.

Proof. $f \in \hat{L}_K$ clearly implies $H(t)f \in \bar{L}_K$. Therefore (A.5′) shows $\hat{L}_K \subset \bar{L}_K$. The opposite inclusion being trivial, this proves the theorem.

Let us now prove the assertion (1.9)–(1.11). If $u \in C_\infty(\bar{\Gamma})$, we have a solution $y(u; t, x)$ of (1.1)–(1.3) which satisfies

$$y(u; \cdot, \cdot) \in C(\bar{G}) \cap C_\infty((0, T] \times \Omega), \tag{A.6}$$

$$u \geqq 0 \Rightarrow y(u; t, x) \geqq 0, \tag{A.7}$$

$$\|y(u; \cdot, \cdot)\|_\infty(\bar{G}) \leqq \|u\|_\infty(\bar{\Gamma}). \tag{A.8}$$

The last two inequalities follow from the maximum principle. Using (1.4)–(1.8′)

we find

$$y(u;t,x)=\sum_k \int_{\Gamma_t} e^{-\lambda_k(t-\tau)} v_k(\xi)u(\tau,\xi)\,d\tau\,d\xi v_k(x). \tag{A.9}$$

Therefore in order to verify (1.11), we must show that we can interchange summation and integration here. (The corresponding formula is *not* true in the case of Dirchlet boundary conditions.) By (1.7) and (1.8) if supp $u \subset [0, t-\delta]\times\bar{\Omega}$, we clearly have

$$y(u;t,x)=\int_{\Gamma_t} g(t,x;\tau,\xi)u(\tau,\xi)\,d\tau\,d\xi. \tag{A.9'}$$

Because of (A.7') this proves (1.9). Now let us choose a suitable $h_\delta \in C_\infty[0, T]$ such that

$$0 \leqq h_\delta \leqq 1,$$

$$h_\delta(\tau)=\begin{cases}0 & \text{if } \tau > t-\delta/2,\\ 1 & \text{if } \tau < t-\delta.\end{cases}$$

Then $0 \leqq y(h_\delta;t,x) \leqq 1$ which implies (1.10) (use (1.9) and the monotone convergence theorem). Finally we want to prove (1.11) for $u \in C_\infty(\bar{\Gamma})$. By the triangle inequality and (A.9') we have

$$\begin{aligned}\left| y(u;t,x)-\int_{\Gamma_t} g(t,x;\tau,\xi)u(\tau,\xi)\,d\tau\,d\xi\right| &\leqq |y(u;t,x)-y(u,h_\delta;t,x)|\\ &\quad+\left|\int_{\Gamma_t} g(t,x;\tau,\xi)(1-h_\delta(\tau))u(\tau,\xi)\,d\tau\,d\xi\right|.\end{aligned}$$

By (1.10) the second term converges to zero. From (A.8) we see that the first term can be estimated as follows:

$$|y(u(1-h_\delta);t,x)| \leqq \|u\|_\infty \cdot y(1;\delta,x).$$

Therefore (1.11) follows from Lemma A.2.

LEMMA A.2.

$$\lim_{\delta\to 0} \|y(1;\delta,\cdot)\|_\infty(\bar{\Omega}) = 0.$$

Proof. For $\mu := \sup a(x)$, there exists $v \in C_\infty(\bar{\Omega})$ satisfying

$$Lv-\mu v = 0, \qquad Bv = 1.$$

Introducing

$$z(t,x) := y(1;t,x)-e^{\mu t}\cdot v(x),$$

we see that z solves

$$\begin{aligned}\partial_t z - L_x z &= 0,\\ B_\xi z &= 0,\\ z(0,x) &= -v(x).\end{aligned}$$

Therefore

$$z(t, x) = -H(t)v(x),$$

and we find

$$\|y(1; \delta, \cdot)\|_\infty \leqq \|H(\delta)v - v\|_\infty + (e^{\mu\delta} - 1) \cdot \|v\|_\infty \to 0.$$

Appendix B. We consider the control of the parabolic equation by means of the boundary condition (3.11). The corresponding control problem can be written in the following form:

$$(P_M) \quad \text{Minimize} \left\| \sum_{i=1}^{M} R_i u^i - z \right\|_\infty \text{ under the constraints } u^i \in U,\ i = 1, \cdots, M.$$

Here R_i is the operator defined by (3.1), (3.2), (3.3) when $g(\xi)$ is replaced by $g^i(\xi)$. With the same arguments as those used for the proof of Lemma 6 (given in Theorem 8) one can show: If $\hat{u} = (\hat{u}^1, \cdots, \hat{u}^M)$ is an optimal solution of (P_M) and if

$$\|\sum R_i \hat{u}^i - z\|_\infty = \hat{\rho} > 0,$$

then there is a functional $\hat{\alpha} \in C(\bar{\Omega})'$ (independent of $\hat{u}$) such that

$$\sup_{u \in U} (R_i' \hat{\alpha}, u) = (R_i' \hat{\alpha}, \hat{u}^i)$$

and

$$\left\langle \hat{\alpha}, \sum_{1}^{M} R_i \hat{u}^i - z \right\rangle \neq 0.$$

We define $l_i \in L_1(0, T)$ by $l_i = R_i' \hat{\alpha}$, $i = 1, \cdots, M$, and conclude that

$$\hat{u}^i(t) = \operatorname{sgn} l_i(t), \qquad i = 1, \cdots, m, \tag{B.1}$$

for all $t \in (0, T) \setminus N_i$, where N_i is the set of zeros of l_i.

Let $\{\mu_k\}_{k \geqq 1}$ be the sequence of *distinct* eigenvalues of (1.4), (1.5). It is known that each μ_k has finite multiplicity $m(k)$: there are $m(k)$ linear independent eigenfunctions

$$w_{k1}, w_{k2}, \cdots, w_{km(k)} \tag{B.2}$$

spanning the eigenspace belonging to μ_k. Let $\{v_j\}_{j \geqq 1}$ be the sequence of *all* eigenfunctions in the following order:

$$w_{11}, \cdots, w_{1m(1)}, w_{21}, \cdots, w_{2m(2)}, w_{31}, \cdots, w_{3m(3)}, w_{41}, \cdots.$$

We define

$$K(g) = \left\{ j \in \mathbb{N} \;\middle|\; \int_{\partial\Omega} g^i(\xi) v_j(\xi)\, d\xi \neq 0 \text{ for some } i \in \{1, 2, \cdots, M\} \right\}.$$

The spaces $\hat{L}_g$ and L_g are defined by means of $K(g)$ as in (3.9), (3.10), and Lemma 7 holds also in this case.

Now we write down the sequence of eigenfunctions which is obtained from (B.2) by omitting the functions v_j, $j \notin K(g)$, and get (possibly after a renumbering) the sequence

$$w_{11}, \cdots, w_{1\sigma(1)}, w_{21}, \cdots, w_{2\sigma(2)}, w_{31}, \cdots, w_{3\sigma(3)}, w_{41}, \cdots,$$

where $0 \leqq \sigma(k) \leqq m(k)$. With these eigenfunctions, we define for each $k \geqq 1$, the $M \times \sigma(k)$-matrix G_k by

$$g_k^{i,j} = \int_{\partial\Omega} g^i(\xi) w_{kj}(\xi)\, d\xi, \qquad i = 1, \cdots, M, \quad j = 1, \cdots, \sigma(k).$$

Then we have the following analogue to Lemma 8.

LEMMA B.1. *We assume*

(i) rank $G_k = \sigma(k)$, $k = 1, 2, 3, \cdots$,

(ii) $(z, v_k) = 0$ *if* $k \notin K(g)$,

(iii) $\inf \|\sum R_i u^i - z\|_\infty = \hat{\rho} > 0$, $u^1, \cdots, u^M \in U$.

Then there is an index $\hat{m} \in \{1, \cdots, M\}$ *such that the set of zeros of* $l_{\hat{m}}$ *has measure zero.*

Proof. By definition of $l_i = R'\hat{\alpha}$,

$$l_i(t) = \sum_{k \in \mathbb{N}} \left(\sum_{j=1}^{m(k)} g_k^{i,j} \langle \hat{\alpha}, v_{kj} \rangle \right) e^{-\mu_k (T-t)}$$

$$= \sum_{k \in \mathbb{N}} \left(\sum_{j=1}^{\alpha(k)} g_k^{i,j} \langle \hat{\alpha}, w_{kj} \rangle \right) e^{-\mu_k (T-t)}, \qquad t < T, \quad 1 \leqq i \leqq M.$$

If we assume that $l_i(t) \equiv 0$ for all $i = 1, \cdots, M$, then we get with the same arguments as in the proof of Lemma 8:

$$\sum_{j=1}^{\sigma(k)} g_k^{i,j} \langle \hat{\alpha}, w_{kj} \rangle = 0, \qquad i = 1, \cdots, M,$$

for any $k \geqq 1$. Because of (i) this implies

$$\langle \hat{\alpha}, w_{kj} \rangle = 0, \qquad k = 1, 2, 3, \cdots, \quad i = 1, 2, 3, \cdots, \sigma(k).$$

Now the proof can be completed with the same arguments as in the proof of Lemma 8.

We remark that assumption (i) of Lemma B.1 is a generalization of the controllability conditions given by Fattorini [5] (see also Sakawa [14]):

$$\{R_1 u^1 + R_2 u^2 + \cdots + R_M u^M / u^i \in L_\infty(0, T),\ i = 1, \cdots, M\}$$

is dense—with respect to the $C(\bar{\Omega})$-norm—in $\hat{L}_g$.

Using (B.1) and Lemma B.1 we get immediately the next theorem.

THEOREM B.1. *Under the assumption of Lemma* B.1, *to each solution* $\hat{u} = (\hat{u}^1, \cdots, \hat{u}^M)$ *of* ((P_M) *there is an index* $\hat{m} \in \{1, \cdots, M\}$ *such that* $\hat{u}^{\hat{m}}$ *is piecewise constant equal to* $+1$ *or* -1 *with finitely many "jumps" on any interval* $[0, T-\varepsilon]$, $\varepsilon > 0$.

The theorem holds also in the case of a p-norm ($1 \leqq p < \infty$) in the formulation of (P_M) and $z \in L_p$; compare the remarks following Theorem 8.

REFERENCES

[1] YU. V. YEGOROV, *Some problems in the theory of optimal control*, Z. Vyčisl. Mat. i Mat. Fiz., 3 (1963), pp. 887–904 = U.S.S.R. Computational Math. and Math. Phys., 3 (1963), pp. 1209–1232.

[2] V. I. PLOTNIKOV, *The convergence of finite dimensional approximations (in the problem of the optimal heating of an inhomogeneous body of arbitrary shape)*, Ibid., 8 (1968), pp. 136–157 = U.S.S.R. Computational Math. and Math. Phys., 8 (1968), pp. 182–211.

[3] A. G. BUTKOVSKIY, *Distributed Control Systems*, American Elsevier, New York, 1969.

[4] H. O. FATTORINI, *Time optimal control of solutions of operational differential equations*, this Journal, 2 (1964), pp. 54–59.

[5] ———, *Boundary control systems*, Ibid., 6 (1968), pp. 349–385.

[6] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

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

[8] F. PROTTER AND H. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, N.J., 1967.

[9] S. MIZOHATA, *Unicité du prolongement des solutions pour quelques opérateurs différentiels paraboliques*, Mem. Coll. Sci. Kyoto Univ., A 31 (1958), pp. 219–239.

[10] H. TANABE, *On differentiability and analyticity of solutions of weighted elliptic boundary value problems*, Osaka J. Math., 2 (1965), pp. 163–190.

[11] K. TSUJIOKA, *Remarks on controllability of second order evolution equations in Hilbert spaces*, this Journal, 8 (1970), pp. 90–99.

[12] A. P. ROBERTSON AND W. J. ROBERTSON, *Topological vector spaces*, Cambridge Tracts in Mathematics and Mathematical Physics, Cambridge University Press, Cambridge, 1966.

[13] H. TRIEBEL, *Höhere Analysis*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1972.

[14] Y. SAKAWA, *Controllability for partial differential equations of parabolic type*, this Journal, 12 (1974), pp. 389–399.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

THE GENERALIZED PROBLEM OF BOLZA*

FRANK H. CLARKE†

Abstract. We consider the problem of minimizing a functional of the type

$$l(x(0), x(1)) + \int_0^1 L(t, x, \dot{x})\, dt,$$

where l and L are permitted to attain the value $+\infty$. We show that many standard variational and optimal control problems may be expressed in this form. In terms of certain generalized gradients, we obtain necessary conditions satisfied by solutions to the problem, in the form of a generalized Euler–Lagrange equation. We also extend the necessary condition of Weierstrass to this setting. The results obtained allow one to treat not only the standard problems but others as well, bringing under one roof the classical (differentiable) situation, the cases where convexity assumptions replace differentiability, and new problems where neither intervene. We apply the results in the final section to derive a new version of the maximum principle of optimal control theory.

1. Introduction. This paper will be concerned with problems of the following kind:

Minimize

$$l(x(0), x(1)) + \int_0^1 L(t, x(t), \dot{x}(t))\, dt, \tag{1}$$

where x is an absolutely continuous function from $[0, 1]$ to R^n with derivative $\dot{x}$ (almost everywhere), and where

$$L : [0, 1] \times R^n \times R^n \to (-\infty, \infty] \quad \text{and} \quad l : R^n \times R^n \to (-\infty, \infty]$$

are given functions. The form of this problem is superficially that of a problem of Bolza in the calculus of variations. Note however that l and L are extended-real-valued; this fact greatly increases the versatility of the problem. As we shall see, many apparently different classical problems and problems of optimal control may be placed within the framework of the above *generalized problem of Bolza.*

Our main concern will be the derivation of necessary conditions satisfied by a solution to the problem. We shall not impose differentiability; the conditions we obtain are given in terms of certain "generalized gradients" developed by the author in [2]. These generalize the usual derivative as well as the subgradients of convex analysis [9]. The main result (Theorem 1) incorporates as special cases necessary conditions for the following situations: classical problems incorporating various types of constraints, problems involving differential inclusions, optimal control problems (see Examples 1–3). The theorem also permits consideration of otherwise standard problems in which nondifferentiable functions appear.

* Received by the editors February 4, 1975, and in revised form August 4, 1975.

† Department of Mathematics, University of British Columbia, Vancouver, Canada. Now at U.E.R. Mathématiques de la Décision, Université Paris IX (Dauphine), Paris 16ème France. This research was supported in part under the National Research Council of Canada Grant A 9 082.

An extensive theory (including duality) for a generalized problem of Bolza was developed by R. T. Rockafellar in [7] and [8], under the assumption that $L(t, \cdot, \cdot)$ and $l(\cdot, \cdot)$ are convex functions. We shall see (§ 2) that Theorem 1 subsumes the part of that work dealing with necessary conditions. It should be pointed out, however, that the convex case lies at the heart of the proof of our result; Rockafellar's results were used in the Author's paper [4] which provides the main tool used here.

The plan of the paper is as follows. In § 2 we discuss the generalized problem of Bolza and some well-known problems reducible to it; we state the main result and discuss the hypotheses. Section 3 is devoted to proving Theorem 1. In § 4 we generalize the necessary condition of Weierstrass to our problem (Theorem 2), and in § 5 we combine this with Theorem 1 to obtain a form of the maximum principle of optimal control theory (Theorem 3).

We now complete the introduction by stating a few facts about generalized gradients. Details and proofs in the same notation may be found in [2].

Let C be a closed nonempty subset of R^n, and let c be a point in C. We define the *normal cone* to C at c, denoted $N_C(c)$, by

$$N_C(c) = \text{cl co}\{\lim_{i\to\infty} s_i(x_i - c_i)\},$$

where we consider all sequences of points $(s_i, x_i, c_i) \in [0, \infty) \times R^n \times R^n$ such that x_i converges to c, x_i has closest point c_i in C, and the indicated limit exists. We obtain in this way a closed convex cone. The cone dual to $N_C(c)$ may be looked upon as a cone of tangents. Although the above notions of tangency and normality are distinct from the several such notions found in variational theory, we may show that $N_C(c)$ reduces to the classical normal space if C is a C^1 manifold, and to the normals in the sense of convex analysis if C is convex.

Now let $f : R^n \to (-\infty, \infty]$ be a lower-semicontinuous (l.s.c.) function, and let x be a point where f is finite. We define the *generalized gradient* of f at x, denoted $\partial f(x)$, by

$$\partial f(x) = \{p \in R^n : (p, -1) \in N_C(x, f(x))\},$$

where C is the *epigraph* of f, i.e., the set

$$\text{epi } f = \{(s, r) \in R^n \times R : f(s) \leqq r\}.$$

Then if f is C^1 at x, $\partial f(x) = \{\nabla f(x)\}$, and if f is convex, $\partial f(x)$ is the *subdifferential* of f at x [9].

Among the properties of this set-valued gradient are the following:

(2a) Let f be locally Lipschitz. Then $\partial f(x)$ is the convex hull of all limits of the form

$$\lim_{i\to\infty} \nabla f(x_i),$$

where x_i is such that $\nabla f(x_i)$ exists, and where x_i converges to x.

(2b) Let f be the *indicator function* of a closed set C; i.e., $f(x)$ is 0 if x belongs to C and $+\infty$ otherwise. Then for c in C, $\partial f(c) = N_C(c)$.

2. Discussion of the problem. Main results. We now give three examples of problems which may be reframed as that of minimizing a Bolza functional of the form (1). As mentioned earlier, the minimization problems we consider involve the class of absolutely continuous functions $x : [0, 1] \to R^n$. We call such an x an *arc*, and we denote the class of arcs A_n.

Example 1. *A problem of Lagrange.* We seek to minimize

$$\int_0^1 g_0(t, x, \dot{x})\, dt$$

subject to the following constraints on the arc x:

$$x(0) = a, \quad x(1) = b, \quad g_1(t, x, \dot{x}) \leqq 0,$$

where g_0 and g_1 are, respectively, real- and vector-valued functions. We define l and L as follows:

$$L(t, s, v) = g_0(t, s, v) \quad \text{if } g_1(t, s, v) \leqq 0,$$

and L is $+\infty$ otherwise;

$$l(x_0, x_1) = 0 \quad \text{if } (x_0, x_1) = (a, b),$$

and l is $+\infty$ otherwise.

It is easy to see that the resulting generalized problem of Bolza is equivalent to the given problem of Lagrange.

Example 2. *Differential inclusion.* We are given a *multifunction* $E : [0, 1] \times R^n \to R^n$ (i.e., for each (t, s), $E(t, s)$ is a subset of R^n), and a subset C of $R^n \times R^n$. The problem is to minimize $g(x(1))$ over the arcs x satisfying

$$(x(0), x(1)) \in C$$

and also the *differential inclusion*

$$\dot{x}(t) \in E(t, x(t)) \quad \text{a.e.}$$

We set

$$L(t, s, v) = 0 \quad \text{if } v \in E(t, s),$$

and L is $+\infty$ otherwise; we define l by

$$l(x_0, x_1) = g(x_1) \quad \text{if } (x_0, x_1) \in C,$$

and $+\infty$ otherwise.

Example 3. *Optimal control.* To every integrable function $u(t)$ taking values in a given subset $U(t)$ of R^m we associate the solution x to the differential equation

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

The problem: minimize over all such pairs (x, u) satisfying

$$(x(0), x(1)) \in C$$

the functional

$$\int_0^1 g(t, x(t), u(t))\, dt.$$

We reframe the problem to one on R^{n+m} by the following definitions ((s, s*) represents a point in $R^n \times R^m$):

$$l(s_0, s_0^*, s_1, s_1^*) = 0$$

if (s_0, s_1) lies in C, and $+\infty$ otherwise;

$$L(t, s, s^*, v, v^*) = g(t, s, v^*)$$

if v^* lies in $U(t)$ and $v = f(t, s, v^*)$, and $+\infty$ otherwise.

Then the pair (z, ν) solves the optimal control problem iff (z, z^*) solves the above generalized problem of Bolza, where

$$z^*(t) = \int_0^t \nu(r)\, dr.$$

DEFINITION 1. The function L is *epi-Lipschitz* at the arc z if there exists an integrable function $k : [0, 1] \to R$ and a positive ε satisfying the following condition: for almost all t in $[0, 1]$, given two points s_1 and s_2 within ε of $z(t)$ and v_1 such that $L(t, s_1, v_1)$ is finite, there exist a point v_2 and a $\delta \geqq 0$ such that $L(t, s_2, v_2)$ is finite and

$$|(v_1 - v_2, L(t, s_1, v_1) - L(t, s_2, v_2) - \delta)| \leqq k(t)|s_1 - s_2|.$$

The above definition is equivalent to saying that the multifunction

$$E(t, s) = \operatorname{epi} L(t, s, \cdot)$$

is Lipschitz in s in the Hausdorff metric, which accounts for the terminology.

We shall adopt the following convention: if for a given arc x the integral or the sum in (1) is not defined, we set the functional (1) equal to $+\infty$. To say that the arc z *solves* the generalized problem of Bolza will mean the following: for $x = z$ the integral in (1) is defined and finite and $l(z(0), z(1))$ is finite; for any other arc x for which $l(x(0), x(1))$ is finite and the integral in (1) defined, the value of the Bolza functional (1) is no less than its value at z. We do not rule out the possibility that the integral in (1) equals $-\infty$ for some arc x. However, if a solution to the problem exists, this can only happen for an arc x for which $l(x(0), x(1))$ equals $+\infty$.

With the above convention for evaluating the Bolza functional, and for any ε in $(0, \infty]$ and s in R^n, we define

$$\Phi_\varepsilon^0(s) = \inf\left\{ l(x(0) + s, x(1)) + \int_0^1 L(t, x, \dot{x})\, dt : x \in A_n, |x - z| < \varepsilon \text{ a.e.} \right\},$$

and we define $\Phi_\varepsilon^1(s)$ similarly for $l(x(0) + s, x(1))$ replaced by $l(x(0), x(1) + s)$.

The infimum in the original problem is then $\Phi_\infty^0(0) = \Phi_\infty^1(0)$, which we assume finite.

DEFINITION 2. The generalized problem of Bolza is *calm* at z if for some ε in $(0, \infty]$, for $i = 0$ *or* 1, we have

$$\liminf_{s \to 0} [\Phi_\varepsilon^i(s) - \Phi_\varepsilon^i(0)]/|s| > -\infty.$$

We now give the measurability hypothesis we shall be imposing.

DEFINITION 3. L is said to be *epi-measurable* (in t) if for each s in R^n the multifunction $E(t, s) = \operatorname{epi} L(t, s, \cdot)$ is Lebesgue measurable in t (we refer to [6] for relevant definitions).

The above is equivalent to $L(\cdot, s, \cdot)$ being measurable with respect to the σ-field generated by products of Lebesgue sets in $[0, 1]$ and Borel sets in R^n. A sufficient (but not necessary) condition assuring that L be epi-measurable is that it be l.s.c. in t and v.

The notation ∂l will denote the generalized gradient of l with respect to both its variables. The notation ∂L will denote the generalized gradient of the function $L(t, \cdot, \cdot)$; we shall not have occasion to consider gradients with respect to t.

THEOREM 1. *Let the arc z solve the generalized problem of Bolza, where the problem is calm at z. Suppose that l is l.s.c., and that $L(t, s, v)$ is epi-measurable in t, l.s.c. in (s, v) and epi-Lipschitz at z. Then there exists an arc p such that*

$$(\dot{p}(t), p(t)) \in \partial L(t, z(t), \dot{z}(t)) \quad \text{a.e.}, \tag{3}$$

$$(p(0), -p(1)) \in \partial l(z(0), z(1)). \tag{4}$$

Remarks. The generalized gradient relationships (3) and (4) are counterparts of, respectively, the usual Euler–Lagrange equation and what are referred to in the theory of optimal control as *transversality conditions.*

Although Theorem 1 appears to apply only to global minima, it is easily adapted to local minima. Suppose, for example, that the arc z is optimal only with respect to the arcs x satisfying

$$|x(t) - z(t)| \leqq \delta.$$

We may redefine $L(t, s, v)$ to be $+\infty$ if $|s - z(t)| > \delta$ (which preserves the lower-semicontinuity of $L(t, \cdot, \cdot)$) and obtain a global minimum. Note however that the hypothesis that L be epi-Lipschitz rules out "state constraints". That is, for all points s near $z(t)$, there must be some point v (possibly depending on t and s) such that $L(t, s, v)$ is finite.

Theorem 1 applies to "fixed-time problems"; of course the normalization to the interval $[0, 1]$ is merely a convenience.

In the remainder of this section, we shall single out some special cases in which the requirements of Theorem 1 are met, and then give a specific example of its use.

PROPOSITION 1. *If either of the following is satisfied, the generalized problem of Bolza is calm at z:*

(a) $$l(x_0, x_1) = l_0(x_0) + l_1(x_0, x_1),$$

where l_1 is finite and Lipschitz in x_1 in a neighborhood of $(z(0), z(1))$;

(b) $$l(x_0, x_1) = l_1(x_1) + l_0(x_0, x_1),$$

where l_0 is finite and Lipschitz in x_0 in a neighborhood of $(z(0), z(1))$.

Proof. (a) We choose ε such that l_1 is finite and Lipschitz in x_1 within 2ε of $(z(0), z(1))$. If K is the Lipschitz constant, we derive easily

$$\Phi^1_\varepsilon(s) - \Phi^1_\varepsilon(0) \geqq -K|s|,$$

for all s within ε of 0. Calmness follows. The proof of (b) is similar. Q.E.D.

The following gives a condition that may be used instead of the epi-Lipschitz hypothesis.

PROPOSITION 2. *Let the hypotheses of Theorem* 1 *hold, with the epi-Lipschitz condition replaced by the following: there is a positive* ε, *a measurable positive function* $\beta(t)$ *and an integrable function* $k(t)$ *such that for each* t, *the function* $L(t, s, v)$ *is Lipschitz in* s *on the set*

$$\{(s, v) : |s - z(t)| \leqq \varepsilon, |v - z(t)| \leqq \beta(t)\},$$

with Lipschitz constant $k(t)$. *Then the conclusions of Theorem* 1 *remain valid.*

Proof. Define L^* as follows:

$$L^*(t, s, v) = L(t, s, v)$$

if $|v - \dot{z}(t)| \leqq \beta(t)$, and $+\infty$ otherwise. The arc z continues to solve the problem of Bolza with l and L^*, and it is easy to see that all the hypotheses of Theorem 1 hold, once we establish that L^* is epi-Lipschitz at z. This we do now.

Let s_1 and s_2 within ε of $z(t)$ be given, as well as $(v_1, L^*(t, s_1, v_1))$. We have necessarily $|v_1 - \dot{z}(t)| \leqq \beta(t)$. Let $v_2 = v_1$. Then $L^*(t, s_2, v_2) = L(t, s_2, v_1)$ and hence

$$|(v_1 - v_2, L^*(t, s_1, v_1) - L^*(t, s_2, v_2))| \leqq k(t)|s_1 - s_2|.$$

We may now apply Theorem 1 for L^* and z. But $L(t, \cdot, \cdot)$ and $L^*(t, \cdot, \cdot)$ agree on a neighborhood of $(z(t), \dot{z}(t))$, and hence

$$\partial L(t, z(t), \dot{z}(t)) = \partial L^*(t, z(t), \dot{z}(t)). \qquad \text{Q.E.D.}$$

The condition in Proposition 2 requires that L be finite near z. This is certainly not necessary for the function L to be epi-Lipschitz at z, as the next result shows.

PROPOSITION 3. *Let* L *have the form of Example* 3, *where for each* t *and* u *in* $U(t)$, $f(t, \cdot, u)$ *and* $g(t, \cdot, u)$ *are Lipschitz within* ε *of* $z(t)$. *We suppose also that the Lipschitz constant* $k(t)$ *(the same for each* u *in* $U(t)$*) is integrable. Then* L *is epi-Lipschitz at* (z, z^*).

Proof. Let (s_1, s_1^*) and (s_2, s_2^*) lie within ε of $(z(t), z^*(t))$, and let a point (v_1, v_1^*, r_1) in the epigraph of $L(t, s_1, s_1^*, \cdot, \cdot)$ be given. Then for some u in $U(t)$ and nonnegative δ we have

$$v_1^* = u, \quad v_1 = f(t, s_1, u), \quad r_1 = g(t, s_1, u) + \delta.$$

Let us put

$$v_2^* = u, \quad v_2 = f(t, s_2, u), \quad r_2 = g(t, s_2, u) + \delta.$$

Then (v_2, v_2^*, r_2) is in epi $L(t, s_2, s_2^*, \cdot, \cdot)$ and

$$|(v_2 - v_1, v_2^* - v_1^*, r_2 - r_1)| \leqq 2k(t)|(s_1 - s_2, s_1^* - s_2^*)|. \qquad \text{Q.E.D.}$$

The following result gives a Lipschitz condition on L that may replace both the calmness and epi-Lipschitz hypotheses of Theorem 1.

PROPOSITION 4. *In Theorem* 1, *let the epi-Lipschitz and calmness hypotheses be replaced by the following: there exist a positive* δ, *a positive measurable function* $\alpha(t)$ *and an integrable function* $k(t)$ *such that a.e., on the set*

$$\{(s, v) : |s - z(t)| < \delta, |v - \dot{z}(t)| \leqq \alpha(t)\},$$

the function $L(t, \cdot, \cdot)$ *is Lipschitz with constant* $k(t)$. *Then the conclusions remain valid.*

Proof. We may suppose α integrable, bounded by 1, and

$$\int_0^1 \alpha(t)\, dt = \delta < 1.$$

The argument used in Proposition 2 shows that the function L^* defined there for $\beta = \alpha/2$ is epi-Lipschitz. As is again the case, the conclusions of Theorem 1 for L replaced by L^* are equivalent, so we need only show that the new problem is calm at z. Let x be any arc within $\delta/2$ of z for which $L^*(t, x, \dot{x})$ is finite a.e., and let s be any point within $\delta/2$ of 0 for which $l(x(0), x(1)+s)$ is finite. We set

$$y(t) = x(t) + (s/\delta) \int_0^t \alpha(r)\, dr.$$

Then $y(0) = x(0)$, $y(1) = x(1) + s$, x and y are within δ of z, and a.e. we have x and y within $\alpha(t)$ of $\dot{z}(t)$. Thus

$$\begin{aligned} l(x(0), x(1)+s) + \int_0^1 L^*(t, x, \dot{x})\, dt &= l(y(0), y(1)) + \int_0^1 L(t, x, \dot{x})\, dt \\ &\geqq l(y(0), y(1)) + \int_0^1 L(t, y, \dot{y})\, dt - \int_0^1 k(t)|(x - y, \dot{x} - \dot{y})|\, dt \\ &\geqq l(z(0), z(1)) + \int_0^1 L(t, z, \dot{z})\, dt - K|s|, \end{aligned}$$

where $K = (2/\delta) \int_0^1 k(t)\, dt$. If Φ_ε^1 corresponds to the new problem, with $\varepsilon = \delta/2$, we thus derive, for all s small,

$$\Phi_\varepsilon^1(s) \geqq \Phi_\varepsilon^1(0) - K|s|.$$

It follows that the new problem is calm. Q.E.D.

The following result shows that Theorem 1 subsumes the part of Rockafellar's work dealing with necessary conditions for convex problems of Bolza [8, Cor. 1].

PROPOSITION 5. *We assume the following hold*:

(a) *conditions* $A - D$ *of* [8];

(b) $\operatorname{ri}(\operatorname{dom} l) \cap \operatorname{ri}(F_L) \neq \varnothing$,

where

$$\operatorname{dom} l = \{(s_0, s_1) : l(s_0, s_1) < \infty\},$$

$$F_L = \left\{(s_0, s_1) : \textit{for some } x,\ x(0) = s_0,\ x(1) = s_1 \textit{ and } \int_0^1 L(t, x, \dot{x})\, dt < \infty\right\},$$

and ri *denotes "relative interior"* (see [9]);

(c) *for some positive* ε, *for any* s *within* ε *of the set*

$$\{z(t) : 0 \leqq t \leqq 1\},$$

there exist integrable functions v *and* k *such that*

$$L(t, s, v(t)) \leqq k(t).$$

Then the hypotheses of Theorem 1 *are satisfied.*

Remarks. We omit the proof, which is given in [1, Thm. 5.25]. Hypothesis (a) is essentially that l and $L(t, \cdot, \cdot)$ are convex, plus some mild regularity assumptions; (b) yields calmness, while (c) together with the convexity implies that L is epi-Lipschitz. In this case by [2, Prop. 3.19], relations (3) and (4) are in terms of subgradients of convex functions.

Concerning epi-measurability we note two cases, the first without proof.

PROPOSITION 6. *Under the hypotheses of Propositions* 2 *and* 4, *the epi-measurability hypothesis of Theorem* 1 *is satisfied if* L *is Lebesgue measurable as a function of* t.

PROPOSITION 7. *Let the function* L *be defined as in Example* 3, *where*

(a) f *is measurable in* t, *continuous in* u,

(b) g *is l.s.c. in* u,

(c) *for each* s, $g(\cdot, s, \cdot)$ *is measurable with respect to the* σ*-field generated by Lebesgue sets in* $[0, 1]$ *and Borel sets in* R^n (*see Definition* 3),

(d) *the multifunction* U *is measurable and closed-valued.*

Then L *is epi-measurable.*

Proof. Suppressing the fixed (s, s^*), we wish to show that the multifunction epi $L(t, s, s^*, \cdot, \cdot)$ is measurable in t. This last set equals the intersection of the sets

$$F_1(t) = \{(f(t, u), u, r) : u \in U(t), r \in R\},$$

$$F_2(t) = \{(p, u, g(t, u) + \delta) : p \in R^n, \delta \geqq 0\}.$$

Condition (c) implies the measurability of F_2. Since the intersection of measurable multifunctions is measurable, we need only show that F_1 is measurable. This is easy if one uses the fact that there exists a countable family $\{u_i\}$ of measurable functions such that

$$U(t) = \operatorname{cl}\{u_i(t)\} \quad \text{a.e.}$$

(This is a consequence of (d).) Q.E.D.

Example 4. We seek to minimize

$$\int_0^1 -|x(t)|\, dt,$$

where $x(0) = -4$ and $\dot{x}(t)$ is constrained to the interval $[5, 6]$. Following the pattern of Example 1, we find that

$$l(x_0, x_1) = 0 \quad \text{if } x_0 = -4,$$

and $+\infty$ otherwise, and L is given by

$$L(t, s, v) = -|s| \quad \text{if } 5 \leqq v \leqq 6,$$

and L is $+\infty$ otherwise. We verify easily that the hypotheses of Theorem 1 are satisfied (calmness follows from Proposition 1). Note that L is the sum of a Lipschitz function of s and an indicator function in v. Using (2a) and (2b), we derive from (3),

$$\begin{aligned}
\dot{p}(t) &= -1 && \text{if } z(t) > 0,\\
&= 1 && \text{if } z(t) < 0,\\
&\in [-1, 1] && \text{if } z(t) = 0,\\
p(t) &= 0 && \text{if } 5 < \dot{z}(t) < 6,\\
&\leqq 0 && \text{if } \dot{z}(t) = 5,\\
&\geqq 0 && \text{if } \dot{z}(t) = 6.
\end{aligned}$$

These relations allow us to determine the motion of (z, p) with time. For example, if on an interval, (z, p) is interior to the first quadrant, we must have on that interval $\dot{z}(t) = 6$ and $\dot{p}(t) = -1$. Hence (z, p) moves along one of the lines $z + 6p = \text{const}$. The other paths of motion are indicated in Fig. 1.

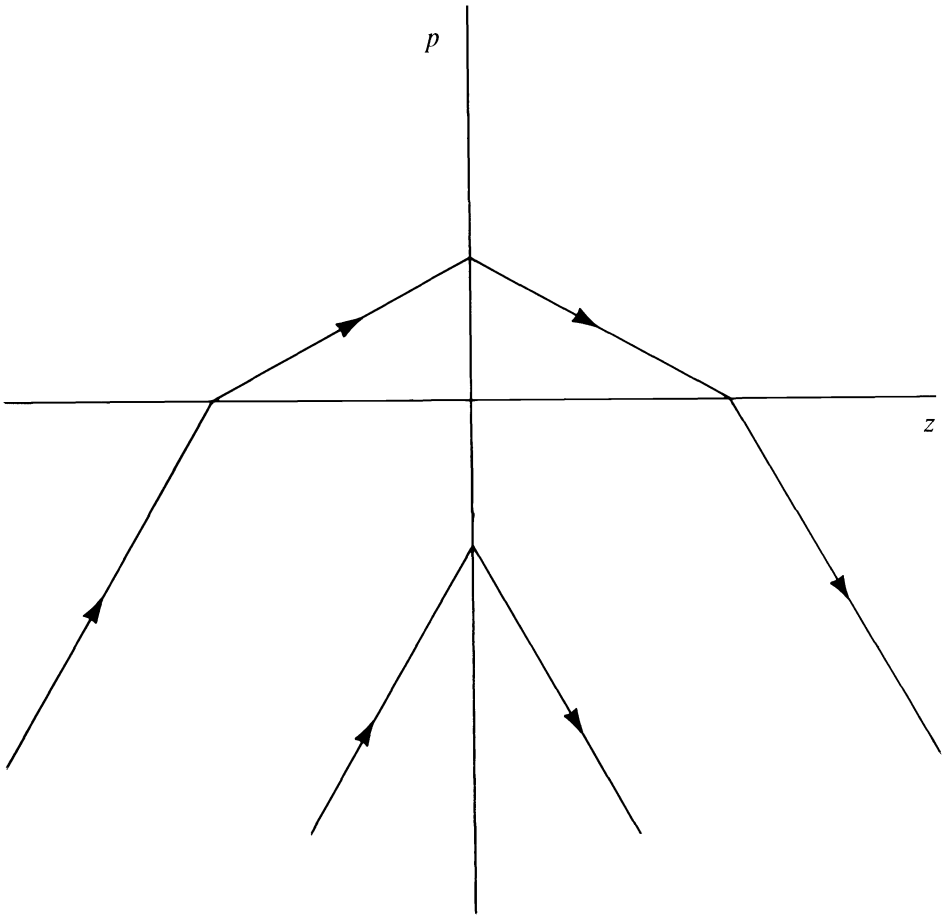


FIG. 1

Relation (4) yields $p(1) = 0$. It now remains to find what (z, p) arcs can begin at $t = 0$ on the line $z = -4$ and, traveling as indicated above, terminate at $t = 1$ on the line $p = 0$. We may easily show that the only way to do this is to begin at the point $(-4, -1/2)$, and that the arc z is then the one having $\dot{z}(t) = 5$ for $0 < t < 1/2$ and $\dot{z}(t) = 6$ for $1/2 < t < 1$.

3. Proof of Theorem 1. We assume that l and L satisfy the hypotheses of Theorem 1.

LEMMA 1. *If $s(t)$ is a continuous function, the multifunction*

$$F(t) = \operatorname{epi} L(t, s(t), \cdot)$$

is measurable.

Proof. The lemma follows immediately from the following result [1, Lemma 3.8]: if a multifunction $E(t, s)$ is measurable in t and continuous in the Hausdorff metric in s, then $E(t, s(t))$ is measurable. We omit the proof, since the usual proof for functions (via step functions) can be mimicked. Q.E.D.

LEMMA 2. *For any arc x, the function f defined by*

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

is measurable.

Proof. The multifunction $F(t) = \operatorname{epi} L(t, x(t), \cdot)$ is measurable by Lemma 1, and for any real number r the multifunction

$$G(t) = \{\dot{x}(t)\} \times (-\infty, r]$$

is measurable. We have

$$\{t : f(t) > r\} = \{t : F(t) \cap G(t) = \varnothing\},$$

and this set is measurable by [6, Cor. 1.3]. Q.E.D.

We now begin the proof of Theorem 1, with the following simplification: we shall assume that the calmness condition of Definition 2 holds for $i = 1$. Were this not the case, we could return to this situation by replacing t by $1 - t$ throughout, and $L(t, s, v)$ by $L(1 - t, s, -v)$. The equivalent transformed problem would satisfy the calmness condition at 1.

We set σ equal to the following finite number:

$$-\min\left\{0, \liminf_{s \to 0} [\Phi_\varepsilon^1(s) - \Phi_\varepsilon^1(0)]/|s|\right\},$$

and we call σ the *sensitivity* of the problem. The ε occurring here may be taken as the same that intervenes in Definitions 1 and 2.

We now adopt the following convention: s^* will refer to a point of the form (s^1, s^2, s^3, s^4) in $R^n \times R \times R^n \times R$. Similarly, an arc x^* has component arcs x^1 and x^3 in A_n and x^2, x^4 in A_1. We define a multifunction E by

$$E(t, s^*) = \{(v, r, 0, 0) \in R^n \times R \times R^n \times R : r \geqq L(t, s^1, v)\}$$

for $|s^1 - z(t)| \leqq \varepsilon/2$, and E is empty otherwise. We also define

$$z^*(t) = \left[z(t), \int_0^t L(r, z, \dot{z})\, dr, z(1), l(z(0), z(1))\right],$$

$$C_0 = \{s^* : l(s^1, s^3) \leqq s^4\},$$

$$C_1 = \{s^* : s^1 = s^3\}.$$

We set $m = 2(\sigma + 1)$.

LEMMA 3. *For some positive δ, the arc z^* minimizes*

$$x^2(1) + x^4(1) + md(x^*(1), C_1)$$

over the arcs x^* satisfying:

$$x^*(0) \in C_0, \quad \dot{x}^* \in E(t, x^*), \quad |x^3 - z(1)| < \delta, \quad |x^1 - z| < \delta.$$

Proof. Let us note first that z^* is feasible for the above problem. Suppose the lemma false. Then for each positive integer j there is an arc x_j^*, satisfying the conditions stated in the lemma for $\delta = 1/j$, such that

$$x_j^2(1) + x_j^4(1) + md(x_j^*(1), C_1) < z^2(1) + z^4(1). \tag{5}$$

We have

$$\dot{x}_j^2(t) \geqq L(t, x_j^1, \dot{x}_j^1),$$

so by Lemma 2 we conclude that $\int_0^1 L(t, x_j^1, \dot{x}_j^1)\, dt$ is defined, possibly as $-\infty$. From (5) and the fact that x_j^3, x_j^4 are constant, we deduce

$$\begin{aligned} l(x_j^1(0), x_j^3(1)) + \int_0^1 L(t, x_j^1, \dot{x}_j^1)\, dt \\ < l(z(0), z(1)) + \int_0^1 L(t, z, \dot{z})\, dt - md(x_j^*(1), C_1). \end{aligned} \tag{6}$$

If we set s_j equal to $x_j^3(1) - x_j^1(1)$, we have

$$|s_j| \leqq 2d(x_j^*(1), C_1) \leqq 2|(x_j^1(1) - z(1), x_j^3(1) - z(1))| < 4/j.$$

Substituting into (6), we arrive at

$$l(x_j^1(0), x_j^1(1) + s_j) + \int_0^1 L(t, x_j^1, \dot{x}_j^1)\, dt < \Phi_\varepsilon^1(0) - |s_j| m/2.$$

Since s_j converges to 0, we deduce from this,

$$\liminf_{s \to 0} [\Phi_\varepsilon^1(s) - \Phi_\varepsilon^1(0)]/|s| < -m/2 = -\sigma - 1,$$

which contradicts the definition of σ. Q.E.D.

The multifunction E is measurable in t and Lipschitz in s^* near z^*, and the sets C_0 and C_1 are closed. These facts, along with the calmness (as defined in [4]) of the problem in Lemma 3, allow us to make use of [4, Thm. 1] to conclude that an arc p^* exists such that

$$(\dot{p}^*(t), p^*(t)) \quad \text{is normal a.e. at } (z^*(t), \dot{z}^*(t)) \tag{7}$$

to the set

$$\{(s^*, v^*) : v^* \in E(t, s^*)\},$$

$$p^*(0) \text{ is normal to } C_0 \text{ at } z^*(0), \tag{8}$$

$$p^2(1) = p^4(1) = -1, \tag{9}$$

$$(-p^1(1), -p^3(1)) \in \partial g(z^1(1), z^1(1)), \tag{10}$$

where g is defined by

$$g(s^1, s^3) = md(s^*, C_1).$$

We deduce from (7) that $\dot{p}^2$, $\dot{p}^3$, $\dot{p}^4$ are 0 a.e., since E depends only on s^1. We derive

$$(\dot{p}^1, p^1, -1) \text{ is normal a.e. at } (z, \dot{z}, L(t, z, \dot{z}))$$

$$\text{to the set epi } L(t, \cdot, \cdot).$$

This gives (3), for $p = p^1$. Because we have $p^3(1)$ equal to $-p^1(1)$ as a consequence of (10), we may conclude from (8) that $(p^1(0), -p^1(1), -1)$ is normal at $(z(0), z(1), l(z(0), z(1)))$ to the set epi l, whence (4). Q.E.D.

In view of the fact that a distance function is Lipschitz with constant 1, relations (10) and (2a) imply the following extra fact which will be used later.

COROLLARY 1. *In Theorem* 1, *we may take p satisfying*

$$|p(1)| \leqq 2(\sigma + 1).$$

We now state two lemmas which will be used later. The first is Theorem 1.21 of [1]; the method of proof is essentially the same as that of [2, Lemma 3.15]. The second is Lemma 6.8 of [1], and may be proved by means of [2, Prop. 3.2].

LEMMA 4. *Let* $f : R^n \to R^m$ *be* C^1 *and* $g : R^n \to R$ *be Lipschitz. Suppose that the point* (α, β, r) *in* $R^n \times R^m \times (-\infty, 0]$ *is normal at* $(x, f(x), g(x))$ *to the set*

$$\{(s, f(s), g(s) + \delta) : s \in R^n, \delta \geqq 0\}.$$

Then

$$\alpha + \beta f'_s(x) \in \partial(-rg)(x).$$

LEMMA 5. *Let E be a closed and convex-valued multifunction from* R^n *to* R^m, *and let* (α, β) *in* $R^n \times R^m$ *be normal at* (s_0, v_o) *to the set*

$$\{(s, v) \in R^n \times R^m : v \in E(s)\}.$$

Then β *is normal at* v_0 *to the set* $E(s_0)$.

4. A Weierstrass-like result. We denote by co$L$ the convexification of the function $L(t, s, v)$ in the v variable. That is, for each t and s, the function

$\mathrm{co}L(t, s, \cdot)$ is the *convex hull* of the function $L(t, s, \cdot)$, the largest convex function majorized by $L(t, s, \cdot)$ (see [9, Cor. 17.1.5]). In general, $\mathrm{co}L$ may have to equal $-\infty$; we always have $\mathrm{co}L \leqq L$.

THEOREM 2. *Under the hypotheses of Theorem* 1, *we may assert the following*:

$$\mathrm{co}L(t, z(t), \dot{z}(t)) = L(t, z(t), \dot{z}(t)) \quad a.e. \tag{11}$$

(12) *the arc z solves, for some positive* δ, *the problem*:

$$\textit{minimize}\,\{l(x(0), x(1)) + \int_0^1 \mathrm{co}L(t, x, \dot{x})\, dt : |x(t) - z(t)| < \delta \quad a.e.\}.$$

Remarks. This is an extension of Theorem 1 of [3], where it was shown that condition (11) is the essence of the necessary condition of Weierstrass in the calculus of variations. Statement (12) is to be interpreted within the convention of § 2. Thus for any arc x within δ of z (as in (12)) for which the integral in (12) is defined, and for which $l(x(0), x(1))$ is finite, the corresponding value of the functional in (12) is no less than its (finite) value at z.

Proof. We apply Lemma 3, § 3, to conclude that z solves the problem given there, where we may assume δ less than ε. This problem falls within the context of [3, Thm. 2], where the W there is in this case the set

$$\{(t, s^*) : |s^1 - z(t)| < \delta, |s^3 - z(1)| < \delta\}.$$

We conclude therefore that z^* continues to solve the problem of Lemma 3 with E replaced by its convex hull. By definition,

$$\mathrm{co}E\,(t, s^*) = \mathrm{epi}\,\mathrm{co}L\,(t, s^1, \cdot) \times \{0\} \times \{0\}.$$

Were it true that $\mathrm{co}L\,(t, z, \dot{z})$ is strictly less than $L(t, z, \dot{z})$ on a set of positive measure, a standard argument (employing [6, Cor. 3.3]) would produce a measurable and integrable function $u(t)$ satisfying

$$\mathrm{co}L(t, z, \dot{z}) \leqq u(t) \leqq L(t, z, \dot{z}),$$

with the second inequality strict on a set of positive measure. Then the arc x^* defined by

$$x^*(t) = \left[z(t), \int_0^t u(r)\, dr, z(1), l(z(0), z(1))\right]$$

would be admissible for the convexified problem of Lemma 3, and in fact strictly better than z^*. This contradiction yields (11).

Now let any arc x within δ of z be given with $l(x(0), x(1))$ finite.

For any integrable function $u(t)$ satisfying

$$\mathrm{co}L\,(t, x, \dot{x}) \leqq u(t),$$

we may define an arc x^* as follows:

$$x^*(t) = \left[x(t), \int_0^1 u(r)\, dr, x(1), l(x(0), x(1))\right].$$

Note that this arc is feasible for the convexified problem of Lemma 3.

The optimality of z^* implies

$$l(x(0), x(1)) + \int_0^1 u(t)\, dt \geqq l(z(0), z(1)) + \int_0^1 L(t, z, \dot{z})\, dt. \tag{13}$$

From Lemma 2 and [9, Cor. 17.1.5] we deduce the measurability of $\mathrm{co}L\ (t, x, \dot{x})$. If $\int_0^1 \mathrm{co}L(t, x, \dot{x})\, dt$ is defined, the value of this integral cannot be $-\infty$, otherwise we could find u as above with $\int_0^1 u(t)\, dt$ arbitrarily close to $-\infty$, making (13) impossible. Thus the value of the integral is either $+\infty$ or a finite number. In the latter case we obtain (13) with $u(t)$ equal to $\mathrm{co}L(t, x, \dot{x})$. Statement (12) of the theorem follows. Q.E.D.

We recall that we are assuming, without loss of generality, the validity of the inequality in Definition 2 for $i = 1$.

COROLLARY 2. *Define $\tilde{\Phi}_\alpha^1$ as in Definition 2 for L replaced by $\mathrm{co}L$, $\alpha = \delta/2$ for the δ of Theorem 2. Then, if m is as defined in Lemma 3,*

$$\liminf_{s \to 0} [\tilde{\Phi}_\alpha^1(s) - \tilde{\Phi}_\alpha^1(0)]/|s| \geqq -m.$$

Hence the relaxed problem is calm at z and its sensitivity $\tilde{\sigma}$ is no greater than $m = 2(\sigma + 1)$.

Proof. Let x be any arc and s any point of R^n such that $l(x(0), x(1)+s)$ is finite, $|x - z| < \alpha$, $|s| < \alpha$, and $\int_0^1 \mathrm{co}L\ (t, x, \dot{x})\, dt$ is defined and less than ∞. The same argument as in the proof of the theorem above shows that the value of the integral cannot be $-\infty$. Hence we may define an arc x^* by

$$x^*(t) = [x(t), \int_0^t \mathrm{co}L\ (r, x, \dot{x})\, dr, x(1) + s, l(x(0), x(1) + s)],$$

an arc feasible for the convexified version of the problem of Lemma 3. Since

$$d(x^*(1), C_1) \leqq |s|,$$

we have, by the optimality of z^*,

$$l(x(0), x(1)+s) + \int_0^1 \mathrm{co}L\ (t, x, \dot{x})\, dt + m|s| \geqq l(z(0), z(1)) + \int_0^1 \mathrm{co}L\ (t, z, \dot{z})\, dt.$$

By Theorem 2, the right-hand side of this last inequality is $\tilde{\Phi}_\alpha^1(0)$. We thus derive, for all s small,

$$\tilde{\Phi}_\alpha^1(s) - \tilde{\Phi}_\alpha^1(0) \geqq -m|s|,$$

whence the corollary. Q.E.D.

5. Optimal control. We consider the following problem of optimal control (see Example 3, § 2): to minimize

$$l(x(0), x(1)) + \int_0^1 g(t, x(t), u(t))\, dt$$

over the couples (x, u) such that x is an arc, u an integrable function such that

$$u(t) \in U(t) \quad \text{a.e.},$$

(where $U : [0, 1] \to R^m$ is a given multifunction) and

$$\dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e.}$$

The calmness of this problem is defined in a manner completely analogous to Definition 2. We shall assume the following:

(14) $f(t, s, u)$ is measurable in t and continuous in u;

(15) $g(t, s, u)$ is l.s.c. in u and satisfies condition (c) of Proposition 7 (§ 2);

(16) $l : R^n \times R^n \to (-\infty, \infty]$ is l.s.c.;

(17) U is measurable and closed-valued.

We shall refer to the above as the (l, f, g, U) *control problem.*

THEOREM 3 (Maximum principle). *Suppose the couple* (z, v) *solves the above problem locally, where*

(18) *the problem is calm at* z,

(19) *for almost all* t, $f(t, \cdot, v(t))$ *is* C^1 *near* $z(t)$,

(20) *for every positive integer* i *there exist a positive* ε *and an integrable function* k *(depending on* i*) such that, for almost all* t, *given two points* s_1 *and* s_2 *within* ε *of* $z(t)$ *and* u *in* $U(t)$ *within* i *of* $v(t)$, *we have*

$$|f(t, s_1, u) - f(t, s_2, u)| \leqq k(t)|s_1 - s_2|,$$

$$|g(t, s_1, u) - g(t, s_2, u)| \leqq k(t)|s_1 - s_2|.$$

Then there exists an arc p *such that*

$$\dot{p}(t) + p(t) f_s'(t, z(t), v(t)) \in \partial_s g(t, z(t), v(t)) \quad \text{a.e.}, \tag{21}$$

$$p(t) \cdot f(t, z(t), v(t)) - g(t, z(t), v(t)) = \max \{p(t) \cdot f(t, z(t), u) - g(t, z(t), u) : u \in U(t)\} \quad \text{a.e.}, \tag{22}$$

$$(p(0), -p(1)) \in \partial l(z(0), z(1)). \tag{23}$$

Remarks. The word "solve" has the same unrestrictive meaning it had in §§ 2 and 4. That the arc z is a local solution means it is optimal relative to other couples (x, u) for which $|x(t) - z(t)|$ is less than some positive number δ. In relation (21), f_s' refers to the usual Jacobian matrix with respect to s, and $\partial_s g$ to the generalized gradient with respect to s. Relation (23), the transversality condition, is susceptible to various interpretations, depending on the nature of l. For example, if l is defined as in Example 3, § 2, (23) says that $(p(0), -p(1))$ is normal to C at $(z(0), z(1))$ (see (2b)).

Although there have been many versions of the maximum principle since Pontryagin's [5], Theorem 3 has something new to offer. The regularity assumptions on f and g are considerably weaker, the function l (and the corresponding

transversality condition) is quite general, and the control set U varies with t. Note also that we do not assume that the optimal control is bounded, as is usually done. This hypothesis is made in the classical case, along with the hypothesis that f'_s and g'_s exist and are continuous in t, s and u. These conditions imply (20). The necessary conditions obtained are "normal"; there is no undetermined constant factor multiplying g in (21) and (22) as is usually the case. This is a consequence of (18). We may thus make the statement "calm problems are normal".

Subsequent to this research, J. Warga [11], [12] has obtained results which resemble Theorem 3 in spirit.

Proof of the theorem. Let r be a given positive integer. We define $U_r(t)$ as follows:

$$U_r(t) = \{u \in U(t) : 1/r \leqq |u - v(t)| \leqq r\} \cup \{v(t)\},$$

and we note that (z, v) solves the control problem (l, f, g, U_r).

Using the notation of Example 3, § 2, we reformulate this problem as a generalized problem of Bolza. The function L is as defined there; we label l^* the function which to a pair of points (s_0, s_0^*) and (s_1, s_1^*) in R^{n+m} assigns the value $l(s_0, s_1)$. It is easy to see (Propositions 3, 7) that the hypotheses of Theorem 2 are present, so that the arc (z, z^*) solves (locally) the new "relaxed" problem in which L is convexified in the variable (v, v^*). Let $\tilde{U}_r(t)$ be the set of points $\tilde{u}$ of the following form:

$$\tilde{u}(t) = (\lambda_1, \lambda_2, \cdots, \lambda_k, u_1, \cdots, u_k), \tag{24}$$

where $k = m + n + 2$, λ_i $(i = 1, 2, \cdots, k)$ are nonnegative numbers whose sum is 1, and u_i $(i = 1, 2, \cdots, k)$ are elements of $U_r(t)$. We define $\tilde{f}$ and $\tilde{g}$, for (t, s) in $[0, 1] \times R^n$ and $\tilde{u}$ in $\tilde{U}_r(t)$ as follows:

$$\tilde{f}(t, s, \tilde{u}) = \sum \lambda_i f(t, s, u_i),$$

$$\tilde{g}(t, s, \tilde{u}) = \sum \lambda_i g(t, s, u_i),$$

and $\tilde{h}(\tilde{u})$ is defined to be $\sum \lambda_i u_i$ (all sums are on i from 1 to k).

It is a consequence of Caratheodory's theorem that the set epi co$L$ $(t, s, s^*, \cdot, \cdot)$ may be expressed as follows:

$$\{(\tilde{f}(t, s, \tilde{u}), \tilde{h}(\tilde{u}), \tilde{g}(t, s, \tilde{u}) + \delta) : \tilde{u} \in \tilde{U}_r(t), \delta \geqq 0\}.$$

Since (z, z^*) solves the relaxed problem, we deduce that the pair $(z, \tilde{v})$ solves the optimal control problem $(l, \tilde{f}, \tilde{g}, \tilde{U}_r)$, where $\tilde{v}(t)$ in $\tilde{U}_r(t)$ is defined by

$$\tilde{v}(t) = (1/k, \cdots, 1/k, v(t), \cdots, v(t)).$$

This in turn implies that a.e. the following set is empty:

$$\{\tilde{u} \in \tilde{U}_r(t) : \tilde{f}(t, z(t), \tilde{u}) = \dot{z}(t), \tilde{g}(t, z(t), \tilde{u}) < \tilde{g}(t, z(t), \tilde{v}(t))\}. \tag{25}$$

Suppose now that the function g is replaced by the function

$$g_r(t, s, u) = g(t, s, u) + |u - v(t)|/r.$$

The hypotheses of the theorem remain in force, including the optimality of (z, ν) for the problem (l, f, g_r, U) and the same argument as above shows that $(z, \tilde{\nu})$ solves the problem $(l, \tilde{f}, \tilde{g}_r, \tilde{U}_r)$. Viewing this as a problem of Bolza, we may apply Theorem 1 to obtain an arc (p, p^*) satisfying (3) and (4). Here, these become

$$(p(0), -p(1)) \in \partial l(z(0), z(1)), \qquad p^*(1) = 0, \tag{26}$$

$(\dot{p}(t), p(t), 0, -1)$ is normal a.e. at the point

$(z(t), \dot{z}(t), \nu(t), \tilde{g}_r(t, z(t), \tilde{\nu}(t)))$ to the set (27)

$$\{(s, \tilde{f}(t, s, \tilde{u}), \tilde{h}(\tilde{u}), \tilde{g}_r(t, s, \tilde{u}) + \delta) : s \in R^n, \tilde{u} \in \tilde{U}_r(t), \delta \geqq 0\}.$$

We now apply Lemma 5 to conclude that $(p(t), 0, -1)$ is normal (in the sense of convex analysis) a.e. at the point $(\dot{z}(t), \nu(t), \tilde{g}_r(t, z(t), \tilde{\nu}(t)))$ to the convex set

$$\{(\tilde{f}(t, z(t), \tilde{u}), \tilde{h}(\tilde{u}), \tilde{g}_r(t, z(t), \tilde{u}) + \delta) : \tilde{u} \in \tilde{U}_r(t), \delta \geqq 0\}.$$

We obtain from this the inequality a.e.

$$p(t) \cdot f(t, z(t), u) - g(t, z(t), u) \leqq p(t) \cdot \dot{z}(t) - g(t, z(t), \nu(t)) + |u - \nu(t)|/r, \tag{28}$$

valid for u in $U_r(t)$.

Suppose now we have the following situation: we have two sequences of points

$$Q_j = (a_j, b_j, c_j, d_j),$$

$$R_j = (s_j, \tilde{f}(s_j, \tilde{u}_j), \tilde{h}(\tilde{u}_j), \tilde{g}_r(s_j, \tilde{u}_j) + \delta_j),$$

for a certain t (suppressed here) for which the set (25) is empty, where $\tilde{u}_j \in \tilde{U}_r(t)$, Q_j and R_j both converge to $(z, \dot{z}, \nu, g(z, \nu))$, $\delta_j \geqq 0$,

$$(Q_j - R_j)/|Q_j - R_j| \to (\alpha, \beta, \gamma, \delta),$$

and Q_j has closest point R_j in the set occurring in (27).

Claim. $\alpha + \beta f'_s(z, \nu) \in \partial_s[-\delta g(z, \nu)]$. Since $\tilde{U}_r$ is compact, we may, if necessary, take a subsequence and assume that $\tilde{u}_j \to \tilde{u}$ in $\tilde{U}_r$. We have $\tilde{f}(z, \tilde{u}) = \dot{z}$ and

$$\tilde{g}_r(z, \tilde{u}) \leqq g(z, \nu).$$

From the fact that the set (25) is empty, we deduce (if $\tilde{u}$ is given by (24))

$$\sum \lambda_i |u_i - \nu(t)| = 0. \tag{29}$$

By the definition of generalized normals we have the vector defined by

$$(\alpha_j, \beta_j, \gamma_j, \delta_j) = (Q_j - R_j)/|Q_j - R_j|$$

normal at R_j to the set

$$\{(s, \tilde{f}(s, \tilde{u}_j), \tilde{h}(\tilde{u}_j), \tilde{g}_r(s, \tilde{u}_j) + \delta) : s \in R^n, \delta \geqq 0\}.$$

From Lemma 4 we conclude:

$$\alpha_j + \beta_j \tilde{f}'_s(s_j, \tilde{u}_j) \in \partial_s[-\delta_j \tilde{g}_r(s_j, \tilde{u}_j)]. \tag{30}$$

The equality (29) implies that either $\lambda_i = 0$ or else, for all j large, $(u_j)_i = \nu$. Taking limits in (30), we consequently obtain the relation claimed (the upper-semicontinuity of the generalized gradient is used here).

The point is that, because of relation (27), $(\dot{p}, p, 0, -1)$ is the limit of convex combinations of points like $(\alpha, \beta, \gamma, \delta)$ above [2, Prop. 3.2]. This allows us to conclude the following:

$$\dot{p}(t)+p(t)f_s'(t, z(t), \nu(t)) \in \partial_s g(t, z(t), \nu(t)) \quad \text{a.e.} \tag{31}$$

The arc p satisfying (26), (28) and (31) actually depends on r, so let us label it p_r. We may assume $|p_r(1)|$ bounded above by $2(\tilde{\sigma}_r + 1)$, where $\tilde{\sigma}_r$ is the sensitivity of the problem $(l, \tilde{f}, \tilde{g}_r, \tilde{U}_r)$ (Corollary 1). In view of Corollary 2, we have $\tilde{\sigma}_r$ bounded above by $2(\sigma_r + 1)$, where σ_r is the sensitivity of the problem (l, f, g_r, U_r). It is easy to see in addition that σ_r is bounded above by σ, the sensitivity of the problem (l, f, g, U). Thus the solutions p_r to (31) are uniformly bounded at 1, and by a theorem of Valadier [10] there exists a subsequence converging uniformly to an absolutely continuous solution p of (21). This p continues to satisfy (23), and satisfies (22). Q.E.D.

Acknowledgment. The author would like to thank Professor R. T. Rockafellar for his assistance and encouragement.

REFERENCES

[1] F. H. CLARKE, *Necessary conditions for nonsmooth problems in optimal control and the calculus of variations*, Ph.D. thesis, University of Washington, Seattle, Wash., 1973.

[2] ———, *Generaiized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[3] ———, *Admissible relaxation in variational and control problems*, J. Math. Anal. Appl., 51 (1975), pp. 557–576.

[4] ———, *Optimal solutions to differential inclusions*, J. Optimization Theory Appl., to appear.

[5] L. S. PONTRYAGIN ET AL., *The Mathematical Theory of Optimal Processes*, Wiley Interscience, New York, 1962.

[6] R. T. ROCKAFELLAR, *Measurable dependence of convex sets*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

[7] ———, *Conjugate convex functions in optimal control and the calculus of variations*, Ibid., 32 (1970), pp. 174–222.

[8] ———, *Existence and duality theorems for convex problem of Bolza*, Trans. Amer. Math. Society, 159 (1971), pp. 1–39.

[9] ———, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[10] M. VALADIER, *Existence globale pour les équations différentielles multivoques*, C.R. Acad. Sci. Paris, 272 (1971), pp. 474–477.

[11] J. WARGA, *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 18 (1975), pp. 41–62.

[12] ———, *Controllability and necessary conditions in unilateral problems without differentiability assumptions*, this Journal, 14 (1976), pp. 546–573.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

GLOBAL CONTROLLABILITY OF NONLINEAR SYSTEMS*

RONALD M. HIRSCHORN†

Abstract. This paper examines the relationship between the structure of the reachable set for nonlinear systems and the properties of the Lie algebras of vector fields associated with nonlinear systems. An expression for the reachable set at time t is obtained for a large class of nonlinear systems using unbounded controls.

1. Introduction. In this paper we study the controllability of nonlinear systems of the form

$$\frac{dx}{dt} = A(x) + u_1 B_1(x) + \cdots + u_m B_m(x)$$

which evolve on a real analytic manifold M. There is a considerable amount of literature which examines some of the basic structural properties of nonlinear systems (see, for example, Brockett [3], [4], Elliot [6], Hermann [9], Haynes and Hermes [7], Hirschorn [11], [12], Lobry [15], Palais [16], Sussmann and Jurdjevic [17], [18]). In contrast with the linear case, the problem of obtaining an expression for the reachable states at time t from some initial state remains unanswered. This question is of practical as well as theoretical interest, and in two special cases, a useful expression for the reachable set at time t is known—these are the symmetric systems, where $A(x) = 0$, and a class of systems where the vector fields $A(x), B(x), \cdots, B_m(x)$ generate a finite-dimensional Lie algebra $\mathscr{L}$. This is the case, for example, with bilinear systems, but the assumption that $\mathscr{L}$ is finite-dimensional is quite restrictive. In particular, the series interconnection of two bilinear systems is a nonlinear system for which $\mathscr{L}$ is infinite-dimensional. The purpose of this paper is to extend these results to a far larger class of systems where $\mathscr{L}$ is possibly infinite-dimensional. The global properties of the state space M are exploited by associating with each nonlinear system a Lie algebra of vector fields $\mathscr{L}$ and transformation group G of the state space. The relationship between the structure of this Lie algebra, the group G and the reachable set is examined, and we obtain a Lie-algebraic criteria for exact time controllability. This global result considerably generalizes the known results.

The organization of this article is as follows: in § 2 we introduce notation and describe some basic results which are used in later sections. Section 3 contains the main theorem and some examples.

2. Preliminaries. We assume that the reader is familiar with the basic notions of differential geometry and Lie theory (cf. [13], [14], [19]).

We shall consider systems of the form

$$(*) \qquad \frac{dx}{dt}(t) = A(x(t)) + \sum_{i=1}^{m} u_i(t) B_i(x(t)), \qquad x(0) = x_0 \in M,$$

* Received by the editors June 16, 1975, and in revised form September 26, 1975.

† Department of Mathematics, Queen's University at Kingston, Kingston, Ontario, Canada K7L 3N6.

where $x \in M$, a real analytic manifold, $A, B_1, \cdots, B_m$ are real analytic vector fields, and the controls $u_i \in \mathcal{P}$, the class of piecewise constant functions from $[0, \infty)$ into R. This last restriction is not necessary—the results remain unchanged if piecewise continuous controls are allowed.

Let $V(M)$ denote the set of real analytic vector fields on M. We regard $V(M)$ as a Lie algebra over the reals. Thus $V(M)$ is a vector space over R with a nonassociative "multiplication" defined as follows: for $X, Y \in V(M)$, the *Lie bracket of X and Y* is

$$[X, Y](m) = X(m)Y - Y(m)X$$

(see [19]). In the case where $M = R^n$, we can consider each vector field X as a mapping from R^n into R^n. Then for each $X \in V(R^n)$ and each real analytic function f on R^n, $X(m)(f) = (df)_m(X(m))$, where $(df)_m$ is the Jacobian of f. In this case,

$$[X, Y](m) = (dY)_m X(m) - (dX)_m Y(m).$$

Let $\mathcal{V}$ be a subset and $\mathcal{L}$ and $\mathcal{H}$ Lie subalgebras of $V(M)$.

$\mathcal{V}(x) = \{Y(x) : Y \in \mathcal{V}\}$;

$\{\mathcal{V}\}_{\mathrm{LS}}$ = linear span of $\mathcal{V}$ in $V(M)$;

$\{\mathcal{V}\}_{\mathrm{LA}}$ = the Lie algebra generated by $\mathcal{V}$, that is, the smallest Lie subalgebra of $V(M)$ containing $\mathcal{V}$;

$[\mathcal{L}, \mathcal{H}] = \{[L, H] : L \in \mathcal{L}, H \in \mathcal{H}\}$.

If S is a subset of a group G, we set

$$\{S\}_G = \text{the subgroup of } G \text{ generated by } S.$$

Let $X, Y \in V(M)$. We define

$$\mathrm{ad}_X^0 Y = Y, \quad \mathrm{ad}_X Y = [X, Y] \quad \text{and} \quad \mathrm{ad}_X^n Y = [X, \mathrm{ad}_X^{n-1} Y].$$

Suppose $X \in V(M)$ is a complete vector field. Then there is a *one-parameter group* X_t of X, i.e., for each $t_1, t_2 \in R$, X_{t_1} and X_{t_2} are diffeomorphisms of M ($X_{t_1}, X_{t_2} \in \mathrm{diff}\,(M)$); $X_{t_1} \circ X_{t_2} = X_{t_1+t_2}$, and for all $m \in M$, $X_0(m) = m$ and

$$\frac{d}{dt} X_t(m) = X(X_t(m)).$$

Assume that the solutions to the differential equation (*) are defined for all $t \geqq 0$. We denote this solution by $\pi(x_0, u, t)$.

If $x, y \in M$, we say that y *is reachable from x at time t* if there exists a $u \in \mathcal{P}$ such that $y = \pi(x, u, t)$. We denote by $\mathcal{R}_t(x)$ *the reachable set from x at time t*, the points in M reachable from x at time t. $\mathcal{R}(x)$ denotes the reachable set from x in positive time, i.e., $\mathcal{R}(x) = \bigcup_{t>0} \mathcal{R}_t(x)$.

It is known that the structure of the reachable set is related to the structure of the Lie algebras:

$$\mathscr{L} = \{A, B_1, \cdots, B_m\}_{\mathrm{LA}},$$
$$\mathscr{L}_0 = \{\mathrm{ad}_A^k B_i : i = 1, \cdots, m \text{ and } k = 0, 1, \cdots\}_{\mathrm{LA}},$$
$$\mathscr{B} = \{B_1, \cdots, B_m\}_{\mathrm{LA}}.$$

We illustrate the construction of these algebras for linear systems of the form

$$\frac{dx}{dt} = Ax + Bu, \tag{1}$$

where $x \in R^n$, A is an $n \times n$ matrix over R, B is an $n \times n$ matrix over R, and $u(t) \in R^m$. Let $B_1, B_2 \cdots, B_m$ denote the columns of B and $u_1, \cdots, u_m$ the coordinates of u. We can rewrite (1) in the form

$$\frac{dx}{dt} = Ax + u_1 B_1 + \cdots + u_m B_m.$$

Thus $A(x) = Ax$, $B_i(x) = B_i$ are linear vector fields. It follows directly from the definitions that $[B_i, B_j] = 0$; $\mathrm{ad}_A B_i(x) = [A, B_i](x) = (dB_i)_x A(x) - (dA)_x B_i(x) = 0 - AB_i(x)$; $\mathrm{ad}_A^k B_i(x) = (-1)^k A^k B_i(x)$; $[\mathrm{ad}_A^k B_i, \mathrm{ad}_A^l B_j] = 0$, and so

$$\mathscr{L} = \{A, A^k B_i : k = 0, 1, \cdots, n-1; i = 1, \cdots, m\}_{\mathrm{LS}},$$
$$\mathscr{L}_0 = \{A^k B_i : k = 0, 1, \cdots, n-1; i = 1, \cdots, m\}_{\mathrm{LS}},$$
$$\mathscr{B} = \{B_1, \cdots, B_m\}_{\mathrm{LS}}.$$

Here $\mathscr{L}$ consists of complete vector fields.

Throughout the rest of this article, we shall assume, as in [18], that $\mathscr{L}$ consists of complete vector fields. This is the case, for example, when M is compact. This assumption is not essential but it considerably simplifies many of the proofs.

Let $\mathscr{H}$ be a lie subalgebra of $V(M)$. For each $x \in M$, $I(\mathscr{H}, x)$ will denote the *maximal integral manifold of $\mathscr{H}$ through x*; i.e., $I(\mathscr{H}, x)$ is the largest connected submanifold N of M which contains x and has the property that for all $y \in N$, the tangent space to N at y is $\mathscr{H}(y)$. Its existence follows from Lobry's global version of Frobenius' theorem [15].

Suppose that $\mathscr{D}$ is a subset of $V(M)$. An integral curve of $\mathscr{D}$ is piecewise smooth curve α with $d/dt(\alpha(t)) \in \mathscr{D}(\alpha(t))$ for each t where the derivative is defined. We will call $\mathscr{D}$ *symmetric* if for each vector field $X \in \mathscr{D}$, $-X \in \mathscr{D}$.

THEOREM 2.1. *Let $\mathscr{D} \subset V(M)$ be symmetric and let $x \in M$. Then for every $y \in I(\mathscr{D}, x)$, there is an integral curve $\alpha : [0, T] \to M$ of $\mathscr{D}$ with $\alpha(0) = x$ and $\alpha(T) = y$. In particular, every point in $I(\{\mathscr{D}\}_{\mathrm{LA}}, x)$ can be reached from x along an integral curve of $\mathscr{D}$* (cf. [9], [18]).

LEMMA 2.2 (Sussmann and Jurdjevic [18]). *Let $\{X_t\}$ and $\{Y_t\}$ denote the one-parameter groups of vector fields X and $Y \in \mathscr{L}$, where $\mathscr{L}$, $\mathscr{L}_0$ and $\mathscr{B}$ are the Lie algebras associated with a nonlinear system* (*). *Then for all $x \in M$ and $t \in R$, $X_t(I(\mathscr{L}_0, x)) = Y_t(I(\mathscr{L}_0, x))$. In particular, $A_t(I(\mathscr{L}_0, x))$ is the unique maximal integral manifold for $\mathscr{L}_0$ through $A_t(x)$.*

This result motivates the following definition:

DEFINITION. Let $\mathcal{H}$ be a Lie subalgebra of $\mathcal{L}$. Then for all $t>0$, we set $I^t(\mathcal{H}, x) = I(\mathcal{H}, A_t(x))$.

THEOREM 2.3 (Sussmann and Jurdjevic [18]). *Consider the nonlinear system* (*). *For all* $x \in M$ *and* $t>0$, $\mathcal{R}_t(x) \subset I^t(\mathcal{L}_0, x)$, *and with respect to the topology of* $I^t(\mathcal{L}_0, x)$, $\mathcal{R}_t(x)$ *is contained in the closure of its interior.*

We conclude this section with a version of the Campbell–Hausdorff formula:

Let $X, Y \in V(M)$. Then

$$X_{t_1} \circ Y_{t_2}(x) = Z_1(x), \quad \text{where} \quad Z = t_1 X + t_2 Y + \frac{t_1 t_2}{2}[X, Y] + \cdots$$

is a formal series which converges for t_1 and t_2 both in some neighborhood of 0.

3. Controllability of nonlinear systems. As noted in §2, $\mathcal{R}_t(x_0)$ has a nonempty interior in $I^t(\mathcal{L}_0, x_0)$. This result follows from the decomposition $\mathcal{L} \supset \mathcal{L}_0 \supset \mathcal{B}$ of $\mathcal{L}$. To determine those nonlinear systems for which

$$\mathcal{R}_t(x_0) = I^t(\mathcal{L}_0, x_0),$$

it is necessary to examine the structure of $\mathcal{L}$ in greater detail. We will construct a new Lie subalgebra of $\mathcal{L}$, which we call the A-radical of $\mathcal{B}$, $\mathcal{R}(A; \mathcal{B})$, where $\mathcal{L} \supset \mathcal{L}_o \supset \mathcal{R}(A; \mathcal{B}) \supset \mathcal{B}$, and prove the following theorems:

THEOREM 3.1. *Consider the nonlinear system* (*), *where*

$$\frac{dx}{dt}(t) = A(x(t)) + \sum_{i=1}^{m} u_i(t) B_i(x(t))$$

with associated Lie algebras of vector fields $\mathcal{L}, \mathcal{L}_0, \mathcal{R}(A; \mathcal{B})$, *and* $\mathcal{B}$. *Then* cl $\mathcal{R}_t(x) \supset I^t(\mathcal{R}(A; \mathcal{B}), x)$ *for all* $x \in M$ *and* $t>0$. (cl = closure in $I^t(\mathcal{L}_0, x)$.)

THEOREM 3.2. *Consider the nonlinear system* (*) *with associated Lie algebras* $\mathcal{L}, \mathcal{L}_0, \mathcal{R}(A; \mathcal{B})$ *and* $\mathcal{B}$. *If* $\mathcal{L}_0(x) = \mathcal{R}(A; \mathcal{B})x$ *for all* x *in* M, *then*

$$\mathcal{R}_t(x) = I^t(\mathcal{L}_0, x)$$

for all $x \in M$, $t>0$.

COROLLARY 1. *Consider the system* (*) *with associated Lie algebras* $\mathcal{L}, \mathcal{L}_0, \mathcal{R}(A; \mathcal{B})$ *and* $\mathcal{B}$. *If* $\mathcal{L}_0 = \mathcal{R}(A; \mathcal{B})$, *then*

$$\mathcal{R}_t(x) = I^t(\mathcal{L}_0, x)$$

for all $x \in M$, $t>0$.

COROLLARY 2. *Suppose* $\mathcal{R}(A; \mathcal{B})(x_0) = \mathcal{L}_0(x_0)$ *for some* x_0 *in* M. *Then* $\mathcal{R}_t(x_0)$ *and the boundary of* cl $\mathcal{R}_t(x_0)$ *are disjoint for all* $t>0$. (cl = closure with respect to the topology of $I^t(\mathcal{L}_0, x_0)$.) *In particular,* $\mathcal{R}_t(x_0)$ *is a closed subset of* $I^t(\mathcal{L}_0, x_0)$ *iff* $\mathcal{R}_t(x_0) = I^t(\mathcal{L}_0, x_0)$.

Remarks. 1. Let $y \in M$. If $\mathcal{R}(A; \mathcal{B})(x) = \mathcal{L}_0(x)$ for all x in $I(\mathcal{L}, y)$, then it follows directly from Theorem 3.2 that $\mathcal{R}_t(x) = I^t(\mathcal{L}_0, x)$ for all $x \in I(\mathcal{L}, y)$. It is not known if this condition is necessary.

2. One can verify that the above theorems holds for systems of the form $dx/dt(t) = A(x(t)) + \sum_{i=1}^{m} f_i(u_i(t)) B_i(x(t))$, where f_i are surjective functions from R onto R.

3. The problem of obtaining an expression for $\mathcal{R}_t(x)$ has been studied in [11] under the assumption that $\mathcal{L}$ is finite-dimensional, and a "finite-dimensional" version of Corollary 2 is proved. The assumption that $\mathcal{L}$ is finite-dimensional greatly simplifies the situation. The system can be reformulated as evolving on a Lie group and described by right-invariant vector fields [11].

The remainder of this section will be devoted to proving this result and presenting some examples. Let $\mathcal{H}$ be a Lie subalgebra of $V(M)$ which consists of complete vector fields. The group of diffeomorphisms of M,

$$G(\mathcal{H}) = \{X_t^\alpha : X^\alpha \in \mathcal{H}, t \in R\}_G,$$

is called the *transformation group generated by* $\mathcal{H}$, and for $x \in M$,

$$G(\mathcal{H}) \cdot x = \{g(x) : g \in G(\mathcal{H})\}$$

is the *orbit of* x *under* $G(\mathcal{H})$.

LEMMA 3.3. *Suppose that* $\mathcal{H}$ *is a Lie subalgebra of* $V(M)$ *which consists of complete vector fields. Then for all* $x \in M$, $G(\mathcal{H}) \cdot x = I(\mathcal{H}, x)$.

Proof. $\mathcal{H} \subset V(M)$ is symmetric. Theorem 2.1 asserts that $G(\mathcal{H}) \cdot x = I(\{\mathcal{H}\}_{\mathrm{LA}}, x) = I(\mathcal{H}, x)$ for each $x \in M$, which proves the lemma.

Remark. If $\mathcal{H}$ contains a vector field X which is not complete, one can associate with X a local one-parameter group $\{X_t\}$ [19]. One can then form a pseudo-group $G(\mathcal{H})$ analogous to the group defined above and $G(\mathcal{H}) \cdot x = I(\mathcal{H}, x)$ as before. In this manner, the results of this section can be carried over to the case where $\mathcal{L}$ contains incomplete vector fields.

Thus $I(\mathcal{L}, x) = G(\mathcal{L}) \cdot x$, $I(\mathcal{L}_0, x) = G(\mathcal{L}_0) \cdot x$, and $\mathcal{R}_t(x) \subset G(\mathcal{L}_0) \cdot A_t(x) = I^t(\mathcal{L}_0, x)$. Clearly the action of $G(\mathcal{L})$ on M is related to the Lie algebraic properties of $\mathcal{L}$. Our aim is to relate the properties of $\mathcal{L}$ to the structure of $\mathcal{R}_t(x_0)$, so it is natural to treat $\mathcal{R}_t(x_0)$ as the orbit of x under some subset of $G(\mathcal{L})$. Since we allow only piecewise constant controls, the appropriate subset of $G(\mathcal{L})$ is

$$S_t = \{(A + c_{11}B_1 + \cdots + c_{1m}B_m)_{t_1} \cdots (A + c_{k1}B_1 + \cdots + c_{km}B_m)_{t_k} : c_{ij} \in R;$$
$$t_1, \cdots, t_k > 0, t_1 + \cdots + t_k = t\}.$$

By definition, $S_t \cdot x = \mathcal{R}_t(x)$ for all $x \in M$, $t > 0$, but the proofs are simplified considerably if we consider instead the subset

$$G_t = \{B^0_{\alpha_0}(A_{t_1}B^1_{\alpha_1}A_{-t_1})(A_{t_1+t_2}B^2_{\alpha_2}A_{-t_1-t_2}) \cdots (A_{t_1+\cdots+t_n}B^n_{\alpha_n}A_{-t_1-\cdots-t_n})A_t,$$
$$B^i \in \mathcal{B}, t_i \geqq 0, \textstyle\sum t_i \leqq t, \alpha_i \in R\}$$

of $G(\mathcal{L})$. The following lemma relates G_t and $\mathcal{R}_t(x)$.

LEMMA 3.4. *Consider the system* (*) *with associated Lie algebras* $\mathcal{L}$, $\mathcal{L}_0$, $\mathcal{B}$, *and let* cl $\mathcal{R}_t(x)$ *denote the closure of the reachable set for* (*) *with respect to the topology of* $I^t(\mathcal{L}_0, x)$. *Then for all* $x \in M$, $t > 0$,

$$\text{cl } \mathcal{R}_t(x) = \text{cl } G_t \cdot x.$$

In particular, cl $\mathcal{R}_t(x)$ *contains* $I^t(\mathcal{B}, x)$.

Proof. We begin by showing that cl $R_t(x)$ contains $I^t(\mathcal{B}, x)$ for all $t>0$. By definition, $\mathcal{R}_t(x)$ contains points of the form

$$\left(A+\frac{c_1}{t_1}B_{k_1}\right)_{t_1}\cdot\left(A+\frac{c_2}{t_2}B_{k_2}\right)_{t_2}\cdots\left(A+\frac{c_n}{t_n}B_{k_n}\right)_{t_n}A_{t-t_1-\cdots-t_n}(x),$$

where $k_i\in\{1,2,\cdots,m\}$, $c_i\in R$, $t_i>0$ and $t_1+\cdots+t_n<t$. Taking the limit as $t_1, t_2, \cdots, t_n\to 0$, we see that it follows from the Campbell–Hausdorff formula that cl $\mathcal{R}_t(x)$ contains points of the form

$$\{(B_{k_1})_{c_1}(B_{k_2})_{c_2}\cdots(B_{k_n})_{c_n}\cdot A_t(x): c_i\in R,\ k_i\in\{1,\cdots,m\}\}.$$

Theorem 2.1 assets that this set is $I(\mathcal{B}, A_t(x))=I^t(\mathcal{B}, x)$.

To prove that cl $\mathcal{R}_t(x)\supset$ cl $G_t\cdot x$, it suffices to show that cl $\mathcal{R}_t(x)\supset G_t\cdot x$. Let $y\in G_t\cdot x$. Then

$$\begin{aligned}y&=g_0(A_{t_1}g_1A_{-t_1})\cdots(A_{t_1+\cdots+t_n}g_nA_{-t_1-\cdots-t_n})\cdot A_t(x)\\&=(g_0A_{t_1})(g_1A_{t_2})(g_2A_{t_3})\cdots(g_{n-1}A_{t_n})(g_nA_{t-t_1-\cdots-t_n})\cdot x,\end{aligned}$$

where $g_i\in G(\mathcal{B})$. Since cl $\mathcal{R}_t(x)\supset I^t(\mathcal{B}, x)=G(\mathcal{B})\cdot A_t(x)$ for all $t>0$, we see that $(g_nA_{t-t_1-\cdots-t_n})(x)\in$ cl $\mathcal{R}_{t-t_1-\cdots-t_n}(x)$, $(g_{n-1}A_{t_n})(g_nA_{t-t_1-\cdots-t_n})(x)\in$ cl $\mathcal{R}_{t-t_1-\cdots-t_{n-1}}(x)$, and after n steps we have $y\in$ cl $\mathcal{R}_t(x)$. Thus cl $\mathcal{R}_t(x)\supset$ cl $G_t\cdot x$.

To complete the proof, we need to show that cl $G_t\cdot x\supset\mathcal{R}_t(x)$. By definition, $G_t\cdot x$ contains points of the form $p_n(x)=(B_{t/n}A_{t/n})^n(x)=(B_{t/n}A_{t/n})(B_{t/n}A_{t/n})\cdots(B_{t/n}A_{t/n})(x)$, where $B\in\mathcal{B}$. As $n\to\infty$, $p_n(x)\to(A+B)_t(x)$ as a consequence of the Campbell–Hausdorff formula. Thus cl $G_t\cdot x$ contains all the elements of $S_t\cdot x$, and $S_t\cdot x=\mathcal{R}_t(x)$. This completes the proof.

Thus it suffices to study the action of G_t on M and try to isolate those algebraic properties of $\mathcal{L}$ which result in $G_t\cdot x=I^t(\mathcal{L}_0, x)$ for $t>0$ and x in M. This is the case for linear systems, where a direct computation shows that $I^t(\mathcal{L}_0, x)=e^{At}x+\text{Range}\,(B, AB, \cdots, A^{n-1}B)$. The Lie algebra $\mathcal{L}$ associated with linear systems was described in § 2, and for this special case, $\mathcal{L}_0$ is Abelian and hence $\mathcal{B}$ is an ideal in $\mathcal{L}_0$. That is, for all $B\in\mathcal{B}$, $L\in\mathcal{L}_0$, $[B, L]\in\mathcal{L}_0$. Thus one is led to the natural conjecture: if (*) is a nonlinear system and $\mathcal{B}$ is an ideal in $\mathcal{L}_0$, then $\mathcal{R}_t(x)=I^t(\mathcal{L}_0, x)$. The proof of this result requires the following lemma:

LEMMA 3.5. *Suppose that $\mathcal{H}$ is a Lie subalgebra of $\mathcal{L}_0$, $\mathcal{H}_0=\{\text{ad}_A^kH: H\in\mathcal{H}, k=0,1,\cdots\}_{\text{LA}}$ and G^t is the group of diffeomorphisms $G^t=\{A_vH_uA_{-v}: 0\leqq v<t, u\in R, H\in\mathcal{H}\}_G$. Then $G^t\cdot x=I(\mathcal{H}_0, x)$ for all $x\in M$ and $t>0$.*

Proof. We begin by showing that $G^t\cdot x\subset I(\mathcal{H}_0, x)$. Since G^t is generated by diffeomorphisms of the form $A_vH_uA_{-v}$, it suffices to show that $A_vH_uA_{-v}\cdot y\in I(\mathcal{H}_0, x)$ for all $u, v\in R$, $x\in M$, $H\in\mathcal{H}$ and $y\in I(\mathcal{H}_0, x)$. Lemma 3.3 asserts that $I(\mathcal{H}_0, x)=G(\mathcal{H}_0)\cdot x\supset G(\mathcal{H})\cdot x=I(\mathcal{H}, x)$. Thus $H_uA_{-v}\cdot y\in G(\mathcal{H})\cdot A_{-v}(y)\subset I(\mathcal{H}_0, A_{-v}(y))$ and $A_vH_uA_{-v}\cdot y\in A_v\cdot I(\mathcal{H}_0, A_{-v}(y))=I(\mathcal{H}_0, y)$ by Lemma 2.2. Since $y\in I(\mathcal{H}_0, x)$, $I(\mathcal{H}_0, x)=I(\mathcal{H}_0, y)$.

We will complete the proof by showing that $G^t\cdot x\supset I(\mathcal{H}_0, x)$.Fix $x\in M$, $t>0$. By definition, G^t contains one-parameter groups of the form $t\to A_vH_t^\alpha A_{-v}$ where $H^\alpha\in\mathcal{H}$, $0\leqq v<t$. Let $X^{\alpha,v}$ denote the vector fields on M induced by those one-parameter groups and let $\hat{\mathcal{H}}_0$ denote the Lie algebra generated by these

vector fields. Since $G^t \cdot x \supset I(\hat{\mathscr{H}}_0, x)$, it suffices to show that $I(\hat{\mathscr{H}}_0, x) \supset I(\mathscr{H}_0, x)$. Now corresponding to each $H^\alpha \in \mathscr{H}$, $0 \leqq v < t$, is the vector field $X^{\alpha,v} \in \hat{\mathscr{H}}_0$, and thus $(X^{\alpha,v}(x) - X^{\alpha,0}(x))/v \in \hat{\mathscr{H}}_0(x)$. Taking a Taylor series expansion for $X^{\alpha,v}(x)$, we find that

$$\begin{aligned}
&\lim_{v\to 0} (X^{\alpha,v}(x) - X^{\alpha,0}(x))/v \\
&\quad = \lim_{v\to 0} \left(\sum_{i=0}^{\infty} \frac{(-v)^i}{i!} \operatorname{ad}_A^i H^\alpha(x) \right) - H^\alpha(x)/v \\
&\quad = -[A, H^\alpha](x) \in \hat{\mathscr{H}}_0(x).
\end{aligned}$$

Repeating this procedure, we see that $\hat{\mathscr{H}}_0(x) \supset E(x)$, where $E = \{\operatorname{ad}_A^k H : H \in \mathscr{H}, k = 0, 1, \cdots\}$. Theorem 2.1 implies that $I(\mathscr{H}_0, x)$ is the set of points which can be joined to x by integral curves of E. We have shown that for all $y \in M$, $\hat{\mathscr{H}}_0(y) \supset E(y)$. Thus $I(\hat{\mathscr{H}}_0, x) \supset I(\mathscr{H}_0, x)$, and the proof is complete.

THEOREM 3.6. *Consider the nonlinear system* (*) *with associated Lie algebras* $\mathscr{L} \supset \mathscr{L}_0 \supset \mathscr{B}$. *If* $[\mathscr{L}_0, \mathscr{B}](x) \subset \mathscr{B}(x)$ *for all* $x \in M$, *then* $\mathscr{R}_t(x) = I^t(\mathscr{L}_0, x)$ *for all* $x \in M$ *and* $t > 0$. *In particular, if* $\mathscr{B}$ *is an ideal in* $\mathscr{L}_0$, *then* $\mathscr{R}_t(x) = I^t(\mathscr{L}_0, x)$ *for* $x \in M$, $t > 0$.

Proof. Suppose that $\mathscr{R}_t(x)$ is a dense subset of the manifold $I^t(\mathscr{L}_0, x)$ for each $t > 0$.

Claim. $\mathscr{R}_t(x) = I^t(\mathscr{L}_0, x)$ for each $t > 0$: let $y \in I^t(\mathscr{L}_0, x)$. The reachable set at time $\frac{1}{2}t$ from y for the system $\dot{x} = -A(x) - \sum u_i B_i(x)$ has a nonempty interior $\mathscr{U}$ in $I^{t/2}(\mathscr{L}_0, x)$ from Theorem 2.3. Since $\mathscr{R}_{t/2}(x)$ is dense in $I^{t/2}(\mathscr{L}_0, x)$, there is a point p common to $\mathscr{R}_{t/2}(x)$ and $\mathscr{U}$. The control which took the negative system from y to p takes the original system from p to y, hence $y \in \mathscr{R}_t(x)$.

We now prove that $\mathscr{R}_t(x)$ is a dense subset of $I^t(\mathscr{L}_0, x)$ for all $t > 0$. Lemma 3.4 shows that $\operatorname{cl} \mathscr{R}_t(x) = \operatorname{cl} G_t \cdot x$, so it suffices to show that $G_t \cdot x = I^t(\mathscr{L}_0, x)$, where

$$\begin{aligned}
G_t = \{ & B_{\alpha_0}^D (A_{t_1} B_{\alpha_1}^1 A_{-t_1}) \cdots (A_{t_1+\cdots+t_n} B_{\alpha_n}^n A_{-t_1-\cdots-t_n}) A_t : \\
& B^i \in \mathscr{B},\ t_i \geqq 0,\ \textstyle\sum t_i \leqq t,\ \alpha_i \in R \}.
\end{aligned}$$

If we let $G^t = \{A_v B_u A_{-v} : 0 \leqq v < t,\ u \in R,\ B \in \mathscr{B}\}_G$, then $G^t \cdot A_t(x) = I(\mathscr{L}_0, A_t(x)) = I^t(\mathscr{L}_0, x)$ by Lemma 3.5. Thus $G_t \cdot x = I(\mathscr{L}_0, A_t(x))$ if $G_t \cdot x = G^t \cdot A_t(x)$, and this is the case if for each y in M; $B^1, B^2 \in \mathscr{B}$; $t_1, t_2, \alpha_1, \alpha_2 \in R$,

$$(A_{t_1+t_2} B_{\alpha_2}^2 A_{-t_1-t_2})(A_{t_1} B_{\alpha_1}^1 A_{-t_1}) \cdot y = (A_t b A_{-t_1})(A_{t_1+t_2} B_{\alpha_2}^2 A_{-t_1-t_2}) \cdot y,$$

where b is a diffeomorphism of M, with $b(z) \in I(\mathscr{B}, z)$ for each $z \in M$. This has the effect of allowing the elements in G_t to appear in any order.

We set $p = (A_{t_2} B_{\alpha_2}^2 A_{-t_2}) B_{\alpha_1}^1 (A_{t_2} B_{-\alpha_2}^2 A_{-t_2})$ and observe that

$$\begin{aligned}
&(A_{t_1+t_2} B_{\alpha_2}^2 A_{-t_1-t_2})(A_{t_1} B_{\alpha_1}^1 A_{-t_1}) \cdot y \\
&\quad = A_{t_1} (A_{t_2} B_{\alpha_2}^2 A_{-t_2}) B_{\alpha_1}^1 (A_{t_2} B_{-\alpha_2}^2 A_{-t_2}) (A_{t_2} B_{\alpha_2}^2 A_{-t_2}) A_{-t_1} \cdot y \\
&\quad = A_{t_1} p (A_{t_2} B_{\alpha_2}^2 A_{-t_2}) A_{-t_1} \cdot y \\
&\quad = (A_{t_1} p A_{-t_1})(A_{t_1+t_2} B_{\alpha_2}^2 A_{-t_1-t_2}) \cdot y.
\end{aligned}$$

This implies that the proof will be complete if we show that $p \cdot x \in I(\mathcal{B}, x)$ for all $x \in M$. The curve $\alpha \to X^t_\alpha = A_t B^2_\alpha A_{-t}$ is an integral curve for a vector field X^t, and $p \cdot x = X^{t_2}_{\alpha_2} \cdot B^1_{\alpha_1} \cdot X^{t_2}_{-\alpha_2} \cdot x$. Let $Y^{\alpha,t}$ denote the vector field with integral curve $C^{\alpha,t} : \tau \to X^t_\alpha B^1_\tau X^t_{-\alpha} \cdot x$. Lemma 2.2 implies that the curve $C^{\alpha,t}$ lies entirely in $I(\mathcal{L}_0, x)$, or equivalently, $Y^{\alpha,t}(x) \in \mathcal{L}_0(x)$ for all $x \in M$. We will now show that $Y^{\alpha,t}(x) \in \mathcal{B}(x)$ for all $x \in M$, $t, \alpha \in R$, which implies that $C^{\alpha,t}$ lies in $I(\mathcal{B}, x)$ and hence $p \cdot x = Y^{\alpha_2,t_2}_{\alpha_1} \cdot x \in I(\mathcal{B}, x)$.

Since A, B^1, B^2 are real analytic vector fields, we can express $Y^{\alpha,t}$ by its Taylor expansion in a neighborhood of $\alpha = 0$ where

$$Y^{\alpha,t} = \left(\sum_{i=0}^{\infty} \frac{\alpha^i}{i!} \operatorname{ad}^i_{X^t} B^1 \right)(x)$$

for all α in this neighborhood. Similarly, for each $x \in M$ there is a neighborhood of $t = 0$ such that

$$X^t(x) = \left(\sum_{j=0}^{\infty} \frac{t^j}{j!} \operatorname{ad}^j_A B^2 \right)(x).$$

Using the fact that $A_t \cdot x$ is jointly analytic in x and t [5], we conclude that for any $x \in M$ there is a neighborhood $\mathcal{N}_x$ of $(0, 0)$ in R^2 such that for all $(t, \alpha) \in \mathcal{N}_x$, the above series representations are valid. Combining these two expressions lets us express $Y^{\alpha,t}(x)$ as a sum of vectors of the form

$$\begin{aligned} V &= [\operatorname{ad}^{k_1}_A B^2, [\operatorname{ad}^{k_2}_A B^2, \cdots, [\operatorname{ad}^{k_n}_A B^2, B^1] \cdots](x) \\ &= [L_1[L_2, [\cdots [L_n, B^1] \cdots], \quad \text{where } L_1, \cdots, L_n \in \mathcal{L}_0. \end{aligned}$$

Assuming $[\mathcal{L}_0, \mathcal{B}](y) \subset \mathcal{B}(y)$ for all $y \in M$, we see that $[L_n, B^1]$ is a vector field in $V(M)$ with $[L_n, B^1](y) \in \mathcal{B}(y)$ for all $y \in M$. A straightforward computation shows that if $B(y) \in \mathcal{B}(y)$ for all $y \in M$ and $L_0 \in \mathcal{L}_0$, then $[L_0, B](y) \in \mathcal{B}(y)$ for all $y \in M$. Thus $[L_{n-1}, [L_n, B^1]](x) \in \mathcal{B}(x)$, and repeating this argument, we find that $[L_1, [\cdots [L_n, B^1] \cdots](x) \in \mathcal{B}(x)$. This means that $Y^{\alpha,t}(x) \in \mathcal{B}(x)$ for $(t, \alpha) \in \mathcal{N}_x$. Since this analytic function is contained in $\mathcal{B}(x)$ for $(t, \alpha) \in \mathcal{N}_x$, it follows from analyticity that $Y^{\alpha,t}(x) \in \mathcal{B}(x)$ for all α, t. This means that $p \cdot x \in I(\mathcal{B}, x)$, which proves the main part of the theorem. If β is an ideal in $\mathcal{L}_0$, then $[\mathcal{B}, \mathcal{L}_0] \subset \mathcal{B}$, so $[\mathcal{B}, \mathcal{L}_0](x) \subset \mathcal{B}(x)$. This completes the proof.

This result reveals a basic link between the structure of $\mathcal{L}$ and the structure of the reachable set. In general, $[\mathcal{L}_0, \mathcal{B}](x)$ is not a subset of $\mathcal{B}(x)$ for all $x \in M$, but some Lie subalgebra $\hat{\mathcal{B}}$ of $\mathcal{B}$ may give rise to a "subsystem" with this property; i.e.,

$$\frac{dx}{dt} = A(x) + u_1 \hat{B}_1(x) + \cdots + u_p \hat{B}_p(x),$$

where $\{\hat{B}_1, \cdots, \hat{B}_p\} \subset \mathcal{B}$, generates a Lie subalgebra $\hat{\mathcal{B}}$ of $\mathcal{B}$ and $\hat{\mathcal{B}}$ is an ideal in $\hat{\mathcal{L}}_0 = \{\operatorname{ad}^k_A \hat{B} : \hat{B} \in \hat{\mathcal{B}}, k = 0, 1, \cdots\}_{\mathrm{LA}}$, or more generally, the distribution corresponding to $[\hat{\mathcal{L}}_0, \hat{\mathcal{B}}]$ is contained in the distribution of $\hat{\mathcal{B}}$. Theorem 3.6 asserts that the reachable set at time t for this "subsystem" is $\hat{\mathcal{R}}_t(x) = I^t(\hat{\mathcal{L}}_0, x)$. This implies that $\operatorname{cl} \mathcal{R}_t(x) \supset I^t(\hat{\mathcal{L}}_0, x)$, and Lemma 3.7 shows that $\operatorname{cl} \mathcal{R}_t(x) \supset I^t(\{\hat{\mathcal{L}}_0, \mathcal{B}\}_{\mathrm{LA}}, x)$. Before going on, we formalize the above constructions.

Let $\mathscr{B}_0$ be a Lie subalgebra of $\mathscr{L}_0$ and suppose $\hat{\mathscr{B}}_0$ is a Lie subalgebra of $\mathscr{B}_0$ with the property that $\hat{\mathscr{B}}_0(x) \supset [\hat{\mathscr{L}}_0, \hat{\mathscr{B}}_0](x)$ for all $x \in M$, where $\hat{\mathscr{L}}_0 = \{\mathrm{ad}_A^k B : B \in \hat{\mathscr{B}}_0, k = 0, 1, \cdots\}_{\mathrm{LA}}$. In this case, we say that $\mathscr{B}_1 = \{\mathscr{B}_0, \hat{\mathscr{L}}_0\}$ is *A-generated from* $\mathscr{B}_0$. For example, for those systems with $\mathscr{B}$ an ideal in $\mathscr{L}_0$, set $\hat{\mathscr{B}}_0 = \mathscr{B}_0 = \mathscr{B}$. Then $\hat{\mathscr{B}}_0$ is an ideal in $\hat{\mathscr{L}}_0 = \{\mathrm{ad}_A^k B : B \in \mathscr{B}, k = 0, 1, \cdots\}_{\mathrm{LA}} = \mathscr{L}_0$, and so $\mathscr{L}_0 = \{\mathscr{B}, \mathscr{L}_0\}_{\mathrm{LA}}$ is A-generated from $\mathscr{B}$. The following lemma shows that if $\mathscr{B}_1$ is A-generated from $\mathscr{B}$, then

$$\mathrm{cl}\, \mathscr{R}_t(x) \supset I^t(\mathscr{B}_1, x).$$

LEMMA 3.7. *Consider the system* (*) *with associated Lie algebras* $\mathscr{L}, \mathscr{L}_0, \mathscr{B}$. *Let* $\mathscr{B}_1$ *be a Lie subalgebra of* $\mathscr{L}_0$ *which is A-generated from* $\mathscr{B}$. *Then for all* $t > 0, x \in M$,

$$\mathrm{cl}\, \mathscr{R}_t(x) \supset I^t(\mathscr{B}_1, x).$$

Proof. Since $\mathscr{B}_1$ is A-generated from $\mathscr{B}$, it follows that for some Lie subalgebra $\hat{\mathscr{B}}$ of $\mathscr{B}$, $\hat{\mathscr{B}}(x) \supset [\hat{\mathscr{L}}_0, \hat{\mathscr{B}}](x)$ for all x in M where $\hat{\mathscr{L}}_0 = \{\mathrm{ad}_A^k B : B \in \hat{\mathscr{B}}, k = 0, 1, \cdots\}_{\mathrm{LA}}$ and $\mathscr{B}_1 = \{\mathscr{B}, \hat{\mathscr{L}}_0\}_{\mathrm{LA}}$. Replacing $\mathscr{B}$ by $\hat{\mathscr{B}}$ in the proof of Theorem 3.6 yields the result that $\mathrm{cl}\, \mathscr{R}_t(x) \supset I^t(\hat{\mathscr{L}}_0, x) = I(\hat{\mathscr{L}}_0, A_t \cdot x)$ for all $x \in M$ and $t > 0$. We know that $\mathrm{cl}\, \mathscr{R}_t(x) \supset I^t(\mathscr{B}, x) = I(\mathscr{B}, A_t \cdot x)$ from Lemma 3.4. Thus $\mathrm{cl}\, \mathscr{R}_t(x)$ contains the set $S(x)$, where

$$S(x) = \{X_{\alpha_1}^1 X_{\alpha_2}^2 \cdots X_{\alpha_n}^n (A_t \cdot x) : X^i \in \hat{\mathscr{L}}_0 \text{ or } X^i \in \mathscr{B}, \alpha_i \in R\}.$$

Theorem 2.1 states that $S(x) = I(\{\hat{\mathscr{L}}_0, \mathscr{B}\}_{\mathrm{LA}}, A_t \cdot x) = I(\mathscr{B}_1, A_t \cdot x) = I^t(\mathscr{B}_1, x)$, which completes the proof.

Now suppose $\mathscr{B}_1$ is A-generated from $\mathscr{B}$, so that $\mathrm{cl}\, \mathscr{R}_t(x) \supset I^t(\mathscr{B}_1, x)$, and let $\mathscr{B}_2$ be A-generated from $\mathscr{B}_1$, so $\mathscr{B} \subset \mathscr{B}_1 \subset \mathscr{B}_2 \subset \mathscr{L}_0$. Repeated application of Lemma 3.7 yields the result that $\mathrm{cl}\, \mathscr{R}_t(x) \supset I^t(\mathscr{B}_2, x)$. More generally, if

$$\mathscr{B} = \mathscr{B}_0 \subset \mathscr{B}_1 \subset \mathscr{B}_2 \subset \cdots \subset \mathscr{B}_n$$

is a chain of Lie subalgebras of $\mathscr{L}_0$ with $\mathscr{B}_i$ A-generated from $\mathscr{B}_{i-1}$ for $i = 1, 2, \cdots, n$, we say $\mathscr{B}_n$ is *A-related* to $\mathscr{B}$. Clearly

$$\mathrm{cl}\, \mathscr{R}_t(x) \supset I^t(\mathscr{B}_n, x).$$

DEFINITION. The *A-radical* for $\mathscr{B}$, $\mathscr{R}(A; \mathscr{B})$, is defined as the smallest Lie subalgebra of $\mathscr{L}_0$ which contains every subalgebra of $\mathscr{L}_0$ that is A-related to $\mathscr{B}$, i.e.,

$$\mathscr{R}(A; \mathscr{B}) = \{\mathscr{H} : \mathscr{H} \text{ is } A\text{-related to } \mathscr{B}\}_{\mathrm{LA}}.$$

We are now in a position to prove our main results.

Proof of Theorem 3.1. Let $\{\mathscr{H}_\alpha : \alpha \in \mathscr{I}\}$ be the collection of Lie subalgebras of $\mathscr{L}_0$ which are A-related to $\mathscr{B}$. By definition, $\mathscr{R}(A; \mathscr{B}) = \{\mathscr{H}_\alpha : \alpha \in \mathscr{I}\}_{\mathrm{LA}}$. As noted above, Lemma 3.7 implies that $\mathrm{cl}\, \mathscr{R}_t(x) \supset I^t(\mathscr{H}_\alpha, x)$ for all $t > 0$, $x \in M$ and $\alpha \in \mathscr{I}$. Since $I^t(\mathscr{H}_\alpha, x) = G(\mathscr{H}_\alpha) \cdot A_t(x)$, we know that

$$\mathrm{cl}\, \mathscr{R}_t(x) \supset \{g_1 g_2 \cdots g_n \cdot A_t(x) : g_i \in G(\mathscr{H}_\alpha) \text{ for some } \alpha \in \mathscr{I}\}.$$

By Theorem 2.1, this set is $G(\{\mathscr{H}_\alpha : \alpha \in \mathscr{I}\}_{\mathrm{LA}}) \cdot A_t(x) = G(\mathscr{R}(A; \mathscr{B})) \cdot A_t(x) = I^t(\mathscr{R}(A; \mathscr{B}), x)$, which completes the proof.

Proof of Theorem 3.2. If $\mathcal{R}(A;\mathcal{B})$ and $\mathcal{L}_0$ define the same distributions on M, then $I(\mathcal{R}(A;\mathcal{B}),x)=I(\mathcal{L}_0,x)$ for all $x\in M$. Thus $\operatorname{cl}\mathcal{R}_t(x)\supset I^t(\mathcal{L}_0,x)$ from Theorem 3.1. In proving Theorem 3.6, we showed that this implies that $\mathcal{R}_t(x)=I^t(\mathcal{L}_0,x)$.

Proof of Theorem 3.2, *Corollary* 1. If $\mathcal{L}_0=\mathcal{R}(A;\mathcal{B})$, then the conditions of Theorem 3.2 are satisfied and the proof is complete.

Proof of Theorem 3.2, *Corollary* 2. Assume that $\mathcal{R}(A;\mathcal{B})(x)=\mathcal{L}_0(x)$ for some fixed $x\in M$, and set $\mathcal{R}=\mathcal{R}(A;\mathcal{B})$. Since $\mathcal{R}\subset\mathcal{L}_0$ it follows that $I(\mathcal{R},x)$ contains an open neighborhood $\mathcal{U}$ of x in $I(\mathcal{L}_0,x)$. Theorem 3.1 implies that $\operatorname{cl}\mathcal{R}_t(x)$ contains $I^t(\mathcal{R},x)=I(\mathcal{R},A_t(x))$ for all $t>0$. Moreover, if $\alpha\to T_\alpha(x)$ is a trajectory for the system (*), then for all $0<\varepsilon<t$,

$$T_{t-\varepsilon}(I(\mathcal{R},A_\varepsilon(x)))\subset\operatorname{cl}\mathcal{R}_t(x).$$

Suppose that $y\in\mathcal{R}_t(x)\cap\partial(\operatorname{cl}\mathcal{R}_t(x))$. Then there exists a trajectory $\alpha\to T_\alpha(x)$ for the system (*) with $T_0(x)=x$ and $T_t(x)=y$. Thus for $0<\varepsilon<t$, $T_{t-\varepsilon}(I(\mathcal{R},A_\varepsilon(x)))\subset\operatorname{cl}\mathcal{R}_t(x)$, and taking the limit as $\varepsilon\to 0$, we see that

$$T_t(I(\mathcal{R},x))\subset\operatorname{cl}\mathcal{R}_t(x).$$

Since $I(\mathcal{R},x)$ contains an open neighborhood $\mathcal{U}$ of x in $I(\mathcal{L}_0,x)$, $\operatorname{cl}\mathcal{R}_t(x)$ contains the open neighborhood $T_t(\mathcal{U})$ of y, so $y\in\operatorname{int}(\operatorname{cl}\mathcal{R}_t(x))$, which contradicts the assumption that $y\in\partial(\operatorname{cl}\mathcal{R}_t(x))$. Thus $\mathcal{R}_t(x)\cap\partial(\operatorname{cl}\mathcal{R}_t(x))=\varnothing$, which proves the main assertion. To complete the proof, we note that $\mathcal{R}_t(x)$ is closed iff $\mathcal{R}_t(x)=\operatorname{cl}\mathcal{R}_t(x)$. By assumption, $\mathcal{R}(A;\mathcal{B})(x)=\mathcal{L}_0(x)$, hence $\mathcal{R}_t(x)\cap\partial(\operatorname{cl}\mathcal{R}_t(x))=\operatorname{cl}\mathcal{R}_t(x)\cap\partial(\operatorname{cl}\mathcal{R}_t(x))=\varnothing$. Moreover, $\partial(\operatorname{cl}\mathcal{R}_t(x))=\operatorname{cl}\mathcal{R}_t(x)\sim\operatorname{int}(\operatorname{cl}\mathcal{R}_t(x))$, thus $\partial(\operatorname{cl}\mathcal{R}_t(x))=\varnothing$ and $\operatorname{cl}\mathcal{R}_t(x)=\operatorname{int}(\operatorname{cl}\mathcal{R}_t(x))$. Since $I^t(\mathcal{L}_0,x)$ is connected, the only nonempty open and closed subset is $I^t(\mathcal{L}_0,x)$ itself. Thus $\operatorname{cl}\mathcal{R}_t(x)=I^t(\mathcal{L}_0,x)$. As we noted in the proof of Theorem 3.6, this implies that $\mathcal{R}_t(x)=I^t(\mathcal{L}_0,x)$, which completes the proof.

Example 1. This system fails to satisfy the hypothesis of Theorem 3.6, but $\mathcal{R}(A;\mathcal{B})$ and $\mathcal{L}_0$ define the same distributions on R^3, and Corollary 1 of Theorem 3.2 shows that $\mathcal{R}_t(x)=R^3$ for all $t>0$ and $x\in R^3$. Consider the system

$$\frac{dx}{dt}=A(x)+u_1B_1(x)+u_2B_2(x),$$

where $x=(x_1,x_2,x_3)\in R^3$, $A(x)=(0,x_1x_2,x_2)$, $B_1(x)=(0,x_1,0)$ and $B_2(x)=(1,0,x_1)$. By direct computation, $B_3=[B_1,B_2]=(0,-1,0)$ and $[B_1,B_3]=[B_2,B_3]=0$. Thus $\mathcal{B}$ has a basis $\{B_1,B_2,B_3\}$. Also,

$$\operatorname{ad}_A^kB_1(x)=(0,(-1)^kx_1^{k+1},(-1)^kx_1^k),$$

so $\mathcal{L}$ is an infinite-dimensional Lie algebra. Since

$$[\operatorname{ad}_AB_1,B_2]=(0,2x_1,1)$$

it follows that $[\operatorname{ad}_AB_1,B_2](0,0,0)=(0,0,1)\notin\mathcal{B}(0,0,0)$, so the hypothesis of Theorem 3.6 is not satisfied; however,

$$[\operatorname{ad}_A^kB_1,B_1](x)=(0,0,0)\quad\text{for all } x\quad\text{in } M=R^3.$$

This means $\hat{\mathcal{B}}=\{B_1\}_{\mathrm{LS}}$ is an ideal in $\hat{\mathcal{L}}_0=\{\operatorname{ad}_A^kB_1:k=0,1,\cdots\}_{\mathrm{LA}}$, so $\mathcal{B}_1=$

$\{\mathscr{B}, \hat{\mathscr{L}}_0\}_{\mathrm{LA}}$ is A-generated from $\mathscr{B}$ and $\mathscr{R}(A;\mathscr{B}) \supset \mathscr{B}_1$. Thus $[\mathrm{ad}_A B_1, B_2](x) = (0, 2x_1, 1) \in \mathscr{R}(A;\mathscr{B})(x)$, and because $\{B_2, B_3. [\mathrm{ad}_A B_1, B_2]\}_{\mathrm{LS}}(x) = R^3$ for all $x \in R^3$, $\mathscr{R}(A;\mathscr{B})(x) = R^3$ for all $x \in R^3$. This means that $\mathscr{R}(A;\mathscr{B})$ and $\mathscr{L}_0$ define the same distributions on R^3, and Corollary 1 of Theorem 3.2 implies that $\mathscr{R}_t(x) = R^3$ for all $t > 0$ and x in R^3. Note that $\dim \mathscr{B}(x) = 2$ for all $x \in R^3$.

Example 2. This system passes the "standard linear test", but $\mathscr{R}(A;\mathscr{B}) \cdot (x_0) \neq \mathscr{L}_0(x_0)$ and $\underline{\mathscr{R}_t(x_0) \neq A_t(I(\mathscr{L}_0, x_0))}$. This points out the relevance of the global object $\mathscr{R}(A;\mathscr{B})$ to the global controllability problem and the irrelevance of the local "standard linear test". Consider the system

$$\frac{dX}{dt}(t) = AX(t) + u(t)BX(t),\ X(0) = X_0 = 1, \tag{1}$$

where

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},\quad B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},\quad I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here

$$[A, B] = AB = BA = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad [[A, B], A] = 0,$$

$[[A, B]B] = -[A, B]$. Thus $\mathscr{L} = \{A, B\}_{\mathrm{LA}}$ has a basis $\{A, B, [A, B]\}$, $\mathscr{L}_0 = \{\mathrm{ad}_A^k \mathscr{B} : k = 0, 1, \cdots\}_{\mathrm{LA}}$ has a basis $\{[A, B], B\}$ and $\mathscr{B} = \{\mathscr{B}\}_{\mathrm{LA}}$ has a basis $\{B\}$. In this example, $\mathscr{B}$ is *not* an ideal in $\mathscr{L}_0$ (i.e., $[[A, B], B] \notin \mathscr{B}$), and $\mathscr{R}(A;\mathscr{B}) = \mathscr{B} \neq \mathscr{L}_0$ and $\mathscr{R}(A;\mathscr{B})(X_0) \neq \mathscr{L}_0(X_0)$.

On the other hand, the "standard linear test" is satisfied, i.e., $\mathscr{L}_0 = \{\mathrm{ad}_A^k B : k = 0, 1, \cdots\}_{\mathrm{LS}}$; hence

$$\dim \mathscr{L}_0(X_0) = \dim \{\mathrm{ad}_A^k B(X_0) : k = 0, 1, \cdots\}_{\mathrm{LS}}.$$

We now show that $\mathscr{R}_t(I) \neq e^{At}\{e^{\mathscr{L}_0}\}G = A_t(I(\mathscr{L}_0, X_0))$. If u is the constant control c, then the corresponding solution to (1) is

$$X(t) = e^{t(A+cB)} = \begin{pmatrix} e^t & 0 & 0 \\ 0 & e^{ct} & \frac{1}{c}(e^{ct} - 1) \\ 0 & 0 & 1 \end{pmatrix}$$

and for each $t \geqq 0$, $X(t)$ is a matrix with nonnegative entries. It follows that trajectories of system (1) evolve in the space of matrices with nonnegative entries, and $\mathscr{R}_t(X_0)$ contains no matrix with negative elements.

Let $L = 2B + 2[A, B] \in \mathscr{L}_0$. Then by direct computation,

$$e^A e^L = \begin{pmatrix} e & 0 & 0 \\ 0 & e^2 & 2 - e^2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Since $e^A e^L \in A_t(I(\mathscr{L}_0, X_0))$ is a matrix with negative entry $2 - e^2$, $e^A e^L \notin \mathscr{R}_t(X_0)$, and thus

$$\mathscr{R}_t(X_0) \neq A_t(I(\mathscr{L}_0, X_0)) = e^{At}\{e^{\mathscr{L}_0}\}_G.$$

Acknowledgment. The author wishes to thank R. W. Brockett and J. Davis for many interesting discussions.

REFERENCES

[1] R. ARENS, *Topologies for homeomorphism groups*, Amer. J. Math., 68 (1946), pp. 593–610.
[2] W. BOOTHBY, *A transitivity problem from control theory*, J. Differential Equations, submitted.
[3] R. BROCKETT, *System theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265–284.
[4] ———, *Control theory on lie groups*, Geometric Methods in Systems Theory, NATO A. S. I. Series, D. Reidel, Boston, 1973.
[5] P. COHN, *Lie Groups*, Cambridge University Press, London, 1957.
[6] D. ELLIOT, *A consequence of controllability*, J. Differential Equations, 10 (1971), pp. 364–370.
[7] G. HAYNES AND HERMES, *Nonlinear controllability via Lie theory*, this Journal, 8 (1970), pp. 450–460.
[8] S. HELGASON, *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1969.
[9] R. HERMANN, *Differential Geometry and the Calculus of Variations*, Academic Press, New York, 1968.
[10] R. HIRSCHORN, *Topological semigroups, sets of generators and controllability*, Duke Math. J., 40 (1973), pp. 937–947.
[11] ———, *Controllability in nonlinear systems*, J. Differential Equations, submitted.
[12] ———, Ph.D. thesis, Harvard Univ., Cambridge, Mass., 1973.
[13] G. HOCHSCHILD, *The Structure of Lie Groups*, Holden-Day, San Francisco, 1965.
[14] S. KOBAYASHI, *Foundations of Differential Geometry*, John Wiley, New York, 1964.
[15] C. LOBRY, *Contrôlabilité des systèmes non linéaires*, this Journal, 8 (1970), pp. 573–605.
[16] R. PALAIS, *A global formulation of the Lie theory of transformation groups*, Mem. Amer. Math. Soc., no. 22, 1957.
[17] H. SUSSMANN AND V. JURDJEVIC, *Control system on Lie groups*, J. Differential Equations, Sept. (1972).
[18] ———, *Controllability of Nonlinear Systems*, Ibid., 12 (1972), pp. 95–116.
[19] F. WARNER, *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman and Co., 1970.
[20] J. WOLF, *Spaces of Constant Curvature*, McGraw-Hill, New York, 1967.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

INPUT–OUTPUT DESCRIPTION OF ROOMY SYSTEMS*

P. DEWILDE†

Abstract. In the classical dynamic systems theory, precise information about a system can be deduced from its input–output map. In fact, for minimal systems, a complete pole-zero theory can be constructed using polynomial coprime factorization techniques, together with algebraic properties of the state space module. In this paper, an infinite-dimensional theory in the same style is presented whereby the input–output maps are assumed to exhibit some energy conservation properties and whereby these maps are ascertained to belong to a class of systems with nontrivial nullspace, called "roomy" systems. As a result, a coprime factorization theory can be deduced based not on properties of polynomials but of analytical functions, a complete polar description of the system can be given and a zero description for a somewhat more restricted class. The mathematical tools used lean heavily on Helson and Lowdenslaeger's invariant subspace theory of Hardy spaces which quite naturally comes into play through the Bochner–Chandrasekharan and the Beurling–Lax theorem. The result, however, is a complete systems description of a Roomy input–output function.

This paper is situated entirely at the input–output level, and avoids using system realizations, to concentrate on properties which can be directly deduced from the input–output map.

1. Introduction. We will investigate the dynamical system properties of an input–output map:

$$\mathcal{S} : L^2_{R^m}[(-\infty, \infty)] \to L^2_{R^n}[(-\infty, \infty)] : a(\cdot) \to \mathcal{S}a(\cdot), \tag{1.1}$$

where $\mathcal{S}$ satisfies following assumptions:

(1.2) $\mathcal{S}$ is linear.

(1.3) $\mathcal{S}$ commutes with translations. Let T_τ be a translation operator $T_\tau f(t) = f(t-\tau)$ on any space L $[(-\infty, \infty)]$; then $\mathcal{S}T_\tau = T_\tau \mathcal{S}$.

(1.4) $\mathcal{S}$ is bounded as a Hilbert space operator.

A natural setting for this kind of map is obtained by Fourier transformation. Let $\mathbb{I}$ be the imaginary axis in $\mathbb{C}$; then we define the Fourier transform

$$\mathcal{F} : L^2_{\mathbb{C}^k}((-\infty, \infty)) \to L^2_{\mathbb{C}^k}(\mathbb{I}):$$

$$f(t) \mapsto F(j\omega) = \lim_{N\to\infty} \int_{-N}^{N} f(t)\, e^{-j\omega t}\, dt$$

and by the reverse Fourier transform (to be used later),

$$\hat{\mathcal{F}} : L^2_{\mathbb{C}^k}((-\infty, \infty)) \to L^2_{\mathbb{C}^k}(\mathbb{I}) :$$

$$f(t) \mapsto \hat{F}(j\omega) = \lim_{N\to\infty} \int_{-N}^{N} f(t)\, e^{j\omega t}\, dt.$$

Clearly $\hat{F}(j\omega) = F(-j\omega)$, $\mathcal{F}$ and $\hat{\mathcal{F}}$ are Hilbert space isomorphisms (Plancherel's theorem). In the sequel we will have to distinguish between a function as a member of some space (e.g., $f \in L^2_{\mathbb{C}^k}((a, b))$) and its functional value as a member of a different space (e.g., $f(t) \in \mathbb{C}^k$). We will try to use consistently the simple symbol ("f") for the first case, and the functional value "$f(t)$" for the second. Also

* Received by the editors October 10, 1974, and in revised form June 18, 1975.

† Faculty of Engineering, Katholieke Universiteit te Leuven, Afdeling Toegepaste Wiskunde en Programmatie, 3030 Heverlee, Belgium. This work was supported by the Belgian National Fund of Scientific Research (NFWO).

while using norms, we will have for $\|f\|$, the norm of f as a member of $L^2_{\mathbb{C}^k}$, while $\|f(t)\|$ is the norm of f as a member of $\mathbb{C}^k$ (Euclidean norm). Other norms will be indicated by suffixes, unless it is clear to which space the element belongs.

By the theorem of Bochner–Chandrasekharan ([1 p. 140 ff.]) and a trivial extension of it to multivariable systems, we have that $\mathcal{S}$ can be represented as a multiplicative operator acting on the Fourier transform of the input function $A = \mathcal{F}a$ to the Fourier transform of the output function $B = \mathcal{F}b$, so that $B(j\omega) = S(j\omega)A(j\omega)$, whereby $S(j\omega)$ is an $n \times m$ matrix function of the frequency ω, $S(j\omega)$ is essentially bounded (hence with entries belonging to L^∞ of the imaginary axis) and

$$\|\mathcal{S}\| = \|S\|_\infty, \tag{1.5}$$

whereby the left-hand side indicates the norm of $\mathcal{S}$ as a Hilbert space operator, and the right-hand side the $H^\infty_{n\times m}$-norm

$$\|S\|_\infty = \operatorname{ess\,sup} \|S(j\omega)\|, \tag{1.6}$$

the essential supremum being taken over the values of the norm of $S(j\omega)$ as a Euclidean map $\mathbb{C}^m \to \mathbb{C}^n$. S is called the "transfer function" of the system and belongs to the subspace $L^\infty_{n\times m}(\mathbb{I})$ of matrices with essentially bounded entries on $\mathbb{I}$ and norm (1.6).

It should be noted that the Bochner–Chandrasekharan theorem provides the means to avoid distribution theory at the (reasonable) cost of knowledge of Hilbert space theory. The proceeding could, of course, be axiomatized completely, thereby avoiding Fourier transforms, or at least introducing them in an abstract way, but since our purpose is mainly practical, we will not indulge in undue abstraction.

We will always think of any $L^2_{\mathbb{C}^k}((a, b))$-space or $L^\infty_{\mathbb{C}^k}((a, b))$-space as natural subspaces of $L^2_{\mathbb{C}^k}((-\infty, \infty))$ or $L^\infty_{\mathbb{C}^k}((-\infty, \infty))$.

For all input–output functions $\mathcal{S}$ considered in the sequel, we will add a fourth property, the property of causality:

$$\mathcal{S}L^2_{R^m}([0, \infty)) \subset L^2_{R^n}([0, \infty)). \tag{1.7}$$

By the Paley–Wiener theorem, we have that $\mathcal{F}L^2_{\mathbb{C}^k}([0, \infty)) = H^2_k$, where the Hardy space H^2_k is defined as the subspace of $L^2_{\mathbb{C}^k}((-j\infty, j\infty))$ of functions f for which there exists an analytic continuation $f(p)$ to the open right complex plane (ORP) with the properties that

(i) $f(\sigma + j\omega)$ is quadratically integrable in ω for all $\sigma > 0$ and

(ii) $f(j\omega)$ is the limit a.e. of $f(\sigma + j\omega)$ for $\sigma \overset{>}{\to} 0$.

Functions belonging to H^2_k will be called "analytic". Likewise K^2_k can be defined symmetrically for functions having analytic continuations to the open left complex plane (OLP) satisfying—mutatis mutandis—conditions (i) and (ii). Functions belonging to K^2_k will be called "conjugate analytic". It is well known [2] that $L^2_{\mathbb{C}^k}((-j\infty, j\infty)) = H^2_k \oplus K^2_k$, so that only the zero function is analytic and conjugate analytic. We can also distinguish "analytic" and "conjugate analytic" functions in $L^\infty_{n\times m}(\mathbb{I})$. By the Hardy space $H^\infty_{n\times m}$ is understood the subspace of $n \times m$ matrices in L^∞_m which are uniform limits a.e. of bounded analytic $n \times m$ matrices in ORP.

Functions in $H^{\infty}_{n\times m}$ will again be called analytic. Likewise for bounded analytic OLP functions, we have $K^{\infty}_{n\times m}$, the subspace of $L^{\infty}_{n\times m}(\mathbb{I})$ of conjugate analytic functions. It is well known [2] that $K^{\infty}_{n\times m}\cap H^{\infty}_{n\times m}$ are exactly constant matrices.

Returning to a causal system $\mathcal{S}$, we will have, because of (2.7), that $SA\in H^2_n$ for all $A\in H^2_m$, or for each component S_{ij} of S, that $S_{ij}F\in H^2$ for all $F\in H^2$. We would like to conclude from this that $S_{ij}\in H^{\infty}$, and for that we would need to put $F=1$. The trouble is that $F\notin H^2$. One way out is to consider a new measure $d\mu = d\omega/(1+\omega^2)$ on the imaginary axis, and the corresponding subspaces, $L^2_{\mu}(\mathbb{I})$ and H^2_{μ}. Clearly $H^2\subset H^2_{\mu}$, and moreover, H^2 is dense in H^2_{μ} (this is easy to prove). It follows that $S_{ij}F\in H^2_{\mu}$ for all $F\in H^2_{\mu}$. Now $1\in H^2_{\mu}$, and hence $S_{ij}\in H^2_{\mu}$. Since $S_{ij}\in L^{\infty}(\mathbb{I})$, and $H^{\infty}=H^2_{\mu}\cap L^{\infty}(\mathbb{I})$ [3], we finally have that $S_{ij}\in H^{\infty}$. Thus, for a causal system, we have that $S\in H^{\infty}_{n\times m}$.

There is no need to require input and output spaces to be real. We will use, as the input space: $\Omega = L^2_{\mathbb{C}^m}((-\infty,\infty))$ and as output space: $\Gamma = L^2_{\mathbb{C}^n}((-\infty,\infty))$.

For a general introduction to these spaces, and the related Hardy spaces, the reader is referred to the excellent textbooks [2], [3], [4], [5], [6]. Bounded input–output maps in the sense described above, arise especially in network theory (see [7], [8], [9]).

2. Roomy systems. Following Kalman's [10, chap. 10] original methodology in constructing an abstract state space from the input–output map, we restrict our attention to the map

$$(2.1)\qquad \mathcal{S}_0 : \Omega_0 = L^2_{\mathbb{C}^m}((-\infty,0)) \to \Gamma_0 = L^2_{\mathbb{C}^n}([0,\infty)) : a(t)\mapsto \mathcal{S}b(t)|_{[0,\infty)}$$

from the space of input functions $\Omega_0 = L^2_{\mathbb{C}^m}((-\infty,0])$ to the space of output functions $\Gamma_0 = L^2_{\mathbb{C}^n}([0,\infty))$, whereby the action of the system on an input up to time zero (representing, in fact, any t, by shift invariance) is investigated. The information which the system gleams from an input is referred to as the state, and in a very natural way, two input functions $f_1(t)$ and $f_2(t)$ in Ω_0 will produce the same state, if there is no way in which the output (observed after $t=0$) will distinguish between these two, whatever input one may subsequently (i.e., for $t>0$) apply. Because of linearity, a zero input for $t>0$ is as good as any, and we can simply say that f_1 and f_2 in Ω_0 generate the same state or are Nerode equivalent, written

$$f_1 \overset{N}{\sim} f_2 \quad \text{if } \mathcal{S}_0 f_1 = \mathcal{S}_0 f_2.$$

Let

$$(2.2)\qquad \mathcal{M}_1 = \{f\in\Omega_0 : f\overset{N}{\sim}0\}.$$

We will call $\mathcal{M}_1$ the "nullspace" of the system. The natural state space then is the set of Nerode equivalent classes. In this case, $\mathcal{M}_1$ is a closed linear subspace of the Hilbert space Ω_0, and the natural state space $\mathcal{H}_1$ can be taken as its orthogonal complement $\mathcal{M}_1^{\perp}$ (for a module discussion of this result see [11]).

If $m=1$ (so called monovariable systems) either $\mathcal{M}_1=\{0\}$, or there is a nonzero $f\in\Omega_0$ such that $f\in\mathcal{M}_1$.

DEFINITION 2.1. A monovariable system with $\mathcal{M}_1$ nontrivial will be called *roomy*.

More generally, let P be a projection of $\mathbb{C}^m$ on a one-dimensional subspace A_p, and consider the monovariable system $\mathcal{S}_P = \mathcal{S}P$.

DEFINITION 2.2. A system $\mathcal{S}$ will be called roomy if $\mathcal{S}_P$ is roomy for all P.

Note. If P is a projection in $\mathbb{C}^k$, then it can trivially be used as a projection in $L^2_{\mathbb{C}^k}$, by putting $(Pf)(j\omega) = Pf(j\omega)$.

To characterize roomy systems in a more useful way, we transform the spaces considered by means of the reverse Fourier transform $\mathcal{F}$. Let

$$\hat{\mathcal{M}}_1 = \mathcal{F}\mathcal{M}_1. \tag{2.3}$$

Then $\hat{\mathcal{M}}_1$ is a subspace of H^2_m, and $\hat{\mathcal{H}}_1 = \mathcal{F}\mathcal{H}_1 = \hat{\mathcal{M}}_1^{\perp}$. $\mathcal{M}_1$ is left shift invariant since, if $f(t) \in \Omega_0$ produces a zero state, then obviously $T_\tau f(t)$ $(\tau < 0)$ will. Hence $\hat{\mathcal{M}}_1$ is invariant for multiplications with $e^{-j\omega\tau}$ $(e^{-j\omega\tau}\hat{\mathcal{M}}_1 \subset \hat{\mathcal{M}}_1)$, and more generally, $\hat{\mathcal{M}}_1$ is invariant for multiplication with H^∞ functions $(f\hat{\mathcal{M}}_1 \subset \mathcal{M}_1,\ f \in H^\infty)$ by a standard density argument [3]. It follows that $\hat{\mathcal{M}}_1$ is an "invariant subspace" in the sense of Helson–Lowdenslaeger [2, p. 7]. At this point, we want to use the Beurling–Lax theorem to characterize $\hat{\mathcal{M}}_1$. An operator $A : l^2_{\mathbb{C}^k}(\mathbb{I}) \to L^2_{\mathbb{C}^m}(\mathbb{I})$ will be called "multiplicative" if it is represented by an $L^\infty_{m\times k}(\mathbb{I})$ matrix A, so that (with a slight confusion in notation) $(AF)(j\omega) = A(j\omega)F(j\omega)$ for all $F \in L^2_{\mathbb{C}^k}(\mathbb{I})$. Its adjoint is indicated by A^*, and it is easy to see that $A^*(j\omega) = \tilde{A}(j\omega)$, where the $\tilde{}$ indicates Hermitian conjugation. A is an isometry if $A^*A = 1$ or if $\tilde{A}(j\omega)A(j\omega) = 1_k$ a.e. (ω). By the Beurling–Lax theorem [2, p. 61], we now have that $\hat{\mathcal{M}}_1 = V_1 H^2_k$ for some $k \leqq m$, with V_1 a multiplicative isometry $H^2_k \to H^2_m$. Hence $V_1 \in H^\infty_{m\times k}$, and $\tilde{V}_1(j\omega)V_1(j\omega) = 1_k$ a.e. ω. We obtain the following simple characterizations of a roomy system.

PROPOSITION 2.1. *A system is roomy if and only if $k = m$, i.e., if $\hat{\mathcal{M}}_1$ has pointwise full dimension a.e.* (such $\hat{\mathcal{M}}_1$ are said to have "full range" [2]).

Proof. The proof is in Appendix A.

Let $\{e_i\}$ be a basis in $\mathbb{C}^m$ and $\{f_j^*\}$ a basis in $(\mathbb{C}^n)^*$, and consider one-dimensional systems: $\mathcal{S}_{ji} = f_j^* \mathcal{S} P_i$, where P_i is the orthogonal projection on e_i in $\mathbb{C}^m$. We have the following corollary.

COROLLARY 2.1. *$\mathcal{S}$ is roomy if and only if the systems $\mathcal{S}_{ij}$ are.*

Proof. Again the proof is given in Appendix A.

For roomy systems such that $m = n = 1$, we have that $\hat{\mathcal{M}}_1 = \phi H^2$, where $\phi \in H^\infty$ and $|\phi(j\omega)| = 1$. Such a function is called "inner". More generally, if $\mathcal{S}$ is roomy, and thus $\hat{\mathcal{M}}_1 = U_1 H^2_m$, then U_1 is unitary in $L^2_{\mathbb{C}^m}(\mathbb{I})$. Also, $U_1(j\omega)$ is a.e. unitary in $\mathbb{C}^m$. Such a U_1 is likewise called "inner".

Suppose $\mathcal{S}$ is monovariable. Then, either $\hat{\mathcal{M}}_1$ is empty (and the system is not roomy) or $\hat{\mathcal{M}}_1$ is infinite-dimensional and the system is roomy. So either the natural state space fills the input space or there is an infinite-dimensional nullspace.

3. Coprime factorization of roomy systems. Very useful to the subsequent results will be the notion of coprime factorizations. Polynomial coprime factorization has been used extensively in systems theory [12], [13], [14]. Roughly, the technique is based on factoring a rational matrix $R(p) = L_1^{-1}(p)\,\Delta_1(p) =$

$\Delta_2(p)L_2^{-1}(p)$, where L_i, Δ_i $(i=1,2)$ are polynomial matrices, and the factorization is "minimal" in a certain sense. This minimality is expressed by a notion of coprimeness: L_1 and Δ_1 are left coprime, while Δ_2 and L_2 are right coprime. We will use a similar technique, replacing "polynomial" by "analytic". It will turn out that our factors are closely related to the inner functions defining nullspaces.

We first introduce some notions about coprime factorization of H^∞ matrices. Let $A \in H^\infty_{n\times k}$ and $B \in H^\infty_{n\times m}$. We will say that A and B have a common left inner divisor (CLID) if there exists an $n\times n$ function $U(j\omega)$ such that

$$A = UA_1, \quad B = UB_1, \quad A_1 \in H^\infty_{n\times k}, \quad B_1 \in H^\infty_{n\times m}. \tag{3.1a}$$

$$U \text{ is inner, i.e., } U \in H^\infty_{n\times n} \text{ and } U^*U = 1. \tag{3.1b}$$

U will be called greatest common left inner divisor (GCLID) if, for any CLID U_1, there is an inner function U_2, such that

$$U = U_1U_2. \tag{3.2}$$

The above definitions make sense only if the rows of the matrix $[A, B]$ span $\mathbb{C}^n$ almost everywhere (which will always be the case in our theory). We have then (with " $\vee$ " indicating the sum of subspaces) Proposition 3.1.

PROPOSITION 3.1. *Let the columns of the matrix $[A, B]$ span $\mathbb{C}^n$ almost everywhere. Then A and B have a unique* GCLID *U. With $\mathcal{M}_A = \overline{AH_k^2}$ and $\mathcal{M}_B = \overline{BH_m^2}$, we have that $\mathcal{M}_A \vee \mathcal{M}_B = UH_n^2$.*

Proof. The proof is in Appendix B.

Note. U is unique except for a trivial right constant unitary factor. A and B will be called left inner coprime (LIC) if their GCLID is a constant (obviously unitary) matrix. We now have the following theorem.

THEOREM 3.1. *A and B are* LIC *if and only if there exist sequences of matrices M_i and N_i with entries in H^∞ such that*

$$\lim_{i\to\infty} (AM_i + BN_i) = 1_n, \tag{3.3}$$

the limit standing for columnwise $L^2(d\omega/(1+\omega^2))$ convergence (and hence pointwise convergence a.e., or uniform convergence on compact subsets of the ORP as well).

Proof. The proof is in Appendix B.

THEOREM 3.2. *Let $T(j\omega)$ be an $n\times m$ matrix function of ω and $T(j\omega) = U^{-1}(j\omega)\,\Delta(j\omega)$, where U is $n\times n$ inner and Δ is analytic. Then if Δ and U are left coprime, the factorization is unique up to a left constant unitary matrix.*

Proof. The proof is in Appendix B.

We are now in a position to discuss the coprime factorization theory for a roomy system $\mathcal{S}$.

THEOREM 3.3. *Let $\mathcal{S}$ be a roomy system and $\hat{\mathcal{M}}_1 = U_1H_m^2$. Then $S(-j\omega)U_1(j\omega) = \Delta_1(j\omega)$ is analytic, and U_1 and Δ_1 are right inner coprime* (RIC).

Proof. The proof is in Appendix B.

Hence $S(-j\omega) = \Delta_1(j\omega)U_1(j\omega)^{-1} = \Delta_1(j\omega)\tilde{U}_1(j\omega)$ or

$$S(j\omega) = \Delta_1(-j\omega)U_1(-j\omega)^{-1} \quad \text{a.e.} \tag{3.4}$$

Formula (3.4) is peculiar in the following sense: $S(j\omega)$ is $H^{\infty}_{n\times m}$ and has an analytic continuation to the ORP. Both $\Delta_1(-j\omega)$ and $U_1(-j\omega)$, however, are conjugate analytic and thus have an analytic continuation to the OLP. Moreover, $U_1(-j\omega)$ is unitary (in $L^2_{n\times m}(\mathbb{I})$), so that its inverse $U_1^{-1}(-j\omega)$ is quite well behaved in the OLP. Formula (3.4) provides a "pseudo-analytic" continuation [15], [16] of S in the left half-plane. U_1 contains all the "polar" information about $\mathcal{S}$. It describes its natural state space completely (through $\hat{\mathcal{M}}_1$), and from 4.4, it appears that it describes completely the singularities in its pseudo-analytical continuation. We have obtained the following result.

THEOREM 3.4. *A roomy system $\mathcal{S}$ has a pseudo-analytical continuation of the form $\Delta_1(-j\omega)U_1^{-1}(-j\omega)$, where Δ_1 and U_1 are analytic functions with entries in H^{∞}, U_1 is inner and Δ_1 and U_1 are right coprime. Any other pseudo-analytical continuation of the type $\Delta_2(-j\omega)U_2^{-1}(-j\omega)$ is such that U_1 is a left divisor of U_2 (or there is an inner U_3 such that $U_2 = U_1U_3$ and $\Delta_1 = \Delta_2U_3$).*

Of all the factorizations given by formula (3.4), the one with Δ_1 and U_1 is the minimal one in the sense of the coprime theory developed with analytical functions. It is clear that if $S(j\omega) = \Delta_2(-j\omega)U_2(-j\omega)^{-1}$ is a pseudo-analytical continuation of S, then $\hat{\mathcal{M}}_1 \supset U_2H^2_n$, so that $\hat{\mathcal{M}}_1$ is the largest subspace of Ω_0 which is mapped by $\mathcal{S}$ in $L^2_n((-\infty\ 0])$. A left coprime factorization can be obtained by producing a right coprime factorization on the dual system $\mathcal{S}_d = T_-\mathcal{S}^*T_-$ where "*" indicates the usual Hilbert space dual and T_- indicates time reversal. The transfer function for $\mathcal{S}_d$ is of course $\tilde{S}(-j\omega)$, and by the previous theory we have

$$\tilde{S}(-j\omega) = \Delta_2(-j\omega)U_2(-j\omega)^{-1} \tag{3.5a}$$

or

$$S = [U_2^*]^{-1}\,\Delta_2^* = U_2\,\Delta_2^* \tag{3.5b}$$

with $\tilde{U}_2(-j\omega)$ and $\tilde{\Delta}_2(-j\omega)$ left coprime. The relation between U_2 and U_1 is given by the following theorem.

THEOREM 3.5. *Given the right and the left coprime factorization for $S = \Delta_1(-j\omega)U_1^{-1}(-j\omega) = [\tilde{U}_2(j\omega)]^{-1}\,\tilde{\Delta}_2(j\omega)$, we have that* $\det U_1(-j\omega) = \det \tilde{U}_2(j\omega)$.

Proof. The proof is given in Appendix B.

Note. Theorem 3.5 and the theory of coprime factorization in this context was developed independently by P. Fuhrmann [17]. Fuhrmann bases his approach on an elegant theory of Smith-like canonical forms due to Nordgren [18]. Our approach is more elementary, and the proof of Theorem 3.5 does not use any of the deeper properties of H^{∞} algebras.

Since Δ_1 and U_1 are analytic, we have analytic continuations $\Delta_1(p)$, $U_1(p)$, $p \in \mathrm{ORP}$ for them. The pseudo-analytic continuation for S is then $\Delta_1(-p)U_1^{-1}(-p)$, and the singularities of $U_1^{-1}(-p)$, $p \in \mathrm{OLP} \cup \mathbb{I}$ are characteristic for the system and will be discussed in the following paragraph. It will appear that, because of Theorem 3.5, the singularities of $U_1^{-1}(-p)$, $p \in \mathrm{OLP} \cup \mathbb{I}$, and of $U_2(-p^*)$, where p^* is the complex conjugate of p, are essentially the same. In fact, we have, by analytic continuation, that

$$\det U_1(-p) = \det \tilde{U}_2(-p^*) \tag{3.6}$$

for all $p \in \mathrm{OLP} \cup \mathbb{I}$. We will see that the determinant of an inner function characterizes completely its behavior.

4. Pole-zero theory of roomy systems. We start off with a "degree theory" for inner functions. Such a theory is very similar to a classical "index" theory, but we will view it more from an applied point of view, and we will take care to stick closely to the classical system theoretical notion of degree. Consider the class of $n \times n$ inner functions $\mathcal{T}_n$. $\mathcal{T}_n$ is a monoid for ordinary matrix multiplication.

We will call δ a "degree" map if following conditions are satisfied:

$$\delta : \mathcal{T}_n \to \mathcal{T}_1. \tag{4.1}$$

$$\delta(U_1 U_2) = \delta(U_1)\delta(U_2); \quad \text{and} \quad \delta(1_n) = 1, \tag{4.2}$$

(δ is a semigroup isomorphism).

$$\delta \begin{bmatrix} \phi & & & \\ & 1 & & \\ & & \cdot & \\ & & & \cdot \\ & & & 1 \end{bmatrix} = \phi. \tag{4.3}$$

PROPOSITION 4.1. $\delta(U_1) = \det U_1$.

Proof. The proof is in Appendix C.

In systems with finite-dimensional state space, we have for $U \in \mathcal{T}_n$ [9, p. 229] that deg U = deg (det U) and that the poles of U coincide with the poles of det U and have same degree. For a rational inner function U, it is also easy to define zeros as the poles of $U^{-1} = \tilde{U}(-p^*)$, and hence for every pole p_0 of U, we have a zero $-p_0^*$ of same degree. This definition of zero will be coherent with a more general one to be given below.

The justification for the extension of the notion to infinite state systems is given mainly by properties (4.1)–(4.3) and by the theory of product decomposition of inner functions [19].

We conclude this discussion with the next proposition.

PROPOSITION 4.2. *When the state space of $\mathcal{S}$ is finite-dimensional and a coprime factorization is given so that*

$$S = \Delta_1(-j\omega)U_1^{-1}(-j\omega),$$

then

$$\dim \mathcal{H} = \deg[\delta(U_1)] = \deg[\det U_1].$$

Proof. The proof is in Appendix C.

The preceding propositions show several things:

(i) In finite-dimensional systems, the classical degree (in the Smith–Macmillan sense) is exactly equal to the degree of det U_1, or U_2. It is also the dimension of the natural state space $\mathcal{H}_1$ (or of $\mathcal{H}_2$).

(ii) In infinite-dimensional systems, the notion of isomorphism of state spaces is not anymore useful, because subspaces with very different state spaces of

infinite dimension can be made isomorphic. However, det U_1 and det U_2 stay good measure of systems complexity, because of (4.2). Using them as a measure of degree makes good sense, especially in view of embedding realization techniques which will be discussed later on.

(iii) In Fuhrmann [20], this same notion is used in connection with shift realizations. There, the notion of quasi-equivalence of shifts as introduced by Moore and Nordgren [21] is found to produce the same kind of characterization.

Next we consider the notion of natural response and "natural frequency" in this more abstract set up. We show, first of all, that the "natural" response for a state x by a roomy $n \times m$ transfer function S with nullspace $\hat{\mathcal{M}}_1 = U_1 H_m^2$, $\hat{\mathcal{M}}_2 = U_2 H_m^2$, is, in a very natural way, a member of the space $\hat{\mathcal{H}}_2 = \hat{\mathcal{M}}_2^{\perp}$, where $\hat{\mathcal{M}}_2 = (U_2 H_n^2)^{\perp}$. In fact, let $x(t) \in L^2_{\mathbb{C}^m}[(-\infty, 0)]$ be an input function which reaches the state x at time $t = 0$, and let $y_1(t) \in L^2_{\mathbb{C}^n}([0, \infty))$ be the response function with no excitation: $y_1(t) = \mathcal{S}_0 x(t)$. We use a left coprime factorization for S:

$$\begin{aligned} S(j\omega) &= [U_2^*(j\omega)]^{-1} \Delta_2^*(j\omega) \\ &= U_2(j\omega)\, \Delta_2^*(j\omega), \end{aligned} \tag{4.4}$$

and with $\mathcal{S}x(t) = y_1(t) + y_2(t)$, $y_2(t) = 0$ for $t \geqq 0$ we have:

$$S(j\omega)X(j\omega) = U_2(j\omega)\, \Delta_2^*(j\omega)X(j\omega) = Y_1(j\omega) + Y_2(j\omega), \tag{4.5}$$

where $Y_1(j\omega)$ and $Y_2(j\omega)$ are Fourier transforms of $y_1(t)$ and $y_2(t)$. Hence,

$$\Delta_2^*(j\omega)X(j\omega) = U_2^*(j\omega)\, Y_1(j\omega) + Y_2^*(j\omega)\, Y_2(j\omega). \tag{4.6}$$

In (4.6), only the term $U_2^*(j\omega)\, Y_1(j\omega)$ might not be conjugate analytic, but since it is equal to the difference of two conjugate analytic terms, we have

$$U_2^*(j\omega)\, Y_1(j\omega) \in K_n^2. \tag{4.7}$$

It follows that $Y_1(j\omega) \in \hat{\mathcal{H}}_2$ since $Y_1(j\omega)$ is analytic and orthogonal, by (4.7) on $\hat{\mathcal{M}}_2 = U_2 H_n^2$.

Furthermore, the set of responses $\{y_1(t)\}$, which is not necessarily a closed set, is dense in $\mathcal{H}_2$, for suppose on the contrary that there is an

$$F(j\omega) \in H_n^2 \quad \text{such that } F \in \hat{\mathcal{H}}_2 \quad \text{and} \quad (F, SX) = 0 \quad \text{for all } X \in K_n^2.$$

Then clearly $G = \Delta_2 U_2^* F \in H_m^2$. Now, since U_2 and Δ_2 are right coprime, there exist, by Theorem 4.1, sequences M_i and N_i with entries in H^∞ such that

$$\lim_{i \to \infty} [N_i U_2 + M_i\, \Delta_2] = 1_n \tag{4.8}$$

or

$$\lim_{i \to \infty} [N_i + M_i S^*] = U_2^*. \tag{4.9}$$

Using (4.9) on F we get

$$\lim_{i \to \infty} [N_i F + M_i S^* F] = U_2^* F. \tag{4.10}$$

The right member of (4.10) is conjugate analytic while the left member is analytic. Hence,

$$U_2^* F = 0, F = 0.$$

We have obtained the following proposition.

PROPOSITION 4.3. *The closure of the set of natural responses of $\mathcal{S}$ has as Fourier transform, the set*

$$\hat{\mathcal{H}}_2 = [U_2 H^2_{R^n}]^\perp.$$

Next, the question of natural modes arises. The physical feeling one has here is that a natural response exhibits a natural mode if it stays equal to itself (modulo a constant) under the action of left shifts, followed by restriction of the shifted function to Γ_0.

This physical insight turns out to be capable of mathematical formulation. Let $\mathcal{H}_2$ ($\hat{\mathcal{H}}_2$) be adjoint state space (or its Fourier transform). Since $\mathcal{M}_2$ ($\hat{\mathcal{M}}_2$) is right invariant (invariant for multiplication with analytic functions), we have, denoting right shifts by T_τ, that

$$T_\tau \mathcal{M}_2 \subset \mathcal{M}_2 \qquad (e^{-j\omega\tau}\hat{\mathcal{M}}_2 \subset \hat{\mathcal{M}}_2)$$

and hence that

$$T_\tau^* \mathcal{H}_2 \subset \mathcal{H}_2 \qquad ((e^{-j\omega\tau})^* \hat{\mathcal{H}}_2 \subset \hat{\mathcal{H}}_2),$$

where the *-operator refers to Γ_0 (H_n^2) and not $L^2_{\mathbb{C}^n}((-\infty, \infty))$. In fact, $T_\tau^* f(t) = f(t+\tau)|_{\Gamma_0} \cdot T_\tau^*$ so defined is a strongly continuous semigroup of operators which we restrict here to $\mathcal{H}_2$. Using ideas of Moeller, Lax–Phillips and Helson summarized in [2], we gave for the spectrum of the generator A of the semigroup T_τ^* Theorem 4.3.

THEOREM 4.3. *The spectrum of A acting in $\mathcal{H}_2$ consist exactly of those complex numbers p such that* Re $p < 0$ *and $U_2(-p^*)$ is not invertible, and those p with* Re $p = 0$ *such that $U_2(p)$ cannot be continued analytically across p.*

Proof. The proof is given in Appendix C.

It should be noted, that since U_2 is unitary, it has a pseudo-analytical continuation $U_2(p) = \tilde{U}_2^{-1}(-p^*)$, and hence the poles of S coincide with the poles of U_2 in the open left half-plane. It is also true that, for the eigenvalues in Re $p < 0$ (poles of U_2 in Re $p < 0$), the multiplicities coincide as well, but the proof of this fact is too technical to be given here.

We have thus interpreted the poles of the pseudo-analytical continuation of S as eigenvalues of the generator of the semigroup of the adjoint shift operator acting on the space of natural responses which we have called the adjoint state space. Similarly, the eigenvalues of the generator of the semigroup of the adjoint shift operator acting on the state space as a subspace of Ω_0 generates the singularities of $U_1(-p)$.

We obtain a dual notion (with a little more evolved physical interpretation): a function $x(t) \in \Omega_0$ is a "natural state generator" if the state which it generates in the interval $(-\infty, 0]$ stays proportional to itself. These notions link the singularities of the pseudo-analytical continuations of S to the spectral properties of shift operators. We have already discovered two invariant subspaces defined by

our system. There are two more which we now proceed to introduce and which are related to the notion of "zeros" of $\mathscr{S}$.

Let $\mathscr{N}_1 = \{\mathscr{S}(g),\ g \in \mathscr{M}_1\}$ in $L^2_n((-\infty, 0])$. Then $\mathscr{N}_1$ is an invariant subspace for left shifts. Also, let $\mathscr{N}_2$ be the set of inputs in $L^2_m([0, \infty))$ such that Sf is in the closure of the set of natural outputs ($\mathscr{M}_2$), i.e., the set of outputs resulting from a certain state with not subsequent excitation. Also, in the coprime factorization

$$S(j\omega) = \Delta_1(-j\omega)U_1^{-1}(-j\omega) = [U_2^*(j\omega)]^{-1}\,\Delta_2^*(j\omega), \tag{4.11}$$

we can factor zeros out of $\Delta_1(-j\omega)$ and $\Delta_2^*(j\omega)$, but, in view of the fact that S might be a rectangular matrix, we have to be a bit careful. Let $m \geqq n$. Then Δ_2 is $n \times m$ and can be factored canonically into its inner and outer parts

$$\Delta_2 = V_2\,\Delta_2'.$$

As for Δ_1, we have that $\overline{\Delta_1 H_n^2}$ is an invariant subspace of less than full range and hence generates a range function [2, p. 91] which we will call $\mathscr{R}$. $\mathscr{R}^\perp$ is also a range function and we have the next proposition.

PROPOSITION 4.4. *$\mathscr{R}^\perp$ is an analytic range function if and only if Δ_1 is roomy.*

Proof. The proof is given in Appendix C.

Since $\mathscr{R}^\perp$ is analytic, there is an $m \times (m-n)$ isometric operator W_2' such that $\mathscr{R}^\perp = W_2' H^2_{m-n}$. In the class of analytic invariant subspaces contained in $\mathscr{R}^\perp$, let us take the maximal element $W_2 H^2_{m-n}$. Also, let $\overline{\Delta_1 H_n^2} = V_1 H_m^2$. V_1 is analytic and isometric by the Beuling–Lax theorem.

Then

$$W = [V_1 \quad W_2] \tag{4.12}$$

is inner, and $\mathscr{R} \subset W H_m^2$. Hence $W^{-1}\,\Delta_1 = \Delta_1'$ is analytic, and we have

$$\begin{aligned} S(j\omega) &= W(-j\omega)\,\Delta_1'(-j\omega)U_1^{-1}(-j\omega) \\ &= [U_2^*(j\omega)]^{-1}\,\Delta_2'^*(j\omega)\,V_2^*(j\omega). \end{aligned} \tag{4.13}$$

THEOREM 4.4. *$\mathscr{N}_1$ and $\mathscr{N}_2$ are invariant subspaces such that $\hat{\mathscr{N}}_1 = V_1 H_m^2$ and $\hat{\mathscr{N}}_2 = V_2 H_m^2$. Moreover, in the case $m = n$, we have*

$$\delta(W(-p)) = \delta(\tilde{V}_2(p^*)).$$

Proof. The proof is in Appendix C.

It is obvious that $U_1^{-1}(-p)$ and $\tilde{U}_2(-p^*)^{-1}$ do not produce any zeros in the left half-plane. Moreover, there is no cancellation between Δ_1 and U_2 or between Δ_2 and U_2 so that the singularities of W_1^{-1} and V_2^{-1} are "genuine". In analogy with the finite-dimensional state space case, we call, when $m = n$, $\delta(W(-p)) = \delta(\tilde{V}_2(-p^*))$ the "left half-plane zero degree maps" because they characterize the "zero" behavior of the system in the left half open plane. In analogy to Theorem 4.3, we can consider here also the spectrum of the cogenerator of T_τ^* acting in $\mathscr{N}_1$ and $\mathscr{N}_2$ and find coincidence between the spectrum and singularities of V_1 and V_2. The physical interpretation of this is again in terms of "natural zero modes"; e.g., in the case of $\mathscr{N}_2^\perp$, we have that a "natural zero mode" induces a "natural output" to stay proportional to itself.

Zeros located in the right half open plane (and incidentally a zero measure also) can also be dealt with as follows. Let's suppose (for definiteness) that $n \leqq m$

and that $\mathcal{S}$ is not outer (if $n \geqq m$, one has to work on $\mathcal{S}_d$). Then $\mathcal{F} = \overline{S(j\omega)H_m^2} = W \cdot H_n^2$, where W is a nontrivial inner function. Also, let $S(j\omega) = [U_2^*(j\omega)]^{-1}\,\Delta_2^*(j\omega)$ be a coprime factorization of $\mathcal{S}_d$, and let $\hat{\mathcal{M}}_2 = U_2 H_n^2$. As we will prove in the sequel, W consists essentially of two parts (or there is an inner factorization of W corresponding to two different types of behavior): one part produces a "genuine" degree reduction in the sense that the action of factoring that part out reduces both $S(j\omega)$ and its state space by the same inner function; the other part, however, does not reduce the state space (keeps it equal), but, when factored out of $S(j\omega)$, just replaces a right half-plane zero p_0 (or zero measure $\mu(j\omega)$) by a left half-plane zero $(-p_0^*)$ symmetrically located with respect to the imaginary axis (a zero measure $-\mu(j\omega)$ which procludes a conjugate analytic zero measure), and hence reduces the analysis to the case of left half plane zeros.

PROPOSITION 4.5. *Let* $\hat{\mathcal{M}}_2 \vee \mathcal{F} = VH_n^2 \neq H_n^2$, V *inner. Then* V *is the* GCLID *of* S *and* U_2. *Let* $U_2 = VU_2'$ *and* $W = VW'$. *Then* $S_1(j\omega) = V^{-1}(j\omega)\ S(j\omega) = [U_2'^*(j\omega)]^{-1}\,\Delta_2^*(j\omega)$, $U_2'^*(j\omega)$ *and* $\Delta_2^*(j\omega)$ *form a coprime factorion of* $\mathcal{S}_{1d}$ *and* $\hat{\mathcal{M}}_2' \vee \mathcal{F}_1 = H_n^2$ *with* $\hat{\mathcal{M}}_2' = U_2' H_n^2$ *and* $\mathcal{F}_1 = S_1(j\omega)H_m^2$.

Proof. All the facts stated are obvious from the previous discussion and previous theorems.

It should be noted that the relation $U_2 = VU_2'$, shows that the state space (which is taken to be $\mathcal{H}_2$ here) is strictly reduced by V, since $\delta(U_2) = \delta(V) \cdot \delta(U_2')$.

Now let's study a system such that $\overline{\mathcal{M}_2 \vee \mathcal{F}} = H_n^2$. Then $S_1(j\omega) = W^{-1}(j\omega)S(j\omega)$ is of course still analytic. Then we can claim the following.

PROPOSITION 4.6. *The state space* $\hat{\mathcal{H}}_1'$ *of* $\mathcal{S}_1$ *is equal to the state space* $\hat{\mathcal{H}}_1$ *of* $\mathcal{S}$.

Proof. The proof is in Appendix C.

Hence, with $W = VW_2$, where $VH_{\mathbb{C}^n}^2 = \overline{\mathcal{M}_2 \vee \mathcal{F}}$, we obtain all right half-plane zeros as (i) zeros (and zero measure) belonging to the strict unitary part of $\mathcal{S}$ and (ii) zeros which do not belong to any degree reducing unitary part of $\mathcal{S}$ and which, in the context of circuit theory, have been called as "nonminimal reactance" zeros [9, p. 150].

In classical circuit theory, there is a third kind of zero, namely, a zero on the imaginary axis. The strict H^2 theory is unable to cope with those. They can be dealt with in a more general setup using so called J-unitary coprime factorization. In § 6, we will discuss these applications to a somewhat greater extent. However, when $S(j\omega)$ is boundedly invertible on the imginary axis, the theory is complete as presented. This is the case in the theory of optimal (Wiener) filtering.

5. Examples. A few examples are presented to illustrate the theory.

(I) Let $S(p) = (p+1)/(p+2)$. Then

$$S(p) = \frac{p+1}{p-1} \cdot \frac{1}{(p+2)/(p-2)} \cdot \frac{p-1}{p-2}.$$

Hence $\hat{\mathcal{M}}_1 = \hat{\mathcal{M}}_2 = ((p-2)/(p+2))H^2$ and $\hat{\mathcal{N}}_1 = \hat{\mathcal{N}}_2 = ((p-1)/(p+1))H^2$. The system is roomy and has a complete pole and zero description.

(II) Let $S(p) = p/(p+1)$. Again $\hat{\mathcal{M}}_1 = \hat{\mathcal{M}}_2 = ((p-1)/(p+1))H^2$, but the zero spaces $\hat{\mathcal{N}}_1 = \hat{\mathcal{N}}_2 = \varnothing$. Also, there are no zeros in the right half-plane, and the theory as presented is not able to deal with the $j\omega$-axis zero $p = 0$.

(III) Let $S(p)=e^{-p}$ (pure delay). $S(p)$ is unitary. Hence $\hat{\mathcal{M}}_1=\hat{\mathcal{M}}_2=e^{-p}H^2$, or $\mathcal{M}_1=\mathcal{M}_2=L^2_{\mathbb{C}}(-\infty,-1]$. The zero set is entirely in the right half-plane and is, in fact, a zero measure; we have that the V of Proposition 4.5 is e^{-p} in this case.

(IV) $S(p)=(p-1)/(p+2)$. Of course, as in example I, we have that $\hat{\mathcal{M}}_1=\hat{\mathcal{M}}_2=(p+2)/(p-2)H^2$. However as far as the zeros are concerned, we have no unitary part in S; hence the V of Proposition 4.5 is 1. Proposition 4.6 applies with $W=(p-1)/(p+1)$, and $S(p)$ has the same state space as $S_1(p)=(p+1)/(p+2)$ but a right half-plane zero.

(V) In the time domain, we have

$$\mathcal{H}=\begin{cases}\{k\cdot e^{t}(t\leqq 0)\} & \text{for } S(p)=\dfrac{p}{p+1},\\[2ex] L^2(-1,0) & \text{for } S(p)=e^{-p}.\end{cases}$$

(VI) To illustrate Theorem 3.5, let

$$S(p)=\left[\frac{p}{p+1},\ \frac{1}{p+1}\right].$$

Right and left coprime factorizations are

$$S(p)=[1\quad 0]\begin{bmatrix}\dfrac{p}{p+1} & \dfrac{1}{p+1}\\[2ex] \dfrac{1}{p+1} & \dfrac{p}{p+1}\end{bmatrix}$$

$$=\frac{p-1}{p+1}\left[\frac{p}{p-1}\quad\frac{1}{p-1}\right].$$

Hence

$$U_2=\frac{p-1}{p+1},\qquad U_1=\begin{bmatrix}\dfrac{p}{p+1} & \dfrac{1}{p+1}\\[2ex] \dfrac{1}{p+1} & \dfrac{p}{p+1}\end{bmatrix}$$

and $\det U_1(-p)=\det U_2^*(-p^*)$. Also $\left[\dfrac{p}{p-1}\ \dfrac{1}{p-1}\right]$ and $[1\quad 0]$ have very different zero behavior as pointed out at the end of the proof of Theorem 4.4

(VII) There are nonroomy systems. An easy way to construct them is given by following principles: (a) If $S(j\omega)$ is roomy, then there is an inner $\phi(j\omega)$ such that $\phi(j\omega)S^*(j\omega)$ is analytic. Hence $\phi(j\omega)[S^*(j\omega)+S(j\omega)]$ is analytic, and Re $S(j\omega)$ cannot be zero on a set of nonzero measure. Hence, any $S(j\omega)$ with Re $S(j\omega)=0$ on a set of nonzero measure is not roomy. (This construction is in [2, p. 92]. (b) If S is roomy, then it has a pseudo-analytical continuation. Hence S cannot have branch points in the left half-plane.

Examples. For (a):

$$S_a(p) = \cot^{-1} p = \frac{j}{2} \cdot \ln \frac{p-j}{p+j},$$

for (b):

$$S_b(p) = \frac{1}{\sqrt{p+1}}.$$

The real part of S_a is zero on Re $p = 0$ for $|\text{Im } p| > 1$, while S_b has a branchpoint in the open left half-plane.

(VIII) The introduction of a limiting procedure with H^∞ functions in Theorem 3.1 is essential. Compare this with the following property ([3 p. 96, Problem 9]. Let f_1 and f_2 be analytic functions in the open disc, which have no common zeros in that disc. Then there exist analytic functions g_1 and g_2 in the open disc such that $f_1 g_1 + f_2 g_2 = 1$. This result is of an entirely different nature than the one presented here. For example, e^{-p} and e^{-2p} have no common zero in the open right half-plane, and

$$\tfrac{1}{2} e^{-p} \cdot e^{p} + \tfrac{1}{2} e^{-2p} e^{2p} = 1.$$

However e^{-p} and e^{-2p} are not coprime having a common factor e^{-p}.

6. Some applications of the theory. We will discuss two major applications of the theory, one dealing with the polar structure and the other dealing with the zero structure. Suppose that S is an $n \times n$ contractive transfer function. A unitary embedding for S is a $2n \times 2n$ inner function Σ so that

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{6.1}$$

and $S = \Sigma_{11}$. We want embeddings for S such that (i) Σ_{21} is outer and (ii) Σ is minimal. It will turn out that, in case S is roomy, such an embedding does exist, moreover, has the same degree as S. First, Σ_{21} is obtained by outer spectral factorization. $\Sigma_{21}^* \Sigma_{21} = 1_n - S^* S$, where Σ_{21} is outer. Uniqueness and existence of Σ_{21} is well known [2, Chap. 10]. There is more however.

PROPOSITION 6.1. *If $\hat{\mathcal{M}}_1$ is the nullspace of S, then the nullspace $\hat{\mathcal{M}}_1'$ of Σ_{21} contains $\hat{\mathcal{M}}_1$. The two coincide if and only if S is outer.*

Proof. See Appendix D.

This means, in fact, that the degree of Σ_{21} is smaller than the degree of S, since U_1' divides U_1 and det U_1' divides det U_1.

Consider now the matrix

$$\Sigma_1 = \begin{bmatrix} S \\ \Sigma_{21} \end{bmatrix}. \tag{6.2}$$

If S is roomy, then, by Proposition 6.1, Σ_1 has a full range nullspace $\hat{\mathcal{M}}_1 = U_1 H_n^2$. Of course, Σ_{1d} is also roomy, and there is a $2n \times 2n$ inner matrix U_2 so that $\hat{\mathcal{M}}_2 = U_2 H_{2n}^2$. It turns out that U_2 is precisely a minimal inner embedding for S!

THEOREM 6.1. *Suppose that S is contractive and roomy and Σ_{21} is an outer spectral factor for S. Then a left coprime factorization for*

$$\Sigma_1 = \begin{bmatrix} S \\ \Sigma_{21} \end{bmatrix}$$

is given by

$$\Sigma_1 = \Sigma \begin{bmatrix} 1_n \\ 0_n \end{bmatrix}, \tag{6.3}$$

whereby Σ is a minimal embedding for S (with outer Σ_{21}).

Proof. The proof is based on the foregoing theory and is given in Appendix D.

The zero theory, on the other hand, also provides for a similar theorem, but in the context of J-unitary spaces. This allows for a mechanism to deal with $j\omega$ zeros as well. The theorem to be introduced now, produces an interesting side result: a mechanism to compute a spectral factor for S as well. Let

$$J = \begin{bmatrix} 1_n & \\ & -1_n \end{bmatrix}; \tag{6.4}$$

then

$$\theta_1 = \begin{bmatrix} \Sigma_{21}^{-1} \\ S\Sigma_{21}^{-1} \end{bmatrix} \tag{6.5}$$

is J-expansive in OLP, e.g.,

$$\tilde{\theta}_1 J \theta_1 - 1_n \geqq 0 \quad \text{for } p \in \text{ORP}. \tag{6.6}$$

We will say that a $2n \times 2n$ Θ is J-unitary, if $\tilde{\Theta} J \Theta - J = 0$ for $p \in \mathbb{I}$. Suppose Σ_{21} not identically singular; then to the unitary

$$\Sigma = \begin{bmatrix} S & \Sigma_{21} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{6.7}$$

there corresponds a J-unitary

$$\Theta = \begin{bmatrix} \Sigma_{21}^{-1} & -\Sigma_{21}^{-1} S \\ S\Sigma_{21}^{-1} & S - S\Sigma_{21}^{-1}\Sigma_{22} \end{bmatrix}. \tag{6.8}$$

If Σ is analytic, then Θ is moreover J-expansive in ORP. Such a Θ will be called "passive". Clearly,

$$\theta_1 = \Theta \begin{bmatrix} 1_n \\ 0_n \end{bmatrix}, \tag{6.9}$$

and (6.9) is a "coprime" factorization for θ_1 in a sense analogous to (6.3), but with unitary Σ replaced by J-unitary Θ corresponding to it. It should be noted at this point that the singularities of Θ exhibit the zero structure of Σ_{21} (a roomy outer contraction). The structure of Θ is well known [19]. The interesting point, however, is that Θ can be obtained directly from S, without reference to Σ_{21}.

THEOREM 6.2. *Suppose S is a roomy contraction, and let*

$$\theta_1 = \begin{bmatrix} (1_n - S_* S)^{-1} \\ S(1_n - S_* S)^{-1} \end{bmatrix}. \tag{6.10}$$

Then there exists a unique factorization

$$\theta_1 = \Theta \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}, \tag{6.11}$$

whereby

(i) $\begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$ is J-contractive in ORP, and Θ is passive, J-unitary.

(ii) Θ and $\begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$ are coprime in the sense that, for any factorization

$$\theta_1 = \Theta_1 \begin{bmatrix} A_1' \\ A_2' \end{bmatrix}$$

with $\begin{bmatrix} A_1' \\ A_2' \end{bmatrix}$ J-expansive in the OLP, we have that $\Theta_1 = \Theta\Theta_3$, with Θ_3 passive and J-unitary.

(iii) Θ is the passive, J-unitary matrix corresponding to the minimal embedding Σ with outer Σ_{21}.

Note. Θ is unique except for a trivial constant J-unitary factor on the right.

Proof. The proof is in Appendix D.

Theorem 6.2 shows several things: (i) the spectral factor Σ_{21} for S can be obtained by coprime factorization of the matrix (6.10). This produces a novel algorithm for coprime factorization and is exploited in [22]; (ii) the zero structure of the spectral factor Σ_{21} is completely contained in the product structure of Θ [19]; (iii) the minimal embedding for S can be obtained simultaneously through coprime factorization of θ_1. Theorems 6.1 and 6.2 are striking examples of the strength of coprime factorization techniques in this context. Some further applications in the same vein can be found in [23], [24], [25].

7. Discussion. The present theory, which the author believes is quite complete as far as coprime factorization and polar behavior is concerned, has been inspired, on the one hand, by systems ideas like polynomial coprime factorization [12], [13], input–output state space description as developed by Arbib, Kalman and others [20], [10], and, on the other hand, by invariant subspace theory as developed in Helson [2] and Nagy–Foias [6]. The first attempts at this kind of description appeared in [26] where the important synthesis theorem (inner embedding theorem) is formulated, which was later generalized in [16].

The author believes that the present paper contains several new mathematical and systemic results: (i) Theorem 3.1 on the analytical characterization of left inner coprime matrices is new; (ii) Theorems 3.2, 3.3 and 3.5 on the existence, the uniqueness and the interpretation as pseudo-analytical continuations of coprime

factorizations for bounded L^2 systems seem to have been deduced independently in [22] and in [17]; (iii) Theorem 4.3 is well known in other contexts and has been used also in systems theory, in connection with the construction of shift models for bounded input–output functions [20], [27], [28]. Its salient feature is that, to find the spectral properties of the system one must look to the spectral properties of a restricted shift acting in a state space or zero space rather than to the spectral properties of the system function itself. The idea to do so is probably due to [29] in a different context; (iv) the idea to use the determinant of an inner function as measure for system complexity dates back to the early days of network theory: Propositions 4.1–4.2 merely adapt that idea to the present context; (v) propositions 4.3, 4.4, 4.5, 4.6 and Theorem 4.4 on the zero structure are new; (vi) Theorems 6.1 and 6.2 are new also [25] and show the importance of coprime factorization in the context of cascade synthesis.

The terminology "roomy systems" has been used first in [26]. In [27] the notion of roominess for a system having a representation $[A, B, C, D]$ (most systems don't) appears as "noncyclicity" in the sense of [15]. This notion is more a property of some vectors in a representation than of the system itself and is akin, but not identical, to the notion of cyclicity inherited from module theory (see [10, Chap. 10]). In fact, as used by Baras, cyclicity refers to the input space, while in Kalman it refers to the state space. Kamen [31] has extensively developed structure theories for infinite systems in the case of distribution spaces. To obtain a valuable structure theory, he is forced to restrict the theory to so called "torsion-systems", where the notion of torsion also comes from module theory. "Roominess" is very akin to "torsion", as discussed in [11], and is, in fact, a generalization of the term for non-Noetherian modules. It seems that the use of the term "roomy" (justified by its interpretation that a system is roomy, roughly, when its state space does not fill the input space) is justified to avoid confusion with "torsion" and "cyclicity". Torsion refers to a module, cyclicity to a vector (or set of vectors) in a space and roominess to a system.

This paper was situated purely on the input–output level. We did not discuss any realization theory [20], [27], [28] although the results have some bearing on it. A wealth of information can be deduced from the input–output map alone, without any reference to a realization. More specifically, it follows from the theory that the system has a lossless cascade realization if and only if it is roomy, and that the cascade can be obtained by means of a coprime factorization on an appropriate matrix (Theorems 6.1 and 6.2). Also, polar and zero-degree properties can be deduced directly from input–output considerations without reference to a realization. Of course, it can be argued that there is a subjacent hidden shift realization, but there is no need for explicitation. The only successful attempts to date to deal with zeros on the imaginary axis are through J-unitary factors (Theorem 6.2), and there is still work to be done on the connection between zero and polar degree characteristics.

The theory as presented produces several secondary facts of interest: there is no L^2-system with finite nullspace different from zero; the notion of roominess governs the ability of finding analytic complementary range functions; the degree theory can be extended to nonfinite systems. It is hoped that many more results will follow from this, e.g., in synthesis theory and stability theory.

Appendix A.

Proof of Proposition 2.1 *and Corollary* 2.1. First, given a basis $\{e_j\}$ of $\mathbb{C}^m$ and an appropriate set of projections $\{P_j\}$ and a basis $\{f_k^*\}$ of $\mathbb{C}^n$, we have that $\mathcal{S}$ is roomy if and only if every one input-one output system $\mathcal{S}_{ik} = f_k^* \mathcal{S} P_i$ is. Indeed, suppose that $\mathcal{S}$ is roomy; then $\hat{\mathcal{M}}_1 = UH_m^2$ and (det U) $\cdot H_m^2 \subset \hat{\mathcal{M}}_1$, so that $A_j =$ (det U) $\cdot e_j \in \hat{\mathcal{M}}_1$, and hence $f_k^* \mathcal{S} P_j a_j$ with $a_j = \bar{\mathcal{F}}^{-1} A_j$ zero for $t > 0$. Conversely, suppose that all $\mathcal{S}_{jk} = f_k^* \mathcal{S} P_j$ are roomy. Hence $(\hat{\mathcal{M}}_1)_{jk} = \phi_{jk} H^2$ for some inner ϕ_{jk}. It follows that $\phi = \Pi_{j,k} \phi_{jk}$ is such that $\phi \cdot H^2_{\mathbb{C}^m} \subset \hat{\mathcal{M}}_1$.

Next by the former paragraph, we have reduced both criteria of Proposition 2.1 to the same criterion.

Appendix B.

Proof of Proposition 3.1. We have $\mathcal{M}_A \subset \overline{\mathcal{M}_A \vee \mathcal{M}_B}$ and $\mathcal{M}_B \subset \overline{\mathcal{M}_A \vee \mathcal{M}_B}$. Hence $U^{-1}A$ and $U^{-1}B$ have H^∞ entries, and $A = UA_1$, $B = UB_1$ so that U is a CLID. If V is another CLID, we have that $VH_k^2 \subset \overline{\mathcal{M}_A \vee \mathcal{M}_B}$ so that $U^{-1}V$ is analytic, and hence there exist W such that $V = UW$.

Proof of Theorem 3.1. The "if" portion is trivial. As to the "only if" part, let $\overline{\mathcal{M}_A \vee \mathcal{M}_B} = H_n^2$. Let $\{e_l\}$ be a set of basis vectors in $\mathbb{C}^n$, and let us first prove the property for the Hardy spaces of the unit circle. Consider the sets $e_l + \mathcal{M}_A$ and $\mathcal{M}_B$ in H_n^2. They are closed convex sets. Let d be their distance. I claim that $d = 0$. Suppose on the contrary $d > 0$. Let U_l be a neighborhood of e_l such that its diameter is smaller than d. Then, it follows from the hypothesis $d > 0$ that U_l does not interest $\mathcal{M}_A \vee \mathcal{M}_B$ for else one would have an $x = f_1 + f_2 \in U_l$ so that $e_l - f_1 \in e_l + \mathcal{M}_A$, $f_2 \in \mathcal{M}_2$ and $d \leqq d(e_l - f_1, f_2) = \|e_l - f_1 - f_2\| = \|e_l - x\| < d$. Hence $d = 0$. Now $e_l + AH_k^\infty$ and BH_m^∞ are dense in $e_l + AH_m^2$ so that $d(e_l + AH_k^\infty, BH_m^\infty) = 0$. Hence there exist sequences $-F_i^{(l)}$ and $G_i^{(l)}$ in H_k^∞ and H_m^∞ such that

$$\lim_{i \to \infty} [AF_i^{(l)} + BG_i^{(l)}] = e_l \tag{B.1}$$

for the $L^2_{\mathbb{C}^n}$ topology of the unit circle. Let $M_i = [F_i^{(l)}]$ and $N_i = [G_i^{(l)}]$ be matrices with lth column $F_i^{(l)}$, respectively, $G_i^{(l)}$, then we have

$$\lim_{i \to \infty} (AM_i + BN_i) = I_n. \tag{B.2}$$

A conformal transformation $z \to (p-1)/(p+1)$ from the unit circle to the right half-plane establishes the result.

Proof of Theorem 3.2. Suppose $T = U_1^{-1} \Delta_1 = U_2^{-1} \Delta_2$, where U_1, Δ_1 and U_2, Δ_2 are left coprime. We have, by Theorem 3.1, the existence of sequences M_i and N_i of analytic matrices such that

$$\lim_{i \to \infty} [U_2 M_i + \Delta_2 N_i] = 1_n. \tag{B.3}$$

Premultiplying this with $U_1 U_2^*$ we get

$$\lim_{i \to \infty} [U_1 M_i + U_1 U_2^* \Delta_2 u_i] = U_1 U_2^* \tag{B.4}$$

or

$$\lim_{i\to\infty}[U_1M_i+\Delta_1N_i]=U_1U_2^*. \tag{B.5}$$

Every column of $U_1M_i+\Delta_1N_i$ is in $H_n^2(d\omega/(1+\omega^2))$ for all i, hence the limit also. It follows that $U_1U_2^*$ is analytic. Likewise, $U_2U_1^*$ is analytic, and hence $U_1U_2^*$ is constant. Hence $U_2U_1^*=U$, a constant unitary matrix.

Proof of Theorem 3.3. For any $G\in H_m^2$, we have $U_1G\in\hat{\mathcal{M}}_1$, and hence $\bar{\mathcal{F}}^{-1}(U_1G)\in\mathcal{M}_1$. Thus $\mathcal{S}\bar{\mathcal{F}}^{-1}(U_1G)\in L^2_{\mathbb{C}^n}(-\infty,0])$ and $\bar{\mathcal{F}}\mathcal{S}\bar{\mathcal{F}}^{-1}(U_1G)\in H_n^2$. Hence (taking Fourier transform) we have

$$S(-j\omega)U_1(j\omega)G\in H_n^2 \quad \text{for all } G\in H_m^2. \tag{B.6}$$

Hence $S(-j\omega)U_1=\Delta_1$ is analytic and $S^*=[U_1^*(-j\omega)]^{-1}\,\Delta_1^*(-j\omega)$. To show that $U_1^*(-j\omega)$ and $\Delta_1^*(-j\omega)$ are left coprime, suppose that they are not. Then there exists an inner W such that:

$$U_2^*(-j\omega)=[W^*(-j\omega)]^{-1}U_1^*(-j\omega) \quad \text{and} \quad \Delta_2^*(-j\omega)=[W^*(-j\omega)]^{-1}\,\Delta_1^*(-j\omega) \tag{B.7}$$

are analytic. Then $S^*=[U_2^*(-j\omega)]^{-1}\,\Delta_2^*(-j\omega)$. It follows that for all $G\in H_m^2$, we have that $S(-j\omega)U_2$ is analytic. Hence $U_2H_m^2\subset UH_m^2$. Hence $U_2=U_1$.

Proof of Theorem 3.5. *Let* χ be the minimal inner annihilator of $\hat{\mathcal{H}}_1$, i.e., the smallest inner function such that $S(-j\omega)\chi$ is analytic (for a complete definition see [11]). Then $\chi^*(-j\omega)$ is the minimal inner annihilator of the state space $\hat{\mathcal{H}}_2$ of $\mathcal{S}_d$, and we have

$$S(j\omega)=\chi^{-1}(-j\omega)\,\Delta(-j\omega), \qquad \text{where } \Delta \text{ is analytic.}$$

Also, of course, $S^*(-j\omega)=[\chi^*(j\omega)]^{-1}\,\Delta^*(j\omega)$. These factorizations are not coprime, and hence there exist inner A_1 and A_2 such that

$$\begin{aligned}\Delta^*(-j\omega)&=A_1(j\omega)\,\Delta_1^*(-j\omega),\\ \chi^*(-j\omega)1_m&=A_1(j\omega)U_1^*(-j\omega),\end{aligned} \tag{B.8}$$

$$\begin{aligned}\Delta(j\omega)&=A_2(j\omega)\,\Delta_2^*(-j\omega),\\ \chi(j\omega)1_n&=A_2(j\omega)U_2^*(-j\omega),\end{aligned} \tag{B.9}$$

or else

$$[\chi^*(-j\omega)1_m,\ \Delta^*(-j\omega)\ \triangleq A_1P_1, \tag{B.10}$$

$$[\chi(j\omega)1_n,\ \Delta(j\omega)]\triangleq A_2P_2 \tag{B.11}$$

with

$$\begin{aligned}P_1&=[U_1^*(-j\omega),\ \Delta_1^*(-j\omega)],\\ P_2&=[U_2^*(-j\omega),\ \Delta_2^*(-j\omega)].\end{aligned} \tag{B.12}$$

The submatrices of P_1 and P_2 are coprime by construction, and hence P_1 and P_2 are outer [6, p. 190]. Hence (B.10) and (B.11) are just inner–outer decompositions. I claim that $\det A_1$ ($\det A_2$) is the greatest common inner divisor of all the minors of largest dimension in the matrices (B.10) and (B.11). First, every minor

of largest dimension of the left hand side has det A_1 (det A_2) as an inner factor by the Binet–Cauchy theorem; hence det A_i $(i=1,2)$ divides the minors of the right-hand sides. Moreover there are, by Theorem 3.2, analytic M_i and N_i such that (for an appropriate topology),

$$\lim_{i\to\infty} P_1\begin{bmatrix} M_i \\ N_i \end{bmatrix} = 1_m. \tag{B.13}$$

Hence,

$$\lim_{i\to\infty} \det P_1\begin{bmatrix} M_i \\ N_i \end{bmatrix} = 1. \tag{B.14}$$

Now, the limit on the left-hand side of (B.14) can be expressed homogeneously in terms of the minors of highest rank in P_1.

Suppose these minors have a common inner factor ϕ. Then each det $P_1\begin{bmatrix} M_i \\ N_i \end{bmatrix}$ lies in ϕH^2 and hence the limit as well.

This contradiction proves that det A_1 (det A_2) is the GCID of all the minors of largest dimension in the matrices (B.10), (B.11). But these are in fact the greatest common inner divisors of the following collection: (suppose for definiteness $n \geqq m$):

(for (B.10))

$$[\chi^*(-j\omega)]^m,$$

$$[\text{Minors of order 1 in } \Delta^*(-j\omega)]\cdot[\chi^*(-j\omega)]^{m-1}, \tag{B.15}$$

$$\text{Minors of order } m \text{ in } \Delta^*(-j\omega);$$

(for (B.11))

$$[\chi(j\omega)]^n,$$

$$[\text{Minors of order 1 in } \Delta(j\omega)]\cdot[\chi(j\omega)]^{n-1}, \tag{B.16}$$

$$[\text{Minors of order } m \text{ in } \Delta(j\omega)]\cdot[\chi(j\omega)]^{n-m};$$

Hence with det $A_1 = \phi(j\omega)$, we have

$$\det A_2 = \chi(j\omega)^{n-m}\phi^*(-j\omega) \tag{B.17}$$

and

$$\det U_1(-j\omega) = \det U_2^*(j\omega) = \left[\frac{\chi(-j\omega)}{\phi^*(-j\omega)}\right]^m. \tag{B.18}$$

Appendix C.

Proof of Proposition 4.1. The property is obvious for constant unitary matrices. Next, $\delta(U)$ depends only on the entries in U, and because of (4.2) only on an antisymmetric form of the entries. Hence it depends only on [det U]. Let

$\delta(U) = f(\det U)$. For all inner ϕ_1, ϕ_2, we have

$$f(\phi_1 \cdot \phi_2) = f(\phi_1) \cdot f(\phi_2)$$

by (5.2) so that

$$f(\det U) = [\det U]^k$$

for some k which must be a positive integer if functions of the type

$$1_n - \frac{2\alpha_0 u\tilde{u}}{p + p_0^*}, \quad \alpha_0 = \operatorname{Re} p_0, \quad \tilde{u}u = 1, \tag{C.1}$$

are to have an analytic $\delta(\cdot)$. By (4.3) clearly $k = 1$.

Proof of Proposition 4.2. We will show that $\dim \hat{\mathscr{H}}_1 = \deg \phi$, where $\hat{\mathscr{H}}_1 = [U_1 H_m^2]^\perp$ and $\phi = \det U_1$. For $U \in \mathscr{T}_m$, the map $U \to \dim (UH_m^2)^\perp$ is a monoid homomorphism to the additive monoid of nonnegative integers. Indeed, let $\hat{\mathscr{M}}_{12} = (U_1 U_2 H_m^2)$ and $\hat{\mathscr{H}}_{12} = (\hat{\mathscr{M}}_{12})^\perp$. Let $\hat{\mathscr{M}}_i = U_i H_m^2$ and $\hat{\mathscr{H}}_i = \hat{\mathscr{M}}_i^\perp$ $(i = 1, 2)$; then $\chi : \hat{\mathscr{M}}_1 \to H_m^2 : f_1 \to U_1^{-1} f$ is a Hilbert space isomorphism mapping $\hat{\mathscr{M}}_1 \ominus \hat{\mathscr{M}}_{12}$ onto $\hat{\mathscr{H}}_2$. Now $\hat{\mathscr{H}}_2$ is finite-dimensional (by hypothesis), and hence $\hat{\mathscr{M}}_1 \ominus \hat{\mathscr{M}}_{12}$ also with the same dimension. It follows, for $\hat{\mathscr{H}}_{12} = \hat{\mathscr{H}}_1 \ominus [\hat{\mathscr{M}}_1 \ominus \hat{\mathscr{M}}_{12}]$, that $\dim \hat{\mathscr{H}}_{12} = \dim \hat{\mathscr{H}}_1 + \dim \hat{\mathscr{H}}_2$.

Next, the homomorphism $U \to \delta(U)$ coincides with it on (i) the constant unitary matrices and (ii) the unitary matrices of type (C.1). It is known [8] that elements of the form (C.1) generate the rational unitary matrices as a multiplicative monoid. On the other hand, if $\hat{\mathscr{H}}_1$ is finite-dimensional, then all one-dimensional systems $\mathscr{S}_{ij} = f_j^* \mathscr{S} P_i$ are finite-dimensional (their state space is in a natural way, embedded in $\hat{\mathscr{H}}_1$) and hence are rational, since their nullspace $\hat{\mathscr{M}}_{ij} = \phi H^2$ can have neither an infinite Blashke product nor a singular part. With all $\mathscr{S}_{ij}$ rational, $\mathscr{S}$ automatically is rational as well.

Proof of Theorem 4.3. We will reduce the proof of Theorem 4.3 to the Theorem 13 of [2]. Let T_τ be the right shift acting in $L^2_{\mathbb{C}^n}$. The generator of T_τ is d/dt, its Fourier transform multiplication by $-j\omega$, and the cogenerator [6, p. 141 ff.] in the H_n^2 space, multiplication by $(-j\omega + 1)(-j\omega - 1)^{-1}$. This amounts to multiplication by $(p-1)(p+1)^{-1}$ in H_n^2. The cogenerator of $\hat{T}_\tau^*$ in H_n^2 is then $[(p-1)(p+1)^{-1}]^*$ acting in H_n^2.

Using the conformal transformation $z = (p-1)/(p+1)$, we have that H_n^2 is isomorphic to the $H_n^2(\mathbb{T})$ space of the unit circle $\mathbb{T}$ by $f(p) \mapsto 2/(1-z) f((1+z)/(1-z))$. Under this isomorphism, the cogenerator becomes $(z)^*$. The invariant subspace $\hat{\mathscr{M}}_2 = U_2 H_n^2$ simply becomes $\hat{\mathscr{M}}_2' = U_2((1+z)/(1-z)) H_n^2(\mathbb{T}) = U_2'(z) H_n^2(\mathbb{T})$. We are now in the situation of Theorem 13 of [10], and conclude that the spectrum of the cogenerator of T_τ^* is given by those complex numbers λ such that $U_2'(\lambda) = U_2((1+\bar{\lambda})/(1-\bar{\lambda}))$ is not invertible, and those λ with $|\lambda| = 1$ such that $U_2((1+z)/(1-z))$ cannot be continued analytically across $z = \bar{\lambda}$. The spectral mapping theorem says that, if λ belongs to the spectrum of the cogenerator, then $(\lambda+1)/(\lambda-1)$ belongs to the spectrum of the generator. Hence, $p_0 \in$ spectrum of the generator if either $-p_0^*$ is in the right half open p plane and $U_2(-p_0^*)$ is not invertible or p_0 lays on the imaginary axis and U cannot be continued across p_0.

Proof of Proposition 4.4. Suppose Δ_1 is roomy, and let $UH_l^2 = \overline{\Delta_1 H_n^2}$. We have to show that U can be augmented to an inner function. Let the $n \times l$ U be partitioned as follows:

$$U = \begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix}, \tag{C.2}$$

where U_{21} is $(n-l)\times l$ and U_{11} is $l \times l$. Let us try to augment U to an inner V supposing U roomy. Then V is of the form

$$V = \begin{bmatrix} U_{11} & U'_{12} \\ U_{21} & U'_{22} \end{bmatrix}. \tag{C.3}$$

U'_{22} can be computed directly by spectral factorization [2, p. 111] $U'_{22}U'^{*}_{22} = 1_{n-1} - U_{21}U^{*}_{21}$. Next U'_{12} can be obtained through

$$U^{*}_{11}U'_{12} + U^{*}_{21}U'_{22} = 0. \tag{C.4}$$

(Note U_{11} can always be chosen nonsingular (a.e.)). We have $U'_{12} = -U^{*-1}_{11}U^{*}_{21}U'_{22}$. Since U_{11} and U_{21} are roomy, we can find a $\phi \in H^{\infty}$ such that $\phi \cdot U^{*-1}_{11}U^{*}_{21}$ is analytic, and hence there exist a minimal $(n-l)\times(n-l)$ inner W such that $U^{*-1}_{11}U^{*}_{21}U'_{22}W$ is analytic. Hence,

$$V\begin{bmatrix} 1_l & 0 \\ 0 & W \end{bmatrix} = \begin{bmatrix} U_{11} & U'_{21}W \\ U_{12} & U'_{22}W \end{bmatrix} \tag{C.5}$$

is analytic and produces an augmentation.

Proof of Theorem 4.4. We have $\overline{\Delta_1(j\omega)U_1^{-1}(j\omega)\hat{\mathcal{M}}_1} = \overline{\Delta_1(j\omega)H_m^2} = V_1H_m^2$. Hence $\hat{\mathcal{N}}_1 = V_1H_m^2$. Next, let $f \in (V_2H_m^2)^{\perp}$ or $V_2^{*}f \in K_m^2$. Then $Sf = U_2\Delta_2'^{*}g$ for some $g \in K_m^2$, and $U_2^{*}Sf \in K_n^2$, so that $Sf \in \hat{\mathcal{M}}_2^{\perp}$. This shows that $(V_2H_m^2)^{\perp} \subset \hat{\mathcal{H}}_2^{\perp}$. Conversely, let $f \in \hat{\mathcal{H}}_2^{\perp}$ so that $Sf \in \hat{\mathcal{M}}_2^{\perp}$ or $U_2^{*}Sf \in K_n^2$. Then there is a $g \in K_n^2$ such that $Sf = U_2g = U_2\Delta_2^{*}f$. It follows that $\Delta_2^{*}f = \Delta_2'V_2^{*}f \in K_n^2$. Now, let $V_2^{*}f = h_1 + h_2$ with $h_1 \in K_m^2$ and $h_2 \in H_m^2$. Then $g = \Delta_2'^{*}h_1 + \Delta_2'^{*}h_2$ so that $\Delta_2'^{*}h_2 \in K_n^2$. For all $g \in H_n^2$ we have $(\Delta_2'^{*}h_2, g) = 0$ so that $(h_2, \Delta_2'g) = 0$. Since $\Delta_2'H_n^2$ is dense in H_m^2, we have $h_2 = 0$. It follows that $V_2^{*}f \in K_n^2$ and $f \in (V_2H_m^2)^{\perp}$. This proves that $\hat{\mathcal{N}}_2^{\perp} \subset (V_2H_m^2)^{\perp}$. The remainder of the theorem is based on Theorem 3.5 and the fact that, for $m = n$, we have that $\det(\Delta_1'(-j\omega))$ and $\det(\Delta_2'^{*}(j\omega))$ are outer. Hence in $U_2^{*}(-j\omega)W(j\omega)\Delta_1'(j\omega) = \Delta_2'^{*}(-j\omega)V_2^{*}(-j\omega)U_1(j\omega)$ one must have $\det[U_2^{*}(-j\omega)W(j\omega)] = \det[V_2^{*}(-j\omega)\cdot U_1(j\omega)]$, this being the inner part of the whole expression. It is not true in general that (i) $\delta(\tilde{V}_2(p^{*})) = \mathrm{GCID}[n \times n$ minors of $V_1]$ or that (ii) $\delta(\tilde{V}_2(p^{*})) = \delta(W(-p))$.

Proof of Proposition 4.6. First, if $f \in \hat{\mathcal{M}}_1$, then $S(-j\omega)f$ is analytic. Then $S_1(-j\omega)f = W^{-1}(-j\omega)S(-j\omega)f$ is analytic a fortiori. It follows that $\mathcal{M}_1 \subset \mathcal{M}_1'$. Conversely, suppose there is an $f \in \mathcal{M}_1'$. Then $W^{*}(-j\omega)S(-j\omega)f$ is analytic. Let $S(-j\omega)f = g_1 + g_2$, $g_1 \in K_n^2$, $g_2 \in H_n^2$. We have $W^{*}(-j\omega)S(-j\omega)f = W^{*}(-j\omega)g_1 + W^{*}(-j\omega)g_2$. Hence $W^{*}(-j\omega)g_1 \in H_n^2$. Also, $g_1 + g_2 = U_2(-j\omega)\Delta_2^{*}(-j\omega)f$, or $\Delta_2^{*}(-j\omega)f = U_2^{*}(-j\omega)g_1 + U_2^{*}(-j\omega)g_1$. Hence, $U_2^{*}(-j\omega)g_1 \in H_n^2$. It follows that $g_1(-j\omega) \in H_n^2$ and $g_1(-j\omega) \perp \mathcal{F} = W \cdot H_n^2$ as well as $g_1(-j\omega) \perp \mathcal{M}_2 = U_2H_n^2$. Since $\mathcal{M}_2 \vee \mathcal{F} = H_n^2$, $g_1 = 0$, and $S(-j\omega)f$ is analytic. Hence $\mathcal{M}_1' \subset \mathcal{M}_1$.

Appendix D.

Proof of Proposition 6.1. We show that if $F \in \hat{\mathcal{M}}_1$, then also $F \in \hat{\mathcal{M}}_1'$. Or equivalently, if $S(-j\omega)F$ is analytic, then $\Sigma_{21}(-j\omega)F$ is. If $S(-j\omega)F$ is analytic, then so is $\tilde{S}(-j\omega)S(-j\omega)F$ and following, of course, also $\tilde{\Sigma}_{21}(-j\omega)\Sigma_{21}(-j\omega)F$. We show that, if $\tilde{\Sigma}_{21}(-j\omega)G$ is analytic, then so is G. Suppose not. Then there is a G_1 such that $\tilde{\Sigma}_{21}(-j\omega)G_1$ is analytic with G_1 conjugate analytic. It follows that $G_1(-j\omega) \perp \Sigma_{21}H_n^2$, and since $G_1(-j\omega)$ is now analytic and Σ_{21} outer, $G_1 = 0$. This shows that $\mathcal{M}_1 \subset \mathcal{M}_1'$. If S is outer, then also $\mathcal{M}_1' \subset \mathcal{M}_1$.

Proof of Theorem 6.1. Let $U \cdot [A_{11}^T, A_{21}^T]^T$ be a left coprime factorization for Σ_1 with U inner (lossless) of dimension $2n \times 2n$ and $[A_{11}^T, A_{21}^T]^T$ conjugate analytic. Since U and $[A_{11}^*, A_{21}^*]$ are right coprime, we have by Theorem 3.1, analytic matrices M_i and N_i such that

$$\lim_{i\to\infty} \{M_i[A_{11}^*, A_{21}^*] + N_iU\} = 1_{2n}. \tag{D.1}$$

Postmultiplying this with $[A_{11}^T, A_{21}^T]^T$ we get

$$\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = \lim_{i\to\infty} \{M_i + N_i\Sigma_1\}, \tag{D.2}$$

since

$$\begin{aligned} [A_{11}^*, A_{21}^*]\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} &= [A_{11}^*, A_{21}^*]U_1^*U_1\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} \\ &= \Sigma_1^*\Sigma_1 = 1_n. \end{aligned} \tag{D.3}$$

It follows that $[A_{11}^T, A_{21}^T]^T$ is both analytic and conjugate analytic, and thus constant. Since it is of rank n, there is a constant unitary matrix C with

$$\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = C\begin{bmatrix} 1_n \\ 0_n \end{bmatrix} \tag{D.4}$$

and $UC \cdot [1_n, 0_n]^T$ is clearly also a left coprime factorization for Σ_1. It is clear that UC has the form

$$UC = [\Sigma_1, \Sigma_2] = \begin{bmatrix} S & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \Sigma \tag{D.5}$$

and that $\deg \Sigma = \deg S$ because Σ and Σ_{1d} have exactly the same natural state space.

Proof of Theorem 6.2. For brevity, we will call a $2n \times n$ matrix A "antipassive" if $\tilde{A}JA - 1_n \geqq 0$ in OLP.

We have

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} (1_n - S_*S)^{-1} \\ S(1_n - S_*S)^{-1} \end{bmatrix} = \begin{bmatrix} \Sigma_{21}^{-1} \\ S\Sigma_{21}^{-1} \end{bmatrix}\Sigma_{21*}^{-1}, \tag{D.6}$$

where Σ_{21} is the cofactor for S. It is clear that

$$\Theta^{-1}\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} \Sigma_{21*}^{-1} \\ 0_n \end{bmatrix} \tag{D.7}$$

is antipassive so that the Θ obtained from the minimal embedding with outer Σ_{21} satisfies the requirements for being a factor. It will be *the* cofactor if we prove that, for Θ_1 such that

$$\Theta_1^{-1}\begin{bmatrix}A_1\\A_2\end{bmatrix}=\begin{bmatrix}A_1'\\A_2'\end{bmatrix} \tag{D.8}$$

is antipassive and whose corresponding $(\Sigma_1)_{21}$ is outer, there is a J-unitary passive Θ_3 such that

$$\Theta_1=\Theta\cdot\Theta_3. \tag{D.9}$$

Of course, $\Theta_3=\Theta^{-1}\Theta_1$ is well-defined; the question is whether or not it is passive. Clearly we have

$$\begin{bmatrix}A_1'\\A_2'\end{bmatrix}=\Theta_3^{-1}\begin{bmatrix}\Sigma_{21*}^{-1}\\0_n\end{bmatrix}=J\Theta_{3*}J\begin{bmatrix}\Sigma_{21*}^{-1}\\0_n\end{bmatrix},$$

so that, denoting

$$\Theta_3=\begin{bmatrix}\theta_{11}&\theta_{12}\\\theta_{21}&\theta_{22}\end{bmatrix}$$

and the corresponding unitary

$$\Sigma_3=\begin{bmatrix}\sigma_{11}&\sigma_{12}\\\sigma_{21}&\sigma_{22}\end{bmatrix},$$

we have

$$\begin{bmatrix}A_1'\\A_2'\end{bmatrix}=\begin{bmatrix}\sigma_{21*}^{-1}&\Sigma_{21*}^{-1}\\\sigma_{22*}&\sigma_{21*}^{-1}\end{bmatrix}\begin{matrix}\\\Sigma_{21*}^{-1}\end{matrix}. \tag{D.10}$$

Since (D.10) is supposed to be antipassive, it follows that $\sigma_{21}\Sigma_{21}$ has to be analytic. Since Σ_{21} is outer, σ_{21} has to be analytic as well. Also, σ_{22} has to be analytic because of the antipassivity of (D.10). From the fact that $\Theta_1=\Theta\Theta_3$, we deduce

$$\Sigma_1=\begin{bmatrix}\Sigma_{11}+\Sigma_{12}\sigma_{11}(1_n-\Sigma_{22}\sigma_{11})^{-1}\Sigma_{21}&\Sigma_{12}(1_n-\sigma_{11}\Sigma_{22})^{-1}\sigma_{12}\\\sigma_{21}(1_n-\Sigma_{22}\sigma_{11})^{-1}\Sigma_{22}&\sigma_{22}+\sigma_{21}\Sigma_{22}(1_n-\sigma_{11}\Sigma_{22})^{-1}\sigma_{12}\end{bmatrix}. \tag{D.11}$$

Σ_1 is analytic and such that $(\Sigma_1)_{21}$ is outer, so that we can deduce that

(i) $\sigma_{21}(1_n-\Sigma_{22}\sigma_{11})^{-1}\Sigma_{21}$ is analytic and outer,

(ii) $\sigma_{22}(1_n-\Sigma_{22}\sigma_{11})^{-1}\Sigma_{22}\sigma_{12}$ is analytic,

(iii) $\Sigma_{12}\sigma_{11}(1_n-\Sigma_{22}\sigma_{11})^{-1}$ is analytic,

(iv) $\Sigma_{12}(1_n-\sigma_{11}\Sigma_{22})^{-1}\sigma_{12}$ is analytic.

At this point we have to use two properties of outer functions (A, B and C are $n\times n$ square matrices).

(a) Suppose $A=BC$ is analytic and either B or C is outer; then B and C are analytic.

Proof. Suppose that B is outer. Then by Theorem 3.1, there is a sequence M_i of analytic functions with $\lim_{i\to\infty}M_iB=1_n$. It follows that $C=\lim_{i\to\infty}M_iA$ must be analytic.

(b) Suppose that B and C are analytic matrices and that B and C are outer. Then BC has to be outer.

Proof. A matrix A is outer if and only if its determinant is outer.

From (i) and (b) it follows that $\sigma_{21}(1_n - \Sigma_{22}\sigma_{11})^{-1}$ is analytic and outer, so that

$$\sigma_{21} = T(1_n - \Sigma_{22}\sigma_{11}) \tag{D.12}$$

with T outer. It follows that $1_n - \Sigma_{22}\sigma_{11}$ and also $\Sigma_{22}\sigma_{11}$ are analytic. From (iii) we have that $\Sigma_{12}\sigma_{11}$ is analytic as well. From (ii) $\Sigma_{22}\sigma_{12}$ is analytic because $\sigma_{21}(1_n - \Sigma_{22}\sigma_{11})^{-1}$ is outer, and we have that

$$\Sigma_{12}\sigma_{12} = \Sigma_{12}(1_n - \sigma_{11}\Sigma_{22})^{-1}\sigma_{12} - \Sigma_{12}\sigma_{11}(1_n - \Sigma_{22}\sigma_{11})^{-1}\Sigma_{22}\sigma_{12}$$

is analytic because of (ii) and (iv).

Now, Σ_{12} and Σ_{22} belong to a minimal embedding of S, so they have to be right coprime, for otherwise $\Sigma_{12} = \Sigma'_{12}U$, $\Sigma_{22} = \Sigma'_{22}U$ with Σ'_{12} and Σ'_{22} analytic and

$$\begin{bmatrix} S & \Sigma'_{12} \\ \Sigma_{21} & \Sigma'_{22} \end{bmatrix}$$

would be a minimal embedding with smaller determinant and hence with smaller (generalized) degree. Thus, there are M_i and N_i such that

$$\lim_{i\to\infty} (M_i\Sigma_{12} + N_i\Sigma_{22}) = 1_n,$$

and it follows that:

$$\begin{aligned} \sigma_{11} &= \lim_{i\to\infty} (M_i\Sigma_{12}\sigma_{11} + N_i\Sigma_{22}\sigma_{11}), \\ \sigma_{12} &= \lim_{i\to\infty} (M_i\Sigma_{12}\sigma_{12} + N_i\Sigma_{22}\sigma_{12}) \end{aligned} \tag{D.13}$$

are analytic. This proves the theorem.

REFERENCES

[1] S. BOCHNER AND K. CHANDRASEKHARAN, *Fourier transform*, Annals of Mathematics Studies, Princeton University Press, Princeton, N.J., 1949.

[2] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.

[3] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, N.J., 1962.

[4] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[5] P. L. DÜREN, *Theory of H^p Spaces*, Academic Press, New York, 1970.

[6] B. SZ-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

[7] R. W. NEWCOMB, *Linear Multiport Synthesis*, McGraw-Hill, New York, 1966.

[8] P. DEWILDE, *Cascade scattering matrix synthesis*, Tech. Rep. 6560-21, Information Systems Lab., Stanford University, Stanford, Calif., 1970.

[9] V. BELEVITCH, *Classical Network Synthesis*, Holden-Day, San Francisco, 1968.

[10] R. E. KALMAN, P. L. FALB AND M. ARBIB, *Topics in Mathematical Systems Theory*, McGraw-Hill, New York, 1968.

[11] P. DEWILDE, *On the finite unitary embedding theorem for lossy scattering matrices*, 1974 European Conference on Circuit Theory and Design, IEE, London, 1974.

[12] V. M. POPOV, *Some properties of the control systems with irreducible matrix transfer functions*, Lecture Notes in Mathematics, Seminar on Differential Equations and Dynamical Systems, Springer-Verlag, New York, 1970, pp. 250–261.
[13] H. H. ROSENBROCK, *State Space and Multivariable Theory*, New York, John Wiley, 1970.
[14] F. M. CALLIER AND C. D. NAHUM, *Necessary and sufficient conditions for the complete controllability and observability of systems in series using the coprime factorization of a rational matrix*, IEEE Trans. Circuits and Systems, 22 (1975), pp. 90–95.
[15] R. G. DOUGLAS, H. S. SHAPIRO AND A. L. SHIELDS, *Cyclic vectors and invariant subspaces for the backward shift operators*, Ann. Inst. Fourier (Grenoble), 20 (1970), pp. 37–76.
[16] R. G. DOUGLAS AND J. W. HELTON, *Inner dilations of analytical matrix functions and Darlington synthesis*, Act. Sci. Math., 34 (1973), pp. 301–310.
[17] P. FUHRMANN, *Factorization theorems for a class of bounded measurable operator valued functions*, Tech. Rep., Div. of Eng. and Appl. Phys., Harvard Univ., Cambridge, Mass., to appear.
[18] C. A. NORDGREN, *On quasi-equivalence of matrices over H^∞*, Acta Sci. Math., 34 (1973), pp. 301–310.
[19] V. P. POTAPOV, *The multiplicative structure of J-contractive matrix functions*, American Mathematical Society Translations, Series II, vol. 15, American Mathematical Society, Providence, R.I., 1960, pp. 131–243.
[20] P. FUHRMANN, *On series and parallel coupling of a class of infinite dimensional systems*, Tech. Rep., Div. of Eng. and Appl. Phys. Harvard Univ., Cambridge, Mass., to appear.
[21] B. MOORE III AND C. A. NORDGREN, *On quasi-equivalence and quasi-simularity*, Acta Sci., Math., 34 (1973), pp. 311–316.
[22] P. DEWILDE, *A novel algorithm for spectral factorization*, Tech. Rep., Dept. of Electrical Engineering, Univ. te Leuven, Louvain, Belgium.
[23] P. DEWILDE AND J. BARAS, *Invariant subspace methods in linear multivariable distributed systems and lumped distributed network synthesis*, IEEE Proceedings, 64 (1976), no. 1, pp. 160–178.
[24] P. DEWILDE, *On the synthesis of networks using coprime factorization techniques*, Tech. Rep., Kath. Univ. te Leuven, Louvain, Belgium.
[25] ———, *Input–output L_2 system theory and scattering matrix synthesis*, Tech. Rep., Kath. Univ. te Leuven, Louvain, Belgium.
[26] ———, *Roomy scattering matrix synthesis*, Tech. Rep., Dept. of Mathematics, Univ. of California at Berkeley, Berkeley, 1971.
[27] J. S. BARAS AND R. W. BROCKET, *H^2-functions and infinite dimensional realization theory*, this Journal, 13 (1975), pp. 221–241.
[28] J. W. HELTON, *Discrete time systems, operator models and scattering theory*, J. Functional Analysis, to appear.
[29] P. D. LAX AND R. S. PHILLIPS, *Scattering Theory*, Academic Press, New York, 1967.
[30] M. A. ARBIB, *A common framework for automatic and control theory*, this Journal, 3 (1965), pp. 206–222.
[31] E. W. KAMEN, *Representation of linear continuous-time systems by spaces of distributions*, Tech. Rep., Inst. de Recherche d'Informatique et d'Automatique, France, 1972.
[32] M. S. LIVSIC, *Operators, Oscillations and Waves, Open Systems*, Amer. Mathematical Society Translations Math. Monogr., vol. 34, 1973.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

RELAXED CONTROLS AND THE CONVERGENCE OF OPTIMAL CONTROL ALGORITHMS*

L. J. WILLIAMSON† AND E. POLAK‡

Abstract. This paper presents a framework for the study of the convergence properties of optimal control algorithms and illustrates its use by means of two examples. The framework consists of an algorithm prototype with a convergence theorem, together with some results in relaxed controls theory.

1. Introduction. Most optimal control algorithms construct a sequence of controls whose corresponding costs form a monotonically decreasing, converging sequence. Because of this, it suffices to require that the sequence of controls and initial states constructed have at least one accumulation point and that any accumulation point of this sequence satisfies an optimality condition, rather than to require that it converges.

In studying the convergence properties of nonlinear programming algorithms, to which the preceding remarks also apply, it is assumed that the sequence of points constructed by the algorithm remains in a compact subset of R^n. This guarantees the existence of an accumulation point. With the exception of penalty function methods (which are not iterative procedures; see, for example [1], [3], [11]), it has been common among inventors of iterative optimal control algorithms to assure that the sequences of controls constructed remain in L_∞-bounded sets, and to show that any L_2-accumulation point satisfies the Pontryagin maximum principle or some relating necessary condition of optimality. (In the absence of constructive, generally applicable necessary and sufficient conditions, one cannot expect proofs of convergence to an optimum.) Unfortunately, there is no mathematical basis for assuming that a sequence of controls in an L_∞-bounded set has an L_2-accumulation point.

The purpose of this paper is to present and illustrate a convergence theory for optimal control algorithms using iteration formulas of the form $u_{i+1} \in A(u_i)$, $i = 0, 1, 2, \cdots$, where the u_i are the successively constructed controls and A is a set-valued iteration function. This class of algorithms includes gradient and gradient projection methods, feasible directions methods, strong variations methods and so forth. (It does not include penalty function type methods whose analysis requires a totally different approach). Our theory does not prove that existing optimal control algorithms always construct controls converging to an optimal control. This is clearly false. Instead, our theory examines the properties of accumulation points of control sequences constructed by optimal control algorithms. In particular, it shows that these accumulation points satisfy some optimality condition for the relaxed problem. The optimality condition satisfied

* Received by the editors February 15, 1974, and in final revised form June 24, 1975. This research was sponsored by the National Aeronautics and Space Administration under Grant NGL-05-003-016, the National Science Foundation under Grant GK-37672, and the U.S. Army Research Office—Durham, under Contract DAHC04-73-C-0025.

† Sandia Laboratories, Livermore, California.

‡ Department of Electrical Engineering and Computer Sciences and Electronics Research Laboratory, University of California, Berkeley, California 94720.

differs from algorithm to algorithm. The theory is based on an extension of results in [9] and on the use of a topology, based on relaxed controls [14], [12], [12a], which ensures that accumulation points always exist for L_∞-bounded sequences.

The theory found in Young [14], with some minor modifications, seems to be the most appropriate one for analyzing optimal control algorithms. There were two reasons for the modifications. The first is that Young specifies a priori a fixed set U in which all controls must take their value. This is extremely inconvenient in analyzing algorithms for problems without control constraints. We have therefore changed a number of definitions to make them independent of such a set U. The second reason is that we felt it very important to preserve a connection between the old $(L_2 \cap L_\infty)$ and new convergence results and have, therefore, modified slightly Young's definition of convergence of relaxed controls.

We illustrate the manner in which this new convergence theory is to be used by means of two examples: an analysis of a strong variations algorithm due to Mayne and Polak [7] and of the Pironneau–Polak dual method of feasible directions [8]. The latter, as well as gradient methods, require the development of a special directional derivative. Finally, in Appendix A, we give a short discussion of the use of optimality conditions in the construction of optimization algorithms, and in Appendix B, we establish the relation between the new and the old convergence results.

2. Compactness properties of the relaxed optimal control problem. The algorithms which we are about to discuss solve optimal control problems of the form:

$$\min g_0(\xi, u) \triangleq \int_0^1 L(x(t, \xi, u), u(t), t)\, dt + h_0(x(1, \xi, u)), \tag{1}$$

subject to the constraints

$$\frac{d}{dt}x(t, \xi, u) = f(x(t, \xi, u), u(t), t), \qquad t \in [0, 1] \text{ a.e.}, \tag{2}$$

$$x(0, \xi, u) = \xi, \tag{3}$$

$$g_j(\xi, u) \triangleq h_j(x(1, \xi, u)) \leqq 0, \qquad j = 1, 2, \cdots, p, \tag{4}$$

$$g_j(\xi, u) \triangleq h_j(\xi) \leqq 0, \qquad j = p+1, \cdots, p+q, \tag{5}$$

$$u(t) \in U \subset R^m \qquad \text{for all } t \in [0, 1], \tag{6}$$

where $f : R^n \times R^m \times [0, 1] \to R^n$ and $L : R^n \times R^m \times [0, 1] \to R^1$. The functions g_j, $j = 0, 1, \cdots, p+q$, are real-valued, and u is assumed to be measurable.

The following hypotheses are commonly made, with $T \triangleq [0, 1]$.

Assumption 1. The functions $f : R^n \times R^m \times T \to R^n$ and $L : R^n \times R^m \times T \to R^1$ and their partial derivatives $\partial f/\partial x$, $\partial L/\partial x$ exist and are continuous on $R^n \times R^m \times T$. The functions $h_j : R^n \to R^1$, $j = 0, 1, \cdots, p+q$, and their derivatives $\partial h_j/\partial x$, $j = 0, 1, \cdots, p+q$, exist and are continuous on R^n.

Assumption 2. For each compact $\Omega \subset R^m$, there exists an $M > 0$ such that $\|f(x, u, t)\| \leqq M(\|x\| + 1)$ for all $(x, u, t) \in R^n \times \Omega \times T$ and $\|f(x, u, t) - f(x', u, t)\| \leqq M\|x - x'\|$ for all $x, x' \in R^n$, $u \in \Omega$, $t \in T$.

With the original problem (1)–(6) we associate a relaxed problem, following Young [14], as will be shown after the necessary definitions have been introduced.

As already pointed out in the introduction, the study of optimization algorithms is substantially simplified when a number of definitions used by Young [14] and Warga [12], [12a] are somewhat modified. This is done to avoid the a priori selection of a compact set $U \subset R^m$ such that $u : T \to U$, since an a priori selection of a U contradicts the absence of constraints on $u(t)$ in control unconstrained problems. The reader is therefore cautioned that our definitions differ from those of Young and Warga. However, the following results can be deduced directly from those of Young [14] and Warga [12], [12a] and are presented here, without claims of originality, so as to make the paper readily accessible to the large number of specialists in computational methods who are not familiar with the theory of relaxed controls.

DEFINITION 1. Let V be the set of nonnegative unit measures (probability measures) on R^m and let $T \triangleq [0, 1]$. *A relaxed control* is any function $\mathbf{v}(\cdot) : T \to V$ with the property that for some compact set $U \subset R^m$, the measure $\mathbf{v}(t)$ is wholly concentrated on U for all $t \in T$ (this will be referred to as "$\mathbf{v}(\cdot)$ vanishes outside of U").

Throughout the paper a relaxed control will be denoted by a boldface $\mathbf{u}$ or $\mathbf{v}$ and an ordinary control (measurable function) by an ordinary u or v.

DEFINITION 2. Given a continuous function $\phi(\cdot)$ defined on R^m and a measure $\bar{v} \in V$, we shall write $\phi_r(\bar{v})$ for its integral in the measure $\bar{v}$, i.e., $\phi_r(\bar{v}) \triangleq \int_{R_m} \phi(u)\, d\bar{v}$, whenever that integral is well-defined. More generally, if $\phi(x, u, t)$ is continuous in (x, u, t), the symbol $\phi_r(x, \bar{v}, t)$ denotes, for fixed (x, t), the integral on R^m of $\phi(x, u, t)$ with respect to the probability measure $\bar{v}$, i.e., $\phi_r(x, \bar{v}, t) \triangleq \int_{R^m} \phi(x, u, t)\, d\bar{v}$.

DEFINITION 3. A relaxed control $\mathbf{v}(\cdot)$ will be termed measurable if for every polynomial $p(u)$ in (the components of) u, the function $p_r(\mathbf{v}(t)) \triangleq \int_{R^m} p(u)\, d\mathbf{v}(t)$ of t is measurable.

Remark. From Young [14, p. 290] it follows that if $\mathbf{v}(\cdot)$ is a measurable relaxed control and $g(t, u)$ is a continuous function of (t, u), then the function $g_r(t, \mathbf{v}(t)) \triangleq \int_{R^m} g(t, u)\, d\mathbf{v}(t)$ of t is measurable.

The relaxed problem is obtained from the original problem (1)–(5) by substituting the cost

$$g_0(\xi, \mathbf{v}) \triangleq \int_0^1 L_r(x(t, \xi, \mathbf{v}), \mathbf{v}(t), t)\, dt + h_0(x(1, \xi, \mathbf{v})) \tag{7}$$

for the cost (1), the differential equation

$$\dot{x}(t) = f_r(x(t), \mathbf{v}(t), t) \triangleq \int_{R^m} f(x(t), u, t)\, d\mathbf{v}(t), \tag{8}$$

for the differential equation (2), and the requirement that

$$\mathbf{v}(\cdot) \text{ vanish outside of } U \tag{9}$$

for (6).

We now give an existence and uniqueness theorem for the solution to the relaxed differential equation (8). The proof is found in Young [14, pp. 291–292 and 298] where the theorem is proved under weaker assumptions.

THEOREM 1. *Suppose that Assumptions* 1 *and* 2 *are satisfied. Then for any measurable relaxed control* $\mathbf{v}(\cdot)$, *which vanishes outside some compact set* $U \subset R^m$, *and any initial state* x_0, *there exists an absolutely continuous function* $x(\cdot, x_0, \mathbf{v}) : T \to R^n$ *that is the unique solution to* (8), *satisfying* $x(0, x_0, \mathbf{v}) = x_0$.

In our analysis, in addition to the relaxed optimal control problem, we will also need associated multiplier functions, defined as follows.

DEFINITION 4. For $j = 0, 1, 2, \cdots, p$, let $\lambda_j(\cdot, \xi, \mathbf{v}) : T \to R^n$, denote the solution of

$$-\lambda_j(t, \xi, \mathbf{v}) = \left(\frac{\partial H_j}{\partial x}\right)_r^T (x(t, \xi, \mathbf{v}), \mathbf{v}(t), \lambda_j(t, \xi, \mathbf{v}), t), \tag{10}$$

$$\lambda_j(1, \xi, \mathbf{v}) = \left(\frac{\partial h_j}{\partial x}\right)^T (x(1, \xi, \mathbf{v})), \tag{11}$$

where the superscript T denotes transposition and $H_j : R^n \times R^m \times R^n \times T \to R^1$, $j = 0, 1, \cdots, p$, is defined by

$$H_j(x, u, \lambda, t) \triangleq \lambda^T f(x, u, t) + \delta_{j0} L(x, u, t), \tag{12}$$

where δ_{j0} is the Kronecker delta.

The relaxed optimal control problem leads to two crucial sequential compactness theorems, as we shall shortly see. The first one of these two theorems is due to Young [14], the second one to Warga [12a].

DEFINITION 5. A sequence $\{\mathbf{v}^i(\cdot)\}_{i=0}^{\infty}$ of measurable relaxed controls converges in the sense of control measures (abbreviated i.s.c.m.) to a relaxed control $\bar{\mathbf{v}}(\cdot)$ if for every continuous, real-valued function $g(t, u)$ defined on $T \times R^m$ and every subinterval Δ of T the values $\int_\Delta g_r(t, \mathbf{v}^i(t))\, dt$ converge to $\int_\Delta g_r(t, \bar{\mathbf{v}}(t))\, dt$.

Notation. If $\{\mathbf{v}^i(\cdot)\}$, $i \in K$, converges i.s.c.m. to $\bar{\mathbf{v}}(\cdot)$, we denote that by $\mathbf{v}^i(\cdot) \overset{K}{\to} \bar{\mathbf{v}}(\cdot)$.

The first compactness theorem which we need is proved in Young [14, pp. 301–303].

THEOREM 2. *Let* $\{\mathbf{v}^i(\cdot)\}_{i=0}^{\infty}$ *be a sequence of measurable relaxed controls which vanish outside some fixed compact set U. Then there exists a relaxed control* $\bar{\mathbf{v}}(\cdot)$ *which also vanishes outside of U and a subsequence indexed by a set* $K \subset \{0, 1, 2, \cdots \supset$ *such that* $\mathbf{v}^i(\cdot) \overset{K}{\to} \bar{\mathbf{v}}(\cdot)$.

Notation. Given a sequence of initial states $\{\xi^i\}_{i=0}^{\infty}$ and a sequence of relaxed controls $\{\mathbf{v}^i(\cdot)\}_{i=0}^{\infty}$, we shall denote the corresponding sequences of trajectories and multipliers (determined according to (8) and (3), and (10), (11), respectively) by $\{x^i(\cdot)\}_{i=0}^{\infty}$, $\{\lambda_j^i(\cdot)\}_{i=0}^{\infty}$, $j = 0, 1, \cdots, p$. We shall also use the notation x^u, $x^{\mathbf{u}}$, λ_j^u, $\lambda_j^{\mathbf{u}}$ to denote solutions to (2), (3), (8), (3) and (10), (11) corresponding to a measurable control u or a relaxed control $\mathbf{u}$.[1]

[1] Equation (10) degenerates into an ordinary differential equation when u is an ordinary control.

DEFINITION 6. If $\{(\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i)\}_{i=0}^{\infty}$ is a sequence of initial states, relaxed controls, corresponding trajectories and corresponding multipliers such that $(\xi^i\}$ converges to $\bar{\xi}$, $\{\mathbf{v}^i\}$ converges to $\bar{\mathbf{v}}$ i.s.c.m., $\{x^i\}$ converges to $\bar{x}$ uniformly, and $\{\lambda_j^i\}$ converges to $\bar{\lambda}_j$ uniformly, $j = 0, 1, \cdots, p$, then we denote this by $(\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i) \to (\bar{\xi}, \bar{\mathbf{v}}, \bar{x}, \bar{\lambda}_0, \cdots, \bar{\lambda}_p)$.

DEFINITION 7. $(\bar{\xi}, \bar{\mathbf{v}}, \bar{x}, \bar{\lambda}_0, \cdots, \bar{\lambda}_p)$ is called an *accumulation point* of $\{(\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i)\}_{i=0}^{\infty}$ if there exists a subsequence, indexed by some $K \subset \{0, 1, 2, \cdots\}$ such that $(\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i) \overset{K}{\to} (\bar{\xi}, \bar{\mathbf{v}}, \bar{x}, \bar{\lambda}_0, \cdots, \bar{\lambda}_p)$.

The second compactness theorem will be established as a consequence of the following lemmas.

LEMMA 1. *Let C, U be arbitrary compact sets in R^p, R^m, respectively, and let S be the set of measurable relaxed controls which vanish outside of U. Let g be a continuous function from $R^p \times R^m \times T$ into R^q. Let $Y^i(\cdot)$, $\bar{Y}(\cdot)$ be continuous functions from T into C such that $Y^i(\cdot)$ converges to $\bar{Y}(\cdot)$ uniformly. Let $\{\mathbf{v}^i(\cdot)\}_{i=0}^{\infty}$ be a sequence of relaxed controls that converges i.s.c.m. to a relaxed control $\bar{\mathbf{v}}(\cdot)$. Then for each subinterval Δ of T,*

$$\int_{\Delta} g_r(Y_i(\tau), \mathbf{v}_i(\tau), \tau)\, d\tau \to \int_{\Delta} g_r(\bar{Y}(\tau), \bar{\mathbf{v}}(\tau), \tau)\, d\tau. \tag{13}$$

Proof. Follows immediately from Definition 5 and the uniform continuity of g on $C \times U \times T$.

The following lemma found in Filippov [3a] will also be needed to establish the second compactness theorem.

LEMMA 2. *Let $\{y^i(\cdot)\}_{i \in I}$, where I is some indexing set, be a collection of absolutely continuous functions from T into R^n such that $\{y^i(0)\}_{i \in I}$ or $\{y^i(1)\}_{i \in I}$ is contained in a compact set of R^n. Let functions $Y^i : T \to R$, $i \in I$, be defined by*

$$Y^i(t) = \|y^i(t)\|^2 + 1. \tag{14}$$

If there exists an $M > 0$ such that $|\dot{Y}^i(t)| \leqq MY^i(t)$, for almost all $t \in T$, $i \in I$, then the set $\{y^i(\cdot)\}_{i \in I}$ is equibounded and equicontinuous. Furthermore, if $I = \{0, 1, 2, \cdots\}$, then there exists a subsequence indexed by a set $K \subset \{0, 1, 2, \cdots\}$ and an absolutely continuous function $\bar{y}(\cdot)$ such that $y^i(\cdot)$ converges uniformly to $\bar{y}(\cdot)$ for $i \in K$.

Now making use of Lemmas 1, 2 and Assumption 2 it is straightforward to show that the following compactness result, due to Warga [12a], holds.

THEOREM 3. *Let C and U be arbitrary compact sets in R^n, R^m, respectively, and let S be the set of measurable relaxed controls which vanish outside of U. If $\{(\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i)\}_{i=0}^{\infty}$ is a sequence of initial states, relaxed controls, corresponding trajectories and corresponding multipliers such that $\{\xi^i\}_{i=0}^{\infty} \subset C$, $\{\mathbf{v}^i\} \subset S$, $\{\xi^i\}$ converges to $\bar{\xi}$, $\{\mathbf{v}_i\}$ converges to $\bar{\mathbf{v}}$ i.s.c.m.,[2] $\{x^i\}$ converges to $\bar{x}$ uniformly, and $\{\lambda_j^i\}$ converges to $\bar{\lambda}_j$ uniformly, $j = 0, 1, \cdots, p$; then $\bar{x}(\cdot) = x(\cdot, \bar{\xi}, \bar{\mathbf{v}})$ and $\bar{\lambda}_j(\cdot) = \lambda_j(\cdot, \bar{\xi}, \bar{\mathbf{v}})$, $j = 0, 1, \cdots, p$. Furthermore, given a sequence $\{(\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i)\}_{i=0}^{\infty}$ such that $\{\xi^i\}_{i=0}^{\infty} \subset C$ and $\{\mathbf{v}^i\}_{i=0}^{\infty} \subset S$, there always exists a subsequence that satisfies the above hypotheses and conclusions.*

[2] If $\mathbf{v}^i \to \bar{\mathbf{v}}$ i.s.c.m., with $\{\mathbf{v}^i\} \subset S$, it follows that $\bar{\mathbf{v}} \in S$ also.

3. Algorithm prototypes and convergence theory. The convergence theorems which we find in [9], as well as in other sources, require that the limit points of sequences constructed by an algorithm lie in the domain of the algorithm. Since this may not be true for optimal control algorithms, it is necessary to modify the existing convergence theory just slightly. We now show how it is done for the simplest case treated in [9]. The more complicated cases discussed in [9] can be modified similarly.

The algorithm prototype below extends the algorithm prototype 1.3.9 in [9]. Let Z be a topological space, $\bar{W}$ be a subset of Z, and W be a subset of $\bar{W}$.[3] We use two functions, the search function, $A : W \to 2^W$, and the stop function, $c : \bar{W} \to R^1$. Finally, we let the set of desirable points, Δ, be a nonempty subset of $\bar{W}$. The problem then is to find any point in Δ where it is assumed that we have some way of recognizing points in Δ.

Algorithm prototype.

Step 0. Compute a $z^0 \in W$.

Step 1. Set $i = 0$.

Step 2. Compute a point $y \in A(z^i)$.

Step 3. Set $z^{i+1} = y$.

Step 4. If $c(z^{i+1}) \geqq c(z^i)$, stop; else, set $i = i+1$ and go to Step 2.

The proof of the following convergence results is the same as that of Theorem 1.3.10 in Polak [9], except that one uses sequences instead of closed balls.

THEOREM 4. *Consider the above algorithm. Suppose that*

(i) *for every nondesirable $\bar{z} \in \bar{W}$ and every sequence $\{z^i\}_{i=0}^{\infty} \subset W$ converging to $\bar{z}$, $\{c(z^i)\}_{i=0}^{\infty}$ converges to $c(\bar{z})$;*

(ii) *for every nondesirable $\bar{z} \in \bar{W}$ and every sequence $\{z^i\}_{i=0}^{\infty} \subset W$ converging to $\bar{z}$, there exists an infinite subset $K \subset \{0, 1, 2, \cdots\}$, an integer $N \geqq 0$, and a $\delta(\bar{z}) > 0$ such that*

$$c(z'') - c(z^i) \leqq -\delta(\bar{z}) < 0 \quad \forall i \geqq N, \quad \forall i \in K, \quad \forall z'' \in A(z^i). \tag{15}$$

Then, either the sequence $\{z^i\}_{i=0}^{\infty}$ constructed by the algorithm is finite and its next to last element is desirable, or else it is infinite and every accumulation point in $\bar{W}$ of $\{z^i\}_{i=0}^{\infty}$ is desirable.

With the proper choice of Z, $\bar{W}$ and W this algorithm prototype and convergence theorem can be applied to a large class of optimal control algorithms. This will be demonstrated in the following sections.

4. A strong variations algorithm. In this section, we shall present a proof of convergence for a strong variations algorithm developed by Mayne and Polak [7]. This algorithm is an "$L_\infty \cap L_2$ stabilized" version of a differential dynamic programming algorithm due to Jacobson and Mayne [5]. Differential dynamic programming algorithms are based on fairly complex relationships between changes in Hamiltonians and changes in cost in optimal control problems. The interested reader is referred to the book by Jacobson and Mayne [5], Mayne [6], and to [7] for background material. The gist of these algorithms is generally as follows. Given a control u_i, an approximation to the optimal control, one

[3] Prototype 1.3.9 in [9] applies only when $W = \bar{W}$, and its convergence is established only in terms of a normed topology.

computes the corresponding trajectories and multipliers x^{u_i}, $\lambda_j^{u_i}$ by solving (2), (10), (11). Then one constructs a Hamiltonian $H(x^{u_i}(t), w, \lambda^{u_i}(t), t)$, where λ^{u_i} is a certain convex combination of the $\lambda_j^{u_i}$, and is an approximation at the optimal costate. By minimizing H with respect to $w \in U$, one obtains an intermediate function $\check{u}_i(t)$.[4] For the algorithm to converge, one now has to use a rather complex way of constructing the next control, u_{i+1}, by setting it equal to u_i for some points in T and to $\check{u}_i$ for some other points in T. The specific rule used in [7] is derived from the Armijo [9] step size selection procedure commonly used in nonlinear programming. Figure 1 will perhaps help the reader in understanding the algorithm. Although strong variations (or differential dynamic programming) algorithms are difficult to understand, they have two distinct advantages: (i) they are computationally efficient, and (ii) they solve certain classes of problems which cannot be solved by other algorithms. (In principle all optimal control problems can be solved by means of penalty function methods, but, at least in our experience, penalty function methods have been found to perform unacceptably on quite a few occasions.)

The algorithm to be described solves the problem (1)–(6) under the additional restriction that the system (5) is replaced by $\dot{\xi} = \xi_0$, that the functions $h_j \equiv 0$ for $j = 1, 2, \cdots, p+q$, and that the set U in (6) is compact. In other words, our initial state is fixed, and we have no initial or terminal inequality constraints. Because of this, we will drop any reference to the initial state. In the discussion below, we shall denote by G the set of measurable functions $u : [0, 1] \to U$.

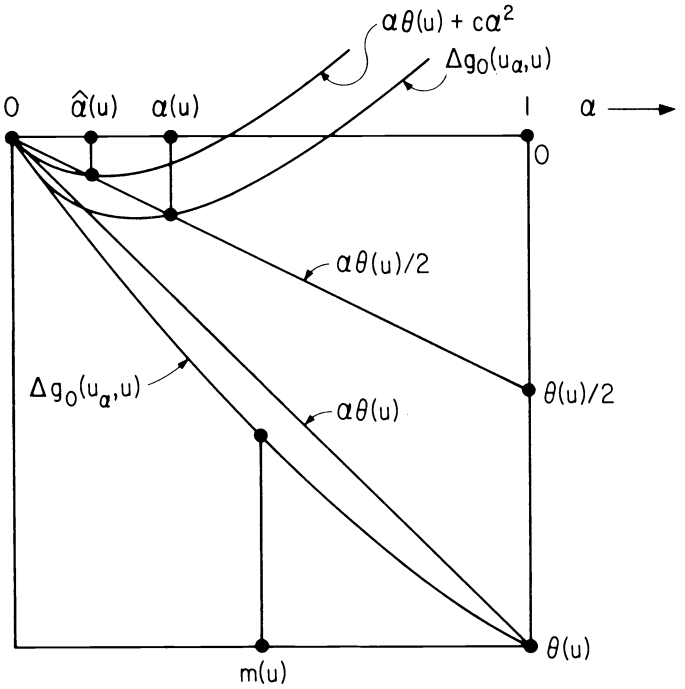


FIG. 1. *Variation of* $\Delta g_0(u_\alpha, u)$ *and* $\Delta g_0(u_\alpha, u)$ *with* α

[4] Thus we can think of these algorithms as being derived from the Pontryagin minimum principle.

To insure that Algorithm 1 is well-defined, we need the following theorem which is a consequence of the McShane–Warfield halfway principle [14].

THEOREM 5. *For any $u \in G$, there exists a $\check{u} \in G$ such that for almost all $t \in T$,*

$$\check{u}(t) \in \check{U}(u, t) \triangleq \arg\min_{w \in U} H_0(x^u(t), w, \lambda_0^u(t), t).^5 \tag{16}$$

Next, let $\bar{H} : G \times T \to R^1$ be defined by

$$\bar{H}(u, t) \triangleq \min_{w \in U} H_0(x^u(t), w, \lambda_0^u(t), t), \tag{17}$$

where H_0 was defined in (12) and let $\theta : G \to R^1$ be defined by

$$\theta(u) \triangleq \int_0^1 [\bar{H}(u, t) - H_0(x^u(t), u(t), \lambda_0^u(t), t)]\, dt. \tag{18}$$

For any $u^1, u^2 \in G$, let $\Delta g_0(u^2, u^1)$ and $\Delta \hat{g}_0(u^2, u^1)$ be defined by

$$\Delta g_0(u^2, u^1) \triangleq g_0(u^2) - g_0(u^1) \tag{19}$$

and

$$\begin{aligned} \Delta \hat{g}_0(u^2, u^1) \triangleq \int_0^1 [&H_0(x^1(t), u^2(t), \lambda_0^1(t), t) \\ &- H_0(x^1(t), u^1(t), \lambda_0^1(t), t)]\, dt, \end{aligned} \tag{20}$$

where g_0 is defined as in (1). (It is shown in [6] that $\Delta \hat{g}_0$ is, in a certain sense, a first order estimate of Δg_0.)

Next, for every $u \in G$, let $\check{U}(u)$, $I_u^{H_0}$, and $m(u)$ be defined, respectively, by

$$\check{U}(u) \triangleq \{v \in G : v(t) \in \arg\min_{w \in U} H_0(x^u(t), w, \lambda_0^u(t), t) \text{ for almost all } t \in T\}; \tag{21}$$

$$I_u^{H_0} \triangleq \{t \in T \mid \bar{H}(u, t) - H_0(x^u(t), u(t), \lambda_0^u(t), t) \leqq \theta(u)\} \tag{22}$$

and

$$m(u) \triangleq \mu(I_u^{H_0}), \tag{23}$$

where μ is Lebesgue measure.

For every $u \in G$ and $\alpha \in [0, 1]$, let $I_{\alpha u}$ be any subset of T having the following properties.

$$\mu(I_{\alpha u}) = \alpha. \tag{24}$$

$$\text{If } \alpha \in [0, m(u)], \quad I_{\alpha u} \subset I_u^{H_0}. \tag{25}$$

$$\text{If } \alpha \in (m(u), 1], \quad I_{\alpha u} \supset I_u^{H_0}. \tag{26}$$

$$\forall \alpha \in [0, m(u)], \quad \{t \in I_u^{H_0}, t' \in I_{\alpha u}, t < t'\} \Rightarrow \{t \in I_{\alpha u}\}. \tag{27}$$

$$\forall \alpha \in (m(u), 1], \quad \{t \in T, t' \in I_{\alpha u} \setminus I_u^{H_0}, t < t'\} \Rightarrow \{t \in I_{\alpha u}\}. \tag{28}$$

[5] Thus $\check{U}(u, t)$ is the set of minimizers.

Next, for any $u \in G$, for any $\alpha \in [0, 1]$, $u_\alpha \in G$ will denote a function with the following properties:

$$u_\alpha(t) \in \check{U}(u, t), \quad \forall t \in I_{\alpha u}, \tag{29}$$

$$u_\alpha(t) = u(t) \quad \forall t \in T \setminus I_{\alpha u}. \tag{30}$$

Finally, let $\alpha : G \to 2^{[0,1]}$ be defined by

$$\alpha(u) = \{\alpha \mid \alpha = \max\{\beta \in [0, 1] \mid \Delta g_0(u_{\beta'}, u) \leqq \beta' \theta(u)/2, \forall \beta' \in [0, \beta]\}\}, \tag{31}$$

where $u_{\beta'} \in G$ is any control that satisfies (29), (30).

ALGORITHM 1 (Mayne and Polak [7]).

Step 0. Select a $u^0 \in G$.

Step 1. Set $i = 0$.

Step 2. Compute x^i by solving (2), with $\xi = \xi_0$.

Step 3. Compute λ_0^i by solving the ordinary control versions of (10) and (11).

Step 4. Compute $\check{u}^i$ such that $\check{u}^i(t) \in U(\check{u}^i, t)$.

Step 5. Compute $\theta(u^i) \triangleq \Delta \hat{g}_0(\check{u}^i, u^i)$ using (20). If $\theta(u^i) = 0$ stop. Else go to Step 6.

Step 6. Compute an $\alpha^i \in \alpha(u^i)$.

Step 7. Set $u^{i+1} = u^i_{\alpha^i}$. Set $i = i + 1$. Go to Step 2.

Algorithm 1 constructs a sequence of ordinary controls. However in proving convergence, we must use relaxed controls. Therefore with each ordinary control u we associate a relaxed control $\mathbf{u}$ which has the property that the measure $\mathbf{u}(t)$ is wholly concentrated at $u(t)$, i.e., $\int_{\{u(t)\}} d\mathbf{u}(t) = 1$ for all $t \in T$. We then see that Algorithm 1 defines a map $A : W \in 2^W$, where W is defined by

$$W \triangleq \{(\mathbf{u}, x^u, \lambda_0^u) \mid u \in G\}. \tag{32}$$

In other words, for any $(\mathbf{u}^i, x^i, \lambda_0^i) \in W$, the set $A((\mathbf{u}^i, x^i, \lambda_0^i))$ consists of the possible $(\mathbf{u}^{i+1}, x^{i+1}, \lambda_0^{i+1})$ which the algorithm can construct from the given point $(\mathbf{u}^i, x^i, \lambda_0^i)$. We will now establish our convergence result for Algorithm 1 using the theory developed in §§ 2 and 3.

As before, let S be the set of measurable relaxed controls which vanish outside of U. We also have to make the straightforward extension of the domain of definition of functions such as θ, $\bar{H}$, etc., to include relaxed controls.

For example,

$$\begin{aligned} \theta(\mathbf{u}) &\triangleq \int_0^1 [\bar{H}(\mathbf{u}, t) - H_{0r}(x^{\mathbf{u}}(t), \mathbf{u}(t), \lambda_0^{\mathbf{u}}(t), t)]\, dt \\ &\triangleq \int_0^1 [\min_{w \in U} H_0(x^{\mathbf{u}}(t), w, \lambda_0^{\mathbf{u}}(t), t) - H_{0r}(x^{\mathbf{u}}(t), \mathbf{u}(t), \lambda_0^{\mathbf{u}}(t), t)]\, dt. \end{aligned} \tag{33}$$

The following lemma is proved in Mayne and Polak [7].

LEMMA 3. *Let $A : W \to 2^W$ be the map defined by Algorithm 1. Then there exists a $c > 0$ such that for all $u \in G$,*

$$\Delta g_0(u', u) \leqq -[\theta(u)]^2/c \quad \forall (\mathbf{u}', x^{\mathbf{u}'}, \lambda_0^{\mathbf{u}'}) \in A((\mathbf{u}, x^{\mathbf{u}}, \lambda_0^{\mathbf{u}})). \tag{34}$$

LEMMA 4. *Let $(\mathbf{u}^i, x^i, \lambda_0^i) \to (\bar{\mathbf{u}}, \bar{x}, \bar{\lambda}_0)$ where $\{\mathbf{u}^i\}_{i=0}^\infty \subset S$. Then $\theta(\mathbf{u}^i) \to \theta(\bar{\mathbf{u}})$.*

Proof. This follows from Lemmas 3 and 4, the continuity of $\min_{w \in U} H_0(x, w, \lambda, t)$ in (x, λ, t), Lemma 1 and Theorem 3.

We can now prove the convergence result. Let W be as in (32), and let Z and $\bar{W}$ be defined by

$$Z = S \times C_n[T] \times C_n[T] \tag{35}$$

and

$$\bar{W} = \{(\mathbf{u}, x^{\mathbf{u}}, \lambda_0^{\mathbf{u}}) \mid \mathbf{u} \in S\}, \tag{36}$$

where $C_n[T]$ is the space of continuous n vector-valued functions on T, with the uniform convergence topology. Let the set of desirable points, Δ, be defined by

$$\Delta = \{(\mathbf{u}, x^{\mathbf{u}}, \lambda_0^{\mathbf{u}}) \in \bar{W} \mid \theta(\mathbf{u}) = 0\}.^{6} \tag{37}$$

THEOREM 6. *Suppose Algorithm* 1 *generates a sequence* $\{(u^i, x^i, \lambda_0^i)\}_{i=0}^{\infty}$; *then either the corresponding sequence* $\{(\mathbf{u}^i, x^i, \lambda_0^i)\}_{i=0}^{\infty}$ *is finite, in which case the last element is desirable, or it is infinite and every accumulation point in* $\bar{W}$ *(at least one exists) is desirable.*

Proof. The above Algorithm 1 is obviously of the form of our Algorithm prototype. Letting W, Z, $\bar{W}$ and Δ be defined, respectively, as in (32), (35), (36) and (37) and c be g_0, we only need to verify conditions (i) and (ii) of Theorem 4 in order to invoke this theorem: (i) If $(\mathbf{u}^i, x^i, \lambda_0^i) \to (\bar{\mathbf{u}}, \bar{x}, \bar{\lambda}_0)$, Lemma 1 immediately implies $g_0(\mathbf{u}^i) \to g_0(\bar{\mathbf{u}})$. (ii) If $(\mathbf{u}^i, x^i, \lambda_0^i) \to (\bar{\mathbf{u}}, \bar{x}, \bar{\lambda}_0)$ with $\theta(\bar{\mathbf{u}}) < 0$, Lemmas 3 and 4 immediately imply that there exists an $N > 0$ such that

$$g_0(\mathbf{u}^{i'}) - g_0(\mathbf{u}^i) \leqq \frac{[\theta(\mathbf{u}^i)]^2}{2c} \leqq -\frac{[\theta(\bar{\mathbf{u}})]^2}{4c} < 0, \tag{38}$$

$$\forall i \geqq N, \quad \forall (\mathbf{u}^{i'}, x^{\mathbf{u}^{i'}}, \lambda^{\mathbf{u}^{i'}}) \in A((\mathbf{u}^i, x^i, \lambda_0^i)).$$

Thus Theorem 4 can be applied. The existence of at least one accumulation point follows from the second half of Theorem 3.

5. A dual method of centers. We shall now consider an algorithm due to Pironneau and Polak [8]. Unlike the algorithm presented in the preceeding section, this one cannot be treated by simply cannibalizing its convergence proof in $L_2 \cap L_\infty$. A special directional derivative must be developed for its analysis.

The algorithm in [8] solves the problem (1)–(6) under the restriction that $h_0 \equiv 0$ and $U = R^m$.

Assumption 3. We will assume that f, L and h_i, $i = 1, \cdots, p+q$, are such that their partials up to second order with respect to x and u exist and are continuous in (x, u, t) on the sets on which they are defined.

The following algorithm is derived from the F. John condition of optimality, as explained in detail in [8]. It is called a "dual" method of feasible directions because it uses multipliers. The "primal", Zoutendijk type methods of feasible directions [9] are derived from the F. John condition in multiplier free form (see [9]), and do not extend to optimal control problems, because the direction finding problems become as difficult as the original problems.

[6] It is shown in Appendix A that $\theta(\mathbf{u}) = 0$ is an optimality condition.

ALGORITHM 2 (Pironneau–Polak [6]) ($\beta \in (0, 1)$ is a step size parameter).

Step 0. Compute a $\xi^0 \in R^n$ and a measurable ordinary control u^0 such that $h_j(\xi^0) \leqq 0$ for $j = p+1, \cdots, p+q$, $h_j(x(1, \xi^0, u^0)) \leqq 0$ for $j = 1, \cdots, p$. Set $i = 0$.

Step 1. Compute $z^i = (\xi^i, u^i, x^i, \lambda_0^i, \cdots, \lambda_p^i)$ according to (2) and the ordinary control versions of (10) and (11).

Step 2. Compute $\nabla g_j(\xi^i, u^i)$, $j = 0, \cdots, p+q$, according to

$$\nabla g_j(\xi^i, u^i) = \left(\lambda_j(0, \xi^i, u^i), \frac{\partial H_j^T}{\partial u}(x(\cdot, \xi^i, u^i), u^i(\cdot), \lambda(\cdot, \xi^i, u^i), \cdot)\right), \tag{39a}$$

$$j = 0, 1, 2, \cdots, p,$$

$$\nabla g_j(\xi^i, u^i) = (\nabla h_j(\xi^i), 0), \qquad j = p+1, p+2, \cdots, p+q. \tag{39b}$$

Step 3. Compute $\mu(z^i) \triangleq (\mu_0(z^i), \cdots, \mu_{p+q}(z^i)) \in R^{p+q+1}$ as a solution of

$$\begin{aligned} \phi(z^i) = \max_{\mu} \Bigg\{ & \sum_{j=1}^{p} \mu_j h_j(x(1, \xi^i, u^i)) \\ & + \sum_{j=p+1}^{p+q} \mu_j h_j(\xi^i) - (1/2) \left\| \sum_{j=0}^{p+q} \mu_j \nabla g_j(\xi^i, u^i) \right\|_2^2 \\ & \Bigg| \sum_{j=0}^{p+q} \mu_j = 1, \mu \geqq 0, j = 0, 1, \cdots, p+q \Bigg\}, \end{aligned} \tag{40}$$

where $\| \cdot \|_2$ denotes the $L_2^{n+m}[0, 1]$-norm.

Step 4. If $\phi(z^i) = 0$, set $\bar{\xi} = \xi^i$ and $\bar{u} = u^i$ and stop; else go to Step 5.

Step 5. Set

$$\omega^i = -\sum_{j=0}^{p} \mu_j(z^i) \lambda_j(0, \xi^i, u^i) - \sum_{j=p+1}^{p+q} \mu_j(z^i) \frac{\partial h_j^T}{\partial x}(\xi^i), \tag{41}$$

$$\nu^i(t, u) = -\sum_{j=0}^{p} \mu_j(z^i) \frac{\partial H_j^T}{\partial u}(x(t, \xi^i, u^i), u, \lambda_j(t, \xi^i, u^i), t) \tag{42a}$$

$$\text{for all } (t, u) \in T \times R^m,$$

$$\bar{\nu}^i(\cdot) = \nu^i(\cdot, u^i(\cdot)). \tag{42b}$$

Step 6. Compute the smallest integer k, such that

$$\begin{aligned} \max \Bigg\{ & \int_0^1 \{L(x(t, \xi^i + \beta^k \omega^i, u^i + \beta^k \bar{\nu}^i), u^i(t) + \beta^k \bar{\nu}^i(t), t) \\ & - L(x(t, \xi^i, u^i), u^i(t), t)\} \, dt; h_j(x(1, \xi^i + \beta^k \omega^i, u^i + \beta^k \bar{\nu}^i)), j = 1, \cdots, p; \\ & h_j(\xi^i + \beta^k \omega^i), j = p+1, \cdots, p+q \Bigg\} - \frac{\beta^k}{2} \phi(z^i) \leqq 0. \end{aligned}$$

Step 7. Set $\xi^{i+1} = \xi^i + \beta^k \omega^i$; set $u^{i+1}(\cdot) = u^i(\cdot) + \beta^k \bar{\nu}^i(\cdot)$, and go to Step 1.

Before proving any convergence results for the above algorithm, we must develop some more theory to make the transition from ordinary controls to relaxed controls. Again this is necessary because we want to study relaxed controls which are accumulation points of a sequence of ordinary controls. In particular,

we need to construct a special directional differential, and we develop a variational equation for this purpose. We first define the following functions which are generalizations of the differentials for functions of ordinary controls.

DEFINITION 8. For any $\xi \in R^n$ and $\mathbf{v}$ a measurable relaxed control, let $\nabla g_j(\xi, \mathbf{v}) : T \times R^m \to R^n \times R^m$, $j = 0, \cdots, p+q$, be defined by

$$\nabla g_j(\xi, \mathbf{v})(t, u) = (\lambda_j(0, \xi, \mathbf{v}), \frac{\partial H_j^T}{\partial u}(x(t, \xi, \mathbf{v}), u, \lambda_j(t, \xi, \mathbf{v}), t)) \tag{43a}$$

for $j = 0, 1, \cdots, p$, and

$$\nabla g_j(\xi, \mathbf{v})(t, u) = \left(\frac{\partial h_j^T}{\partial x}(\xi), 0\right) \tag{43b}$$

for $j = p+1, \cdots, p+q$, where λ_j and H_j are as defined in Definition 4.

It will be shown in Theorem 9 that the $\nabla g_j(\xi, \mathbf{v})$ are analogous to L_∞ gradients (see (56), (57)).

DEFINITION 9. Let $\mathbf{v}$ be a measurable relaxed control, let $\xi, \xi' \in R^n$, and let y, y' be continuous functions from $T \times R^m$ into R^m. Then $\langle (\xi, y), (\xi', y') \rangle_{\mathbf{v}}$ and $|(\xi, y)|_{\mathbf{v}}$ will denote

$$\langle (\xi, y), (\xi', y') \rangle_{\mathbf{v}} = \langle \xi, \xi' \rangle + \int_0^1 \left(\int_{R^m} \langle y(t, u), y'(t, u) \rangle \, d\mathbf{v}(t) \right) dt \tag{44}$$

and

$$|(\xi, y)|_{\mathbf{v}} = \left[\langle \xi, \xi \rangle + \int_0^1 \left(\int_{R^m} \|y(t, u)\|^2 \, d\mathbf{v}(t) \right) dt \right]^{1/2}, \tag{45}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product and $\| \cdot \|$ denotes the Euclidean norm.

DEFINITION 10. Let W and $\bar{W}$ be defined by

$$W \triangleq \{\xi, \mathbf{v}, x^v, \lambda_0^v, \cdots, \lambda_p^v) | \text{ where } \mathbf{v} \text{ is a relaxed control associated with the ordinary control } v, \xi \in R^n \text{ and } g_j(\xi, v) \leqq 0, j = 1, \cdots, p+q\} \tag{46}$$

and

$$\bar{W} \triangleq \{(\xi, \mathbf{v}, x^{\mathbf{v}}, \lambda_0^{\mathbf{v}}, \cdots, \lambda_p^{\mathbf{v}}) | \mathbf{v} \text{ is a measurable relaxed control, } \xi \in R^n \text{ and } g_j(\xi, \mathbf{v}) \leqq 0, j = 1, \cdots, p+q\}. \tag{47}$$

DEFINITION 11. Let Δ, the set of desirable points, be defined by $\Delta = \{(\xi, \mathbf{v}, x^{\mathbf{v}}, \lambda_0^{\mathbf{v}}, \cdots, \lambda_p^{\mathbf{v}}) \in \bar{W} |$ there exists multipliers μ_j, $j = 0, 1, \cdots, p+q$, such that (i) $\mu_j \geqq 0$, $j = 0, 1, \cdots, p+q$, (ii) $\sum_{j=0}^{p+q} \mu_j = 1$, (iii) $\mu_j g_j(\xi, \mathbf{v}) = 0$ for $j = 1, \cdots, p+q$, (iv) $|\sum_{j=0}^{p+q} \mu_j \nabla g_j(\xi, \mathbf{v})|_{\mathbf{v}}^2 = 0\}$.

Assumption 4. The set $\{(\xi, \mathbf{u}, x^u, \lambda_0^u, \cdots, \lambda_p^u) \in W | g_i(\xi, u) < 0, j = 1, \cdots, p+q\} \neq \varnothing$.

The following definition is an extension of ϕ in (40) to relaxed controls.

DEFINITION 12. Let $z^i = (\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \lambda_1^i, \cdots, \lambda_p^i)$. Then

$$\phi(z^i) = \max \Big\{ \sum_{j=1}^{p} \mu_j h_j(x(1, \xi^i, \mathbf{v}^i)) + \sum_{j=p+1}^{p+q} \mu_j h_j(\xi^i) - (1/2) \Big| \sum_{j=0}^{p+q} \mu_j \nabla g_j(\xi^i, \mathbf{v}^i) \Big|_{\mathbf{v}^i}^2 \Big| \sum_{j=0}^{p+q} \mu_j = 1, \mu_j \geqq 0, j = 0, 1, \cdots, p+q \Big\}. \tag{48}$$

Thus Algorithm 2 defines a map $A : W \to W$.

We can establish convergence properties only for bounded infinite sequences $\{\xi^i, u^i, x^i, \lambda_0^i, \cdots, \lambda_{p+q}^i\}$ constructed by Algorithm 2. We therefore introduce an arbitrary compact set $C \subset R^n$ which will be assumed to contain $\{\xi^i\}$ and an arbitrary compact set $U \subset R^m$ which will be assumed to contain $\{u^i(t)\}$, $t \in T$. In addition, we shall make use of an arbitrary compact set D containing C in its interior, and we shall denote by S the set of measurable relaxed controls which vanish outside of U.

LEMMA 5. *Let* $\{\xi^i\}_{i=0}^{\infty} \subset C$, $\{\mathbf{u}^i\}_{i=0}^{\infty} \subset S$ *be such that* $(\xi^i, \mathbf{u}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i) \to (\bar{\xi}, \bar{\mathbf{u}}, \bar{x}, \bar{\lambda}_0, \cdots, \bar{\lambda}_p)$. *Then there exists a subsequence indexed by* $K \subset \{0, 1, 2, \cdots\}$ *such that* $\phi(z^i) \overset{K}{\to} \phi(\bar{z})$, *where* $z^i = (\xi^i, \mathbf{u}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i)$ *and* $\bar{z} = (\bar{\xi}, \bar{\mathbf{u}}, \bar{x}, \bar{\lambda}_0, \cdots, \bar{\lambda}_p)$.

Proof. Let μ^i be a solution to (48). Since $\{\mu^i\}_{i=0}^{\infty}$ is contained in a compact set, there exists a subsequence indexed by $K \subset \{0, 1, 2, \cdots\}$ and a $\bar{\mu} \in R^{p+q+1}$ such that $\mu^i \overset{K}{\to} \bar{\mu}$, $\sum_{j=0}^{p+q} \bar{\mu}_j = 1$, and $\bar{\mu}_j \geqq 0$ for $j = 0, 1, \cdots, p+q$. By Lemma 1, we obtain $\phi(z^i) \overset{K}{\to} (\sum_{j=1}^{p} \bar{\mu}_j h_j(x(1, \bar{\xi}, \bar{\mathbf{u}})) + \sum_{j=p+1}^{p+q} \bar{\mu}_j h_j(\bar{\xi}) - (1/2) |\sum_{j=0}^{p+q} \bar{\mu}_j \nabla g_j(\bar{\xi}, \bar{\mathbf{u}})|_{\bar{\mathbf{u}}}^2)$. Now

$$\phi(\bar{z}) \geqq \Big\{ \sum_{j=1}^{p} \bar{\mu}_j h_j(x(1, \bar{\xi}, \bar{\mathbf{u}})) + \sum_{j=p+1}^{p+q} \bar{\mu}_j h_j(\bar{\xi}) - (1/2) \Big| \sum_{j=0}^{p+q} \bar{\mu}_j \nabla g_j(\bar{\xi}, \bar{\mathbf{u}}) \Big|_{\bar{\mathbf{u}}}^2 \Big\}.$$

Suppose the inequality is strict and let $\bar{\bar{\mu}} = (\bar{\bar{\mu}}_0, \cdots, \bar{\bar{\mu}}_{p+q})$ be a solution of (48), for $z^i = \bar{z}$, that gives $\phi(\bar{z})$. Then we must have that

$$\Big(\sum_{j=1}^{p} \bar{\bar{\mu}}_j h_j(x(1, \xi^i, \mathbf{u}^i)) + \sum_{j=p+1}^{p+q} \bar{\bar{\mu}}_j h_j(\xi^i) - (1/2) \Big| \sum_{j=0}^{p+q} \bar{\bar{\mu}}_j \nabla g_j(\xi^i, \mathbf{u}^i) \Big|_{\mathbf{u}^i}^2 \Big) \overset{K}{\to} \phi(\bar{z}).$$

But this implies that for sufficiently large $i \in K$,

$$\Big(\sum_{j=1}^{p} \bar{\bar{\mu}}_j h_j(x(1, \xi^i, \mathbf{u}^i)) + \sum_{j=p+1}^{p+q} \bar{\bar{\mu}}_j h_j(\xi^i) - (1/2) \Big| \sum_{j=0}^{p+q} \bar{\bar{\mu}}_j \nabla g_j(\xi^i, \mathbf{u}^i) \Big|_{\mathbf{u}^i}^2 \Big) > \phi(z^i).$$

This is a contradiction of (48). Therefore,

$$\phi(\bar{z}) = \sum_{j=1}^{p} \bar{\mu}_j h_j(x(1, \bar{\xi}, \bar{\mathbf{u}})) + \sum_{j=p+1}^{p+q} \bar{\mu}_j h_j(\bar{\xi}) - (1/2) \Big| \sum_{j=0}^{p+q} \bar{\mu}_j \nabla g_j(\bar{\xi}, \bar{\mathbf{u}}) \Big|_{\bar{\mathbf{u}}}^2. \tag{49}$$

Thus $\phi(z^i) \overset{K}{\to} \phi(\bar{z})$.

The following lemma can be deduced from an analogous result in [8].

LEMMA 6. *Let* $\phi : \bar{W} \to R^1$ *be defined as in* (48), *and let* $z \in \bar{W}$ *be arbitrary. Then* $\phi(z) \leqq 0$, *and* $\phi(z) = 0$ *if and only if* $z \in \Delta$.

LEMMA 7 (see [8]). *Suppose that $z \in W$ is such that $\phi(z) < 0$ and that $\mu(z) = (\mu_0(z), \cdots, \mu_{p+q}(z))$ is a solution to (40) for $z^i = z$. Then*

$$\begin{aligned} &\max \left\{ \left\langle \nabla g_0(\xi, u), -\sum_{j=0}^{p+q} \mu_j(z) \nabla g_j(\xi, u) \right\rangle_2 ; \right. \\ &\left. g_j(\xi, u) + \left\langle \nabla g_j(\xi, u), -\sum_{j=0}^{p+q} \mu_j(z) \nabla g_j(\xi, u) \right\rangle_2, j = 1, \cdots, p+q \right\} \\ &\leqq \phi(z) - (1/2) \left\| \sum_{j=0}^{p+q} \mu_j(z) \nabla g_j(\xi, u) \right\|_2^2 < 0. \end{aligned} \tag{50}$$

The following corollary to Lemma 7 is obtained by application of Lemma 1.

COROLLARY 1. *Suppose that $z \in \bar{W}$ is such that $\phi(z) < 0$; then there exists a $\mu(z)$ such that*

$$\begin{aligned} &\max \left\{ \left\langle \nabla g_0(\xi, \mathbf{v}), -\sum_{j=0}^{p+q} \mu_j(z) \nabla g_j(\xi, \mathbf{v}) \right\rangle_{\mathbf{v}} ; \right. \\ &\left. g_j(\xi, \mathbf{v}) + \left\langle \nabla g_j(\xi, \mathbf{v}), -\sum_{j=0}^{p+q} \mu_j(z) \nabla g_j(\xi, \mathbf{v}) \right\rangle_{\mathbf{v}}, j = 1, \cdots, p+q \right\} \\ &\leqq \phi(z) - (1/2) \left| \sum_{j=0}^{p+q} \mu_j(z) \nabla g_j(\xi, \mathbf{v}) \right|_{\mathbf{v}}^2 < 0.^7 \end{aligned} \tag{51}$$

At this point we develop a set of variational equations defining a special directional differential which we shall need to show that Algorithm 2 satisfies (ii) of Theorem 4.

DEFINITION 13. For any $\xi \in R^n$, any measurable relaxed control $\mathbf{v}$, any $\alpha \in [-1, 1]$ and any $y \in C_m[T \times R^m]$, let $x(t, \xi, \mathbf{v}, \alpha, y)$ denote the solution of

$$\begin{aligned} &\frac{d}{dt} x(t, \xi, \mathbf{v}, \alpha, y) = \int_{u \in R^m} f(x(t, \xi, \mathbf{v}, \alpha, y), u + \alpha y(t, u), t) \, d\mathbf{v}(t), \\ &x(0, \xi, \mathbf{v}, \alpha, y) = \xi. \end{aligned} \tag{52}$$

The following results can be established by lengthy, but straightforward calculations. For a proof see [13].

DEFINITION 14. For any $\xi \in R^n$, any measurable relaxed control $\mathbf{v}$, any $\alpha \in [-1, 1]$, any $y \in C_m[T \times R^m]$ and any $\delta\xi \in R^n$, let $x(t) = x(t, \xi, \mathbf{v}, 0, y)$ and let $\delta x(x, \delta\xi, y, \alpha)(\cdot)$ denote the solution of

$$\begin{aligned} \delta\dot{x}(x, \delta\xi, y, \alpha)(t) = \int_{R^m} &\left[\frac{\partial f}{\partial x}(x(t), u, t) \, \delta x(x, \delta\xi, y, \alpha)(t) \right. \\ &\left. + \frac{\partial f}{\partial u}(x(t), u, t) \cdot \alpha y(t, u) \right] d\bar{\mathbf{v}}(t), \qquad t \in T, \end{aligned} \tag{53}$$

and

$$\delta x(x, \delta\xi, y, \alpha)(0) = \delta\xi. \tag{54}$$

[7] *Note.* If $z^i \to z$, where $z^i \in W$, then a $\mu(z)$ that is an accumulation point of $\{\mu(z^i)\}$ satisfies (50).

THEOREM 7. *There exists a* $K>0$ *such that* $\|x(t, \xi+\delta\xi, \mathbf{v}, \alpha, y)-x(t, \xi, \mathbf{v}, 0, y)\| \leqq K(|\alpha|+\|\delta\xi\|)$ *for all* $t \in T$, *for all* $\alpha \in [-1, 1]$, *for all* $\xi \in C$, *for all* $\mathbf{v} \in S$ *and for all* $\delta\xi \in R^n$ *such that* $\xi+\delta\xi \in D$.

THEOREM 8. *Let* $y(\cdot)$ *and* $x(\cdot)$ *be as in Definition* 14. *Then there exists an* $M>0$ *such that* $\|\delta x(x, \delta\xi, y, \alpha)(t)-(x(t, \xi+\delta\xi, \mathbf{v}, \alpha, y)-(x(t, \xi, \mathbf{v}, 0, y)))\| \leqq M(\|\delta\xi\|+|\alpha|)^2$ *for all* $t \in T$, *for all* $\alpha \in [-1, 1]$, *for all* $\xi \in C$, *for all* $\mathbf{v} \in S$, *and for all* $\delta\xi \in R^n$ *such that* $\xi+\delta\xi \in D$.[8]

The following theorem is a consequence of Theorem 8.

THEOREM 9. *Let* $L : R^n \times R^m \times T \to R^1$ *and* $\phi : R^n \to R^1$, $h_j : R^n \to R^1$, $j = 1, 2, \cdots, p$, *be functions whose partial derivatives with respect to* x *and* u *exist and are continuous in* (x, u, t) *up through second order; then there exists a* $P>0$ *such that for all* $\alpha \in [-1, 1]$, $\mathbf{v} \in S$, $\xi \in C$ *and* $\delta\xi$ *such that* $\xi+\delta\xi \in D$,

$$
\begin{aligned}
&\Bigg| \int_0^1 \left(\int_{R^m} L(x(t, \xi+\delta\xi, \mathbf{v}, \alpha, y), u+\alpha y(t, u), t)\, d\mathbf{v}(t) \right) dt \\
&\qquad + \phi(x(1, \xi+\delta\xi, \mathbf{v}, \alpha, y)) \\
&\qquad - \int_0^1 \left(\int_{R^m} L(x(t), u, t)\, d\mathbf{v}(t) \right) dt - \phi(x(1)) \\
&\qquad - \alpha \int_0^1 \left(\int_{R^m} \left\langle \frac{\partial H_0^T}{\partial u}(x(t), u, \lambda(t), t), y(t, u) \right\rangle d\mathbf{v}(t) \right) dt - \langle \lambda_0(0), \delta\xi \rangle \Bigg| \\
&\quad \leqq P(|\alpha|+\|\delta\xi\|)^2,
\end{aligned} \tag{55}
$$

$$
\begin{aligned}
&\Bigg| h_j(x(1, \xi+\delta\xi, \mathbf{v}, \alpha, y)) - h_j(x(1, \xi, \mathbf{v}, 0, y)) \\
&\qquad - \alpha \int_0^1 \left(\int_{R^m} \left\langle \frac{\partial H_j^T}{\partial u}(x(t), u, \lambda_j(t), t), y(t, u) \right\rangle d\mathbf{v}(t) \right) dt - \langle \lambda_j(0), \delta\xi \rangle \Bigg| \\
&\quad \leqq P(|\alpha|+\|\delta\xi\|)^2,
\end{aligned} \tag{56}
$$

where the λ_j, $j = 0, 1, 2, \cdots, p$, and H_j are defined as in Definition 4.

To relate Theorem 9 to Gateaux differentials in $L_2 \cap L_\infty$, we observe that Theorem 9 implies that there exists a P' such that with g_j defined as in (4), (5) and ∇g_j as in (43a), (43b),

$$
|h_j(x(1, \xi+\alpha\delta\xi, \mathbf{v}, \alpha, y)) - h_j(x(1, \xi, \mathbf{v}, 0, y)) - \alpha\langle \nabla g_j(\xi, \mathbf{v}), (\delta\xi, y) \rangle_{\mathbf{v}}| \leqq P'\alpha^2, \quad j = 1, 2, \cdots, p \tag{57}
$$

and

$$
\begin{aligned}
&\Bigg| \int_0^1 \left(\int_{R_m} L(x(t, \xi+\alpha\delta\xi, \mathbf{v}, \alpha, y), u+\alpha y(t, u), t)\, d\mathbf{v}(t) \right) dt \\
&\qquad + \phi(x(1, \xi+\alpha\delta\xi, \mathbf{v}, \alpha, y)) \\
&\qquad - \int_0^1 \left(\int_{R^m} L(x(t, \xi, v, 0, y), u, t)\, d\mathbf{v}(t) \right) dt - \phi(x(1)) \\
&\qquad - \alpha\langle \nabla g_0(\xi, \mathbf{v}), (\delta\xi, y) \rangle_{\mathbf{v}} \Bigg| \leqq P'\alpha^2.
\end{aligned} \tag{58}
$$

[8] Thus $\delta x(x, \delta\xi, y, \alpha)(\cdot)$ is a kind of directional differential.

DEFINITION 15. Let $\theta : R^1 \times (R^n \times C_m[T]) \times C \times S \to R^1$ be defined by

$$\theta(\alpha, (\delta\xi, y), \xi, \mathbf{v}) = \max\left\{\int_0^1 \left[\int_{R^m} (L(x(t, \xi+\alpha\delta\xi, \mathbf{v}, \alpha, y), u+\alpha y(t), t) - L(x(t, \xi, \mathbf{v}, 0, 0), u, t))\, d\mathbf{v}(t)\right] dt;\ h_j(x(1, \xi+\alpha\delta\xi, \mathbf{v}, \alpha, y)),\ j=1, \cdots, p;\ h_j(\xi+\alpha\cdot\delta\xi), j=p+1, \cdots, p+q\right\}. \tag{59}$$

PROPOSITION 1. *Let* $\{z^i\}_{i=0}^{\infty} \triangleq \{(\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i)\} \subset W$ *be a sequence converging to* $\bar{z} = (\bar{\xi}, \bar{\mathbf{v}}, \bar{x}, \bar{\lambda}_0, \cdots, \bar{\lambda}_p)$, *with* $\{\mathbf{v}^i\} \subset S$ *and* $\phi(\bar{z}) < 0$. *Then there exists an integer* $k(\bar{z})$ *such that*

$$\theta(\beta^{k(\bar{z})}, -\sum_{j=0}^{p+q} \mu_j(\bar{z})\nabla g_j(\bar{\xi}, \bar{\mathbf{v}}), \bar{\xi}, \bar{\mathbf{v}}) - \tfrac{3}{4}\beta^{k(\bar{z})}\phi(\bar{z}) \leqq 0, \tag{60}$$

where $\mu(\bar{z})$ *is an accumulation point of a sequence* $\{\mu(z^i)\}$ *corresponding to* $\{z^i\}$.

Proof. This result follows directly from the definition of $\nabla g_j(\bar{z})$, $j = 0, \cdots, p+q$, Theorem 9 and the fact that by inequality (50), $\langle \nabla g_0(\bar{z}), -\sum_{j=0}^{p+q} \mu_j(\bar{z})\nabla g_j(\bar{\xi}, \bar{\mathbf{v}})\rangle_{\bar{\mathbf{v}}} \leqq \phi(\bar{z})$, and $\langle \nabla g_j(\bar{z}), -\sum_{j=0}^{p+q} \mu_j(\bar{z})\nabla g_j(\bar{\xi}, \bar{\mathbf{v}})\rangle_{\bar{\mathbf{v}}} \leqq \phi(\bar{z})$ for all $j \in \{1, \cdots, p+q\}$ such that $g_j(\bar{\xi}, \bar{\mathbf{v}}) = 0$.

The following lemma is obtained by repeated utilization of Lemma 1.

LEMMA 8. *Let* $\{z^i\}_{i=0}^{\infty} \triangleq \{(\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i)\}_{i=0}^{\infty} \subset W$ *be a sequence converging to* $\bar{z} = (\bar{\xi}, \bar{\mathbf{v}}, \bar{x}, \bar{\lambda}_0, \cdots, \bar{\lambda}_p)$, *where* $\{\mathbf{v}^i\}_{i=0}^{\infty} \subset S$, $\{\xi^i\}_{i=0}^{\infty} \subset C$, *and suppose that a corresponding sequence of solutions to* (48), $\{\mu(z^i)\}_{i=0}^{\infty}$, *converges to a* $\mu(\bar{z})$. *Then for any* $\alpha \in [-1, 1]$, *there exists an infinite subset* $J(\alpha) \subset \{0, 1, 2, \cdots\}$ *such that*

$$\theta\Big(\alpha, -\sum_{j=0}^{p+q} \mu_j(z^i)\nabla g_j(\xi^i, \mathbf{v}^i), \xi^i, \mathbf{v}^i\Big) \xrightarrow{J(\alpha)} \theta\Big(\alpha, -\sum_{j=0}^{p+q} \mu_j(\bar{z})\nabla g_j(\bar{\xi}, \bar{\mathbf{v}}), \bar{\xi}, \bar{\mathbf{v}}\Big). \tag{61}$$

LEMMA 9. *Let* $\{(\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i)\}_{i=0}^{\infty}$ *be a sequence in* W, *converging to* $\bar{z} \triangleq (\bar{\xi}, \bar{\mathbf{v}}, \bar{x}, \bar{\lambda}_0, \cdots, \bar{\lambda}_p)$, *that satisfies the hypotheses of* Lemma 8. *Suppose that* $\phi(\bar{z}) < 0$. *Then there exist a* $\delta(\bar{z}) < 0$, *an integer* $M > 0$ *and an infinite subset* $K \subset \{0, 1, 2, \cdots\}$ *such that*

$$g_0(\xi^{i+1}, \mathbf{v}^{i+1}) - g_0(\xi^i, \mathbf{v}^i) \leqq \delta(\bar{z}) \quad \forall i \in K, \quad \forall i \geqq M, \tag{62}$$

where ξ^{i+1} *and* $\mathbf{v}^{i+1}$ *are, respectively, the initial state and the measurable control that Algorithm* 2 *would construct from the control* $\mathbf{v}^i$ *and initial state* ξ^j.

Proof. For $j = 0, 1, 2, \cdots, p$, let $\bar{y}_j(t, u) = (\partial H_j^T/\partial u)(x(t, \bar{\xi}, \bar{\mathbf{v}}), u, \lambda_j(t, \bar{\xi}, \bar{\mathbf{v}}), t)$, and let $y_j^i(t, u) = (\partial H_j^T/\partial u)(x, (t, \xi^i, \mathbf{v}^i), u, \lambda_j(t, \xi^i, \mathbf{v}^i), t)$. Then by Proposition 1, there exists an integer $k(\bar{z}) \geqq 0$ such that

$$\begin{aligned}
&\int_0^1 \left[\int_{R^m} \Big(L\Big\{x\Big(t, \bar{\xi} + \beta^{k(\bar{z})} \cdot \Big[-\sum_{j=0}^{p} \mu_j(\bar{z})\lambda_j(0, \bar{\xi}, \bar{\mathbf{v}}) - \textstyle\sum_{j=p+1}^{p+q} \mu_j(\bar{z})\dfrac{\partial h_j^T}{\partial x}(\bar{\xi})\Big], \bar{\mathbf{v}}, \beta^{k(\bar{z})}, \right.\right. \\
&\qquad \left.\left. -\sum_{j=0}^{p} \mu_j(\bar{z})\bar{y}_j\Big), u + \beta^{k(\bar{z})}\Big[-\sum_{j=0}^{p} \mu_j(\bar{z})\bar{y}_j(t, u)\Big], t\Big\} \right. \\
&\qquad \left. - L(x(t, \bar{\xi}, \bar{\mathbf{v}}, 0, 0), u, t)\Big)\, d\bar{\mathbf{v}}(t)\right] dt \\
&\leqq \theta\Big(\beta^{k(\bar{z})}, -\sum_{j=0}^{p+q} \mu_j(\bar{z})\nabla g_j(\bar{\xi}, \bar{\mathbf{v}}), \bar{\xi}, \bar{\mathbf{v}}\Big) \leqq \tfrac{3}{4}\beta^{k(\bar{z})}\phi(\bar{z}) < 0.
\end{aligned} \tag{63}$$

By Lemmas 8 and 1 and inequality (63), there exists an integer $M' > 0$ such that

$$\begin{aligned}
&\int_0^1 \int_{R^m} L(x(t, \xi^i + \beta^{k(z)}\omega^i, \mathbf{v}^i, u, \beta^{k(\bar{z})}, \mathbf{v}^i), u + \beta^{k(\bar{z})}\nu^i(t,u), t) d\mathbf{v}^i(t)\, dt \\
&\qquad - \int_0^1 \int_{R^m} L(x(t, \xi^i, \mathbf{v}^i, 0, 0), u, t)\, d\mathbf{v}^i(t) \\
&= \int_0^1 \Bigg[\int_{R^m} \Bigg(L\Bigg(x\Bigg(t, \xi^i + \beta^{k(\bar{z})}\Bigg[-\sum_{j=0}^{p} \mu_j(z^i)\lambda_j(0, \xi^i, \mathbf{v}^i) \\
&\qquad - \sum_{j=p+1}^{p+q} \mu_j(z^i)\frac{\partial h_j}{\partial x}(\xi^i)\Bigg], \mathbf{v}^i, \beta^{k(\bar{z})}, -\sum_{j=0}^{p} \mu_j(z^i) y_j^i \Bigg), \\
&\qquad u + \beta^{k(\bar{z})}\Bigg[-\sum_{j=0}^{p} \mu_j(z^i) y_j^i(t,u)\Bigg], t\Bigg) \\
&\qquad - L(x(t, \xi^i, \mathbf{v}^i, 0, 0), u, t)\Bigg) d\mathbf{v}^i(t)\Bigg] dt \\
&\leqq \theta\Bigg(\beta^{k(\bar{z})}, -\sum_{j=0}^{p+q} \mu_j(z^i)\nabla g_j(\xi^i, \mathbf{v}^i), \xi^i, \mathbf{v}^i \Bigg) \\
&\leqq \tfrac{1}{2}\beta^{k(\bar{z})}\phi(z^i) < 0 \qquad\qquad \forall i \geqq M', \quad \forall i \in J(\beta^{k(\bar{z})}).
\end{aligned} \tag{64}$$

Now consider the control $\mathbf{v}^i$ and initial state ξ^i and the control $\mathbf{v}^{i+1}$ and initial state ξ^{i+1} which Algorithm 2 constructs. The control $\mathbf{v}^{i+1}$ is associated with

$$u^{i+1}(\cdot) = u^i(\cdot) + \beta^{k(z^i)}\bar{\nu}^i(\cdot) \tag{65}$$

and initial state by

$$\xi^{i+1} = \xi^i + \beta^{k(z^i)}\omega^i, \tag{66}$$

where $k(z^i)$ is the integer computed in Step 6 of Algorithm 2. It follows from (64) that $\beta^{k(z^i)} \geqq \beta^{k(\bar{z})}$. Therefore by construction, we get

$$\begin{aligned}
g_0(\xi^{i+1}, \mathbf{v}^{i+1}) - g_0(\xi^i, \mathbf{v}^i) &\leqq \theta\Bigg(\beta^{k(z^i)}, -\sum_{j=0}^{p+q} \mu_j(z^i)\nabla g_j(\xi^i, \mathbf{v}^i), \xi^i, \mathbf{v}^i \Bigg) \\
&\leqq \tfrac{1}{2}\beta^{k(z^i)}\phi(z^i) \leqq \tfrac{1}{2}\beta^{k(\bar{z})}\phi(z^i) \\
&\qquad\qquad \forall i \geqq M', \quad \forall i \in J(\beta^{k(\bar{z})}).
\end{aligned} \tag{67}$$

Since $\phi(z^i) \to \phi(\bar{z})$, there exists an integer $M \geqq M'$ such that

$$g_0(\xi^{i+1}, \mathbf{v}^{i+1}) - g_0(\xi^i, \mathbf{v}^i) \leqq \tfrac{1}{4}\beta^{k(\bar{z})}\phi(\bar{z}) \qquad \forall i \geqq M, \quad \forall i \in J(\beta^{k(\bar{z})}), \tag{68}$$

which completes our proof.

We now give the main result of this section.

THEOREM 10. *Let* $\{(\xi^i, \mathbf{v}^i, x^i, \lambda_0^i, \cdots, \lambda_p^i)\}_{i=0}^{\infty}$ *be a sequence of initial states, measurable controls, corresponding trajectories and corresponding multipliers constructed by Algorithm* 2. *If there exist compact sets* $C \subset R^n$, $U \subset R^m$ *such that* $\xi^i \in C$ *and* $u^i(t) \in U$, $t \in T$, *for all* $i = 0, 1, 2, \cdots$, *then either the sequence is finite, in which case the last element is desirable, or it is infinite and every accumulation point of this sequence is desirable. Furthermore, at least one accumulation point exists.*

Proof. The above Algorithm 2 is basically of the form of our Algorithm prototype. With W, $\bar{W}$ and Δ defined as in Definitions 10 and 11, $c \equiv g_0$, we only need verify conditions (i) and (ii) of Theorem 4 in order to invoke Theorem 4. Lemma 2 immediately implies (i) and Lemmas 7 and 9 immediately imply (ii). The existence of at least one accumulation point is guaranteed by Theorem 3.

Conclusion. The two examples we have included in this paper illustrate the use of the new convergence results for optimal control algorithms. Many other and much more complex algorithms can be analyzed in a similar way. The interested reader can find further results in [9]. The net effect of our work is to show that optimal control algorithms are very well-behaved, contrary to the misgivings felt by some theoreticians.

Appendix A. Optimality conditions in optimization algorithms. A careful examination of nonlinear programming algorithms (see, e.g., [9, chap. 4]) shows that they are frequently derived from variants of some basic optimality condition. For example, Rosen's gradient projection method is based on the Kuhn–Tucker conditions in standard form. The Zukhovidskii–Polyak–Primak method of feasible directions is based on the Kuhn–Tucker conditions stated as a multiplier free constrained optimization problem. The Zoutendijk and Demyanov methods of feasible directions are based on the F. John condition stated as a multiplier-free min-max problem, and the Pironneau–Polak method is based on the F. John conditions stated as a max problem with multipliers, which also happens to be the dual of a multiplier-free min–max problem. Many more such examples can be cited.

The same phenomenon holds true in optimal control algorithms, as illustrated by the two algorithms presented in this paper. We shall now show the relationship between the optimality conditions $\theta(z_i)=0$, $\phi(z_i)=0$ used in Algorithms 1 and 2 with the relaxed minimum principle.

THEOREM A.1 (the relaxed minimum principle). *If* $\mathbf{u}$ *is optimal for the relaxed optimal control problem* (7), (8), (9), (4), (5) *and* $x^{\mathbf{u}}$ *is the corresponding optimal trajectory, then* $x^{\mathbf{u}}(1)$ *satisfies* (4), $x^{\mathbf{u}}(0)=\xi$ satisfies (5), $\mathbf{u}(\cdot)$ *satisfies* (9), *and there exist a scalar* λ^0 *and a costate trajectory* $\lambda^{\mathbf{u}}$, *with* $(\lambda^0, \lambda^{\mathbf{u}}(t)) \not\equiv 0$, *such that* $\lambda^0 \geqq 0$ *and*

$$\text{(A.1)} \qquad \frac{d}{dt}\lambda^{\mathbf{u}}(t) = -\lambda^0\left(\frac{\partial L}{\partial x}\right)_r (x^{\mathbf{u}}(t), \mathbf{u}(t), t) - \left\langle \left(\frac{\partial f}{\partial x}\right)_r (x^{\mathbf{u}}(t), \mathbf{u}(t), t), \lambda^{\mathbf{u}}(t)\right\rangle, \qquad t \in T,$$

with

$$\text{(A.2)} \qquad \lambda^{\mathbf{u}}(1) = \sum_{j=0}^{p} \mu_j \nabla h_j(x^{\mathbf{u}}(1)),$$

$$\text{(A.3)} \qquad \lambda^{\mathbf{u}}(0) = -\sum_{j=p+1}^{p+q} \mu_j \nabla h_j(x^{\mathbf{u}}(0)),$$

where $\mu_j \geqq 0$ *for* $j=0, 1, 2, \cdots, p+q$, $\mu_j h_j(x^{\mathbf{u}}(1))=0$ *for* $j=1, 2, \cdots, p$,

$\mu_j h_j(x^{\mathbf{u}}(0)) = 0$ *for* $j = p+1, \cdots, p+q$, *and for all admissible relaxed controls,* $\hat{\mathbf{u}}$,

$$\begin{aligned}&\Delta H(x^{\mathbf{u}}(t), \lambda^{\mathbf{u}}(t), \mathbf{u}(t), \hat{\mathbf{u}}(t), t)\\ &\triangleq \lambda^0 L_r(x^{\mathbf{u}}(t), \mathbf{u}(t), t) + \langle \lambda^{\mathbf{u}}(t), f_r(x^{\mathbf{u}}(t), \mathbf{u}(t), t)\rangle\\ &\quad - (\lambda^0 L_r(x^{\mathbf{u}}(t), \hat{\mathbf{u}}(t), t) + \langle \lambda^{\mathbf{u}}(t), f_r(x^{\mathbf{u}}(t), \hat{\mathbf{u}}(t), t)\rangle) \leqq 0.\end{aligned} \tag{A.4}$$

Now consider Algorithm 1, which solves the fixed initial state, free terminal state problem. Since (see (33)) $\theta(z) \leqq 0$ for all admissible z and since whenever $\theta(z) < 0$, the algorithm will construct a z' resulting in a lower cost, it is clear that if z is optimal, then $\theta(z) = 0$. The relationship of $\theta(z) = 0$ to the relaxed maximum principle is as follows.

THEOREM A.2. *Consider the optimal control problem solved by Algorithm* 1 *with the accompanying assumptions. Suppose that* $\bar{W}$ *is defined as in* (36) *and that* $z = (\mathbf{u}, x^{\mathbf{u}}, \lambda_0^{\mathbf{u}})$ *is such that* $\theta(z) = 0$; *then* $\lambda_0^{\mathbf{u}}$ *satisfies* (A.1) *with* $\lambda^0 = 1$, $\lambda_0^{\mathbf{u}}(1)$ *satisfies* (A.2) *with* $\mu_0 = 1$, $\mu_j = 0$, $j = 1, \cdots, p$ *and* (A.3) *with* $\mu_j = 0$, $j = p+1, \cdots, q$, *and for all admissible relaxed controls* $\hat{\mathbf{u}}$,

$$\int_0^1 \Delta H(x^{\mathbf{u}}(t), \lambda^{\mathbf{u}}(t), \mathbf{u}(t), \hat{\mathbf{u}}(t), t)\, dt \leqq 0. \tag{A.5}$$

Thus, $\theta(z) = 0$ is seen as an integral form of the relaxed maximum principle.

Now consider the problem solved by Algorithm 2. Again by construction, it is clear that $\phi(z) = 0$ (see (40)) is a necessary condition of optimality. Its relation to the relaxed minimum principle is as follows.

THEOREM A.3. *Consider the optimal control problem solved by Algorithm* 2, *with the accompanying assumption. Suppose that* $\bar{W}$ *is defined as in* (45) *and that* $z = (\mathbf{u}, x^{\mathbf{u}}, \lambda_0^{\mathbf{u}}, \cdots, \lambda_p^{\mathbf{u}})$ *is such that* $\phi(z) = 0$, *and let* $\mu_j(z)$, $j = 0, 1, \cdots, p+q$ *be computed according to* (40). *Then the costate* $\lambda^{\mathbf{u}}(t) = \sum_{j=0}^{p} \mu_j(z) \lambda_j^{\mathbf{u}}(t)$ *satisfies* (A.1) *with* $\lambda^0 = \mu_0(z) \geqq 0$, (A.2) *with* $\mu_j = \mu_j(z)$, $j = 0, 1, \cdots, p$, *and* (A.3) *with* $\mu_j = \mu_j(z)$, $j = p+1, \cdots, p+q$. *Furthermore,*

$$\int_0^1 \int_{R^m} \left\| \lambda^0 \left(\frac{\partial L}{\partial u}\right)^T (x^{\mathbf{u}}(t), u, t) + \left(\frac{\partial f}{\partial u}\right)^T (x^{\mathbf{u}}(t), u, t) \lambda^{\mathbf{u}}(t) \right\|^2 d\mathbf{u}(t)\, dt = 0. \tag{A.6}$$

Thus the condition $\phi(z) = 0$ is seen as a weak, or "differential," form of the relaxed minimum principle.

Appendix B. Convergence in L_2 and i.s.c.m. We will now present the link between the convergence of a sequence $\{u^i\}$ of controls in $L_2^m[0, 1] \cap L_\infty^m[0, 1]$ in the L_2-norm and the convergence of the associated sequence of measurable relaxed controls $\{\mathbf{u}^i\}$, in the sense of control measures. We first give the definition of almost uniform convergence of measurable function defined on a closed interval T.

DEFINITION B.1. A sequence of measurable functions, $\{u^i(\cdot)\}_{i=0}^{\infty}$, is said to be *almost uniformly convergent* to a measurable function $\bar{u}(\cdot)$ if for each $\delta > 0$ there is a set E_δ in T with $\mu(E_\delta) < \delta$ such that $u^i(\cdot)$ converges uniformly to $\bar{u}(\cdot)$ on T/E_δ.

The following theorem is found in Bartle [2, chap. 7, p. 75].

THEOREM B.1. *If a sequence of measurable functions, $\{u^i(\cdot)\}_{i=0}^{\infty}$, converges to a measurable function $\bar{u}(\cdot)$ in the L_2-norm, then there exists a subsequence which converges almost uniformly to $\bar{u}(\cdot)$.*

In the standard L_2 theory of convergence of optimal control algorithms, one assumes that the sequence of measurable controls $\{u^i(\cdot)\}_{i=0}^{\infty}$ constructed by an optimal control algorithm, has a subsequence which converges in the L_2-norm to a measurable function $\bar{u}(\cdot)$. Theorem B.1 shows that when the above assumption is made, it is automatically assumed that there exists a subsequence of $\{u^i(\cdot)\}_{i=0}^{\infty}$ which converges almost uniformly to the function $\bar{u}(\cdot)$.

The following theorem shows that almost uniform convergence of measurable controls implies i.s.c.m. convergence of the associated measurable relaxed controls.

THEOREM B.2. *Let $\{u^i\}_{i=0}^{\infty} \subset L_2^m[0, 1] \cap L_\infty^m[0, 1]$ be a sequence of uniformly bounded measurable controls which converges almost uniformly to $\bar{u}$, and let $\{\mathbf{v}^i\}_{i=0}^{\infty}, \bar{\mathbf{v}}$ be associated measurable relaxed controls. Then $\mathbf{v}^i$ converges i.s.c.m. to $\bar{\mathbf{v}}$.*

REFERENCES

[1] A. V. BALAKRISHNAN, *On a new computing technique in optimal control*, this Journal, 6 (1968), pp. 149–173.

[2] R. G. BARTLE, *The Elements of Integration*, John Wiley, New York, 1966.

[3] J. CULLUM, *Penalty functions and nonconvex continuous optimal control problems*, Computing Methods in Optimization–2, L. A. Zadeh, L. W. Neustadt and A. V. Balakrishnan, eds., Academic Press, New York, 1969, pp. 55–67.

[3a] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–80.

[4] R. V. GAMRELIDZE, *On some extremal problems in the theory of differential equations with application to the theory of optimal control*, this Journal, 3 (1965), pp. 106–128.

[5] D. H. JACOBSON AND D. Q. MAYNE, *Differential Dynamic Programming*, American Elsevier, New York, 1970.

[5a] E. J. MCSHANE, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.

[6] D. Q. MAYNE, *Properties of a cost function employed in a second order optimization algorithm*, J. Math. Anal. Appl., 38 (1972), pp. 42–52.

[7] D. Q. MAYNE AND E. POLAK, *First order strong variation algorithms for optimal control*, J. Optimization Theory Appl., 16 (1975), pp. 277–339.

[8] O. PIRONNEAU AND E. POLAK, *A dual method for optimal control problems with initial and final boundary constraints*, this Journal, 11 (1973), pp. 534–549.

[9] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[10] E. POLAK AND D. Q. MAYNE, *First order strong variation algorithms for optimal control problems with terminal inequality constraints*, J. Optimization Theory Appl., to appear.

[11] D. L. RUSSELL, *Penalty functions and bounded phase coordinate control*, J. Soc. Indust. Appl. Math. Ser. A: Control, 2 (1964), pp. 409–422.

[12] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[12a] ———, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1964), pp. 432–455.

[13] L. J. WILLIAMSON, *Convergence properties of optimal control algorithms*, Ph.D. thesis, University of California, Berkeley, 1973.

[14] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

BEHAVIOR OF OBSERVATIONS OF LINEAR CONTROL SYSTEMS*

ROGER C. McCANN†

Abstract. Given two linear control systems with reachable sets $R_1(t)$, $R_2(t)$, respectively, and a constant matrix M, directly verifiable necessary and sufficient conditions are obtained for the coincidence $R_1(t) = MR_2(t)$ for all $t > 0$.

1. Introduction. Consider a pair of linear, autonomous control systems described by

$$\dot{x}(t) = Ax(t) + u(t), \qquad u(t) \in U, \tag{1}$$

$$\dot{y}(t) = By(t) + v(t), \qquad u(t) \in V, \tag{2}$$

where the dimensions are n and m, respectively, and the constraint sets U, V are polytopes; i.e., U and V are each the convex span of a finite number of points. The reachable set and controllability space of (1) are

$$R_1(t) = \left\{ \int_0^t e^{-As} u(s)\, ds : \text{measurable } u : [0, t] \to U \right\}$$

and

$$\mathscr{C}_1 = \text{linear span of } U + AU + \cdots + A^{n-1}U.$$

The reachable set $R_2(t)$ and controllability space $\mathscr{C}_2$ of (2) are defined analogously.

An observed (linear, autonomous) control system may have various state space representations, as in

$$\dot{x}(t) = Ax(t) + u(t), \qquad u(t) \in U,$$
$$z(t) = Mx(t).$$

In interesting cases, the elements A, U, M and possibly even the dimension of x are not given in advance. Common to all descriptions of the observed system are the sets $MR_1(t)$, $t > 0$. (As the control functions $u(t)$ are not necessarily given in advance, one cannot necessarily identify the observations $z(t)$.)

It is natural to inquire into the following question: given two such systems, what conditions on the data A, B, U, V, M_1, M_2 are necessary and sufficient for the coincidence

$$M_1R_1(t) = M_2R_2(t) \quad \text{for all } t > 0. \tag{3}$$

As a first step, the case $M_1 = I = M_2$ has been studied in [3] and [4]; and the case that $M_1 = I$ and $M_2 = M$ has linearly independent columns (i.e., $x \to Mx$ is one-to-one) is treated in [1]. In the latter case, (3) holds, if, and only if,

$$U = MV \tag{4}$$

* Received by the editors March 17, 1975, and in revised form July 18, 1975.

† Department of Mathematics and Statistics, Case Western Reserve University, Cleveland, Ohio 44106.

and

$$AMv = MBv \quad \text{for all } v \in \mathscr{C}_2. \tag{5}$$

If M does not have linearly independent columns, then (4) and (5) are sufficient, but not necessary, for (3) as the following example shows.

Example 1. Let $n=1$, $m=2$, $A=[-2]$, $B=\begin{bmatrix}2 & 0\\ 0 & -1\end{bmatrix}$, $m=[2,2]$, $U=[-2,2]$ and V be the parallelogram in R^2 with vertices $\binom{1}{0}, \binom{-1}{0}, \binom{0}{2}, \binom{0}{-2}$. A simple calculation shows that $R_1(t)=MR_2(t)$ for all $t>0$, but $AM \neq MB$ on $\mathscr{C}_2 = R^2$.

This paper treats the next stage of the problem, in that $M_1 = I$ and $M_2 = M$ is allowed to represent a singular mapping. Essential to the study of this problem are the following properties of reachable sets, say, of system (1): the limit theorem

$$\lim_{t\to 0^+} \frac{R_1(t)}{t} = U \tag{6}$$

and the additive formula

$$R_1(t+s) = R_1(t) + e^{-At}R_1(s) \tag{7}$$

[2, Proposition 1 and Lemma 1].

2. Results.

LEMMA 1. *Let $S \subset R^m$ be a polytope and M an $n \times m$ matrix. Then*

(i) *MS is a polytope,*

(ii) *if z is an extreme point of MS, then there is an extreme point x of S such that $Mx = z$,*

(iii) *if $m = n$ and M is nonsingular, then x is an extreme point of S if, and only if, Mx is an extreme point of MS.*

Proof. Since S is the convex span of its extreme points $\{x_1, x_2, \cdots, x_k\}$, MS is the convex span of $\{Mx_1, Mx_2, \cdots, Mx_k\}$. The assertions follow directly.

LEMMA 2. *$R_1(t) = MR_2(t)$ for all $t>0$ if, and only if, there is an $a>0$ such that $e^{-At}U = Me^{-Bt}V$ for all $t \in [0, a)$. In particular, if $R_1(t) = MR_2(t)$ for all $t>0$, then $U = MV$.*

Proof. Suppose that $R_1(t) = MR_2(t)$ for all $t>0$. Applying the addition formula (7) we have

$$R_1(t) + e^{-At}R_1(s) = R_1(t+s) = MR_2(t+s) = MR_2(t) + Me^{-Bt}R_2(s) \tag{8}$$

for all t, $s>0$. Since all summands in (8) are compact and convex and $R_1(t) = MR_2(t)$, we conclude that $e^{-At}R_1(s) = Me^{-Bt}R_2(s)$ for all t, $s>0$. Applying the limit theorem (6) we have

$$e^{-At}U = \lim_{s\to 0} e^{-At}\frac{R_1(s)}{s} = \lim_{s\to 0} Me^{-Bt}\frac{R_2(s)}{s} = Me^{-Bt}V$$

for all $t>0$ and

$$U = \lim_{s\to 0}\frac{R_1(s)}{s} = \lim_{s\to 0} m\frac{R_2(s)}{s} = MV.$$

Now suppose that $e^{-At}U = Me^{-Bt}V$ on an interval $[0, a)$. Let $t_1 \in (0, a)$ and $u : [\cdot, t_1] \to U$ be measurable. Then there is a measurable function $v : [\cdot, t_1] \to V$ such that $e^{-As}u(s) = Me^{-Bs}v(s)$ for every $s \in [0, t_1]$. Hence $\int_0^t e^{-As}u(s)\,ds = M\int_0^t e^{-Bs}v(s)\,ds$. It follows that $MR_2(t) \supset R_1(t)$ for every $t \in (0, a)$. The opposite inclusion is proved in a similar fashion. Thus $R_1(t) = MR_2(t)$ for every $t \in (0, a)$. Then (8) is valid for all t, $s > 0$ such that $t + s < a$. A simple induction using (7) yields the desired result.

LEMMA 3. *Suppose that $R_1(t) = MR_2(t)$ for all $t > 0$. If u is an extreme point of U, then there exists an extreme point v of V such that $e^{-At}u = Me^{-Bt}v$ for all $t > 0$.*

Proof. Let W denote the set of all extreme points of V. For each $s \geqq 0$, the extreme points of $Me^{-Bs}V$ are among $\{Me^{-Bs}w : w \in W\}$. Also $e^{-As}u$ is an extreme point of $e^{-As}U$ for each $s \geqq 0$. Since $e^{-As}U = Me^{-Bs}V$ on a nonvoid interval $[0, a)$, we have that for any $s \in [0, a)$ there is a $w \in W$ with $e^{-As}u = Me^{-Bs}w$. Since there are only finitely many points w in W, we conclude that, for some $v \in W$, $e^{-As}u = Me^{-Bs}v$ holds for infinitely many $s \in [0, a)$. By analyticity $e^{-As}u = Me^{-Bs}v$ for all $s \in R^1$.

LEMMA 4. *If $u \in R^n$ and $v \in R^m$ are such that $A^j u = MB^j v$ for $j = 0, 1, 2, \cdots, m$, then $A^j u = MB^j v$ for $j = 0, 1, 2, \cdots$.*

Proof. We proceed by induction. Suppose that $A^j u = MB^j v$ for some $j \geqq m$. By the Cayley–Hamilton theorem, B^j may be written as a linear combination $\sum_{h=0}^{m-1} a_h B^h$ for some coefficients a_h. Then

$$\begin{aligned} A^{j+1}u &= AMB^j v = A\sum_{h=0}^{m-1} a_h MB^h v = A\sum_{h=0}^{m-1} a_h A^h Mv \\ &= \sum_{h=0}^{m-1} a_h A^{h+1} Mv = \sum_{h=0}^{m-1} a_h MB^{h+1} v = MB\sum_{h=0}^{m-1} a_h B^h v \\ &= MBB^j V = MB^{j+1} v. \end{aligned}$$

It follows that $A^j u = MB^j v$ for $j = 0, 1, 2, \cdots$.

THEOREM 5. *Let $\{u_1, u_2, \cdots, u_k\}$ be the extreme points of U. Then $R_1(t) = MR_2(t)$ for all $t > 0$ if, and only if, there exists $a > 0$ and extreme points $\{v_1, v_2, \cdots, v_k\}$ of V such that, for $t \in [0, a]$,*

(i) *$\{Me^{-Bt}v_i : i = 1, 2, \cdots, k\}$ are all the extreme points of $Me^{-Bt}V$,*

(ii) *$A^j u_i = MB^j v_i$ for $i = 1, 2, \cdots, k$ and $j = 0, 1, 2, \cdots, m$.*

Proof. Suppose that $R_1(t) = MR_2(t)$ for $t > 0$. Since U has only finitely many extreme points, the existence of points v_i with property (i) and such that $e^{-At}u_i = Me^{-Bt}v_i$ for $t > 0$ follows directly from Lemma 3. Then for $i = 1, 2, \cdots, k$,

$$A^j u_i = (-1)^j \frac{d^j}{dt^j}(e^{-At}u_i)\bigg|_{t=0} = (-u)^j \frac{d^j}{dt^j}(Me^{-Bt}v_i)\bigg|_{t=0} = MB^j v_i.$$

Now suppose properties (i) and (ii) hold. By Lemma 4, we have $A^j u_i = MB^j v_i$ for $i = 1, 2, \cdots, k$ and $j = 0, 1, 2, \cdots$. Hence for any $t \geqq 0$,

$$e^{-At}u_i = \sum_{j=0}^{\infty} \frac{(-1)^j A^j t^j}{j!} u_i = M\sum_{j=0}^{\infty} \frac{(-1)^j B^j t^j}{j!} v_i = Me^{-Bt}v_i.$$

Then

$$e^{-At}U = \operatorname{conv}\{e^{-At}u_i : i = 1, 2, \cdots, k\} = \operatorname{conv}\{Me^{-Bt}v_i : i = 1, 2, \cdots, k\} = Me^{-Bt}V$$

for every $t \geqq 0$. The desired result now follows directly from Lemma 2.

COROLLARY 6. *Let v_i be as in Theorem 5. If $R_1(t) = MR_2(t)$ for all $t > 0$, then for $i = 1, 2, \cdots, k$ and $j = 0, 1, 2, \cdots$,*

$$A^jMv_i = MB^jv_i.$$

Proof. From Theorem 5 (ii) we have $u_i = Mv_i$ when $j = 0$ and $MB^jv_i = A^ju_i = A^jMv_i$ for $j = 0, 1, \cdots, m$. The desired result follows from Lemma 4.

It is of interest to determine when conditions (4) and (5) are necessary for $R_1(t) = MR_2(t)$ for all $t > 0$. In what follows, $\tilde{V}$ will denote the convex span of v_1, $v_2, \cdots, v_k$, where the v_i are as in Theorem 5.

LEMMA 7. *Suppose $R_1(t) = MR_2(t)$ for all $t > 0$. Then every $v \in V$ can be written in the form $v = w_1 + w_2$, where $w_1 \in \tilde{V}$ and $w_2 \in \ker M \cap [V - \tilde{V}]$ (i.e., $V \subset \tilde{V} + \ker M$).*

Proof. Let $v \in V$. Since $\{Mv_i : i = 1, 2, \cdots, k\}$ are the extreme points of MV, there exists $a_1, a_2, \cdots, a_k \geqq 0$ such that $\sum_{i=1}^k a_i = 1$ and $Mv = \sum_{i=1}^k a_iMv_i$. Set $w_1 = \sum_{i=1}^k a_iv_i$ and $w_2 = v - w_1$. Evidently $w_1 \in \tilde{V}$ and $w_2 \in \ker M \cap [V - \tilde{V}]$.

THEOREM 8. *Suppose $R_1(t) = MR_2(t)$ for all $t > 0$. Then $AM = MB$ on $\mathscr{C}_2$ if, and only if,*

$$B^j(\ker M \cap [V - \tilde{V}]) \subset \ker M \tag{9}$$

for $j = 0, 1, 2, \cdots, m$.

Proof. Let $v \in V$. Then $v = w_1 + w_2$ where $w_1 \in \tilde{V}$ and $w_2 \in M \cap [V - \tilde{V}]$ (Lemma 7). Suppose (9) holds. Then for $j = 0, 1, 2, \cdots, m$ we have

$$A^jMv = A^jMw_1 + A^jMw_2 = A^jMw_1$$

and

$$MB^jv = MB^jw_1 + MB^jw_2 = MB^jw_1.$$

It follows directly from Corollary 6 that $A^jM = MB^j$ on $\tilde{V}$. Hence $A^jMv = A^jMw_1 = MB^jw_1 = MB^jv$. This proves $A^jM = MB^j$ on V for $j = 0, 1, 2, \cdots, m$. Let $w \in \mathscr{C}_2$. Then $w = \sum_{j=0}^{m-1} a_jB^jw_j$ for some $w_j \in V$ and $a_j \in R^1$. We have

$$MBw = \sum_{j=0}^{m-1} a_jMB^{j+1}w_j = \sum_{j=0}^{m-1} a_jA^{j+1}Mw_j = A\sum_{j=0}^{m-1} a_jA^jMw_j$$

$$= A\sum_{j=0}^{m-1} a_jMB^jw_j = AM\sum_{j=0}^{m-1} a_jB^jw_j = AMw.$$

Now suppose $AM = MB$ on $\mathscr{C}_2$. Since $V \subset \mathscr{C}_2$, $AM = MB$ on V. Suppose that $A^jM = MB^j$ on V. Then, for every $v \in V$,

$$MB^{j+1}v = MB(B^jv) = AMB^jv = AA^jMv = A^{j+1}Mv.$$

It follows that $MB^j = A^jM$ on V for $j = 0, 1, 2, \cdots$. Since $\tilde{V} \subset V$, we also have that $MB^j = A^jM$ on $V - \tilde{V}$. Let $w \in (\ker M \cap [V - \tilde{V}])$. Then $0 = A^jMw$, since $w \in \ker M$, and $A^jMw = MB^jw$, since $w \in V - \tilde{V}$. Hence $MB^jw = 0$. This verifies (9).

Henceforth $R_3(t)$ and $\mathscr{C}_3$ will denote the reachable set and controllability space of the control system

$$\dot{y}(t) = By(t) + v(t), \qquad v(t) \in \tilde{V}. \tag{10}$$

LEMMA 9. *If $R_1(t) = MR_2(t)$ for all $t > 0$, then $MR_2(t) = MR_3(t)$ for all $t > 0$.*

Proof. Since $\{v_i : i = 1, 2, \cdots, k\}$ are the extreme points of $\tilde{V}$, we have $\operatorname{conv}\{Me^{-Bs}v_i : i = 1, 2, \cdots, k\} = Me^{-Bs}\tilde{V}$ for every $s \geqq 0$. Also, by Theorem 5 (i), $\operatorname{conv}\{Me^{-Bs}v_i : i = 1, 2, \cdots, k\} = Me^{-Bs}V$ for every s in an interval of the form $[0, a)$. Hence for every $s \in [0, a)$, we have $Me^{-Bs}V = Me^{-Bs}\tilde{V}$. Let $t < a$. For every measurable $v : [0, t] \to V$, there exists a measurable $v_1 : [0, t] \to \tilde{V}$ such that $Me^{-Bs}v(s) = Me^{-Bs}v_1(s)$ for every $s \in [0, t]$. Then $M\int_0^t e^{-Bs}v(s)\,ds = M\int_0^t e^{-Bs}v_1(s)\,ds$. This proves $MR_2(t) \subset MR_3(t)$ for all $t \in [0, a)$. The opposite inclusion is proved similarly. Hence $MR_2(t) = MR_3(t)$ for all $t \in [0, a)$. An induction argument using the addition formula, (7), yields the desired result.

Assembling the properties of $\tilde{V}$ we have

THEOREM 10. *Suppose $R_1(t) = MR_2(t)$ for all $t > 0$; then there is a polytope $\tilde{V} \subset V$ such that*

(i) *the extreme points of $\tilde{V}$ are extreme points of V,*

(ii) *$R_1(t) = MR_3(t)$ for all $t > 0$,*

(iii) *$U = M\tilde{V}$,*

(iv) *$AMv = MBv$ for all $v \in \mathscr{C}_3$.*

Proof. Only (iv) remains unproved, and it follows directly from Theorem 8.

It should be noted that $\tilde{V}$ need not be unique.

Example 2. Let $A = [1]$, $B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $M = [1, 0]$, $U = [-1, 1]$ and V be the square in R^2 with vertices $\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}$. A short calculation shows that $e^{-At}U = Me^{-Bt}V$ for all $t \geqq 0$, that $Me^{-Bt}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = Me^{-Bt}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = e^{-t}$, and that $Me^{-Bt}\begin{pmatrix} -1 \\ 1 \end{pmatrix} = Me^{-Bt}\begin{pmatrix} -1 \\ -1 \end{pmatrix} = -e^{-t}$. Thus $\tilde{V}$ could be chosen as the convex of v_1 and v_2, where $v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ or $v_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $v_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ or $v_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$. There are four choices for $\tilde{V}$.

If v is an extreme point of V and Mv is an extreme point of MV, it is not necessarily true that $Me^{-Bt}v$ is an extreme point of $Me^{-Bt}V$.

Example 3. Let $A = [2]$, $B = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $M = [1 \quad -1]$, $U = [-1, 1]$ and V be the square in R^2 with vertices $\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}$. A short calculation shows that

$$Me^{-Bt}V = [-e^{-t}, e^{-t}] = e^{-At}U, \qquad Me^{-Bt}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = e^{-2t}, \qquad \text{and} \qquad Me^{-Bt}\begin{pmatrix} 0 \\ -1 \end{pmatrix} = e^{-t}.$$

$Me^{-Bt}\begin{pmatrix} 0 \\ -1 \end{pmatrix}$ is an extreme point of $Me^{-Bt}V$ for all $t \geqq 0$, while $Me^{-Bt}\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is an extreme point of $Me^{-Bt}V$ only when $t = 0$.

REFERENCES

[1] E. N. CHUKWU, *Symmetries of autonomous linear control systems*, this Journal, 12 (1974), pp. 436–448.
[2] O. HÁJEK, *On differentiability of the minimal time function*, Math. Systems Theory, to appear.
[3] ———, *Identification of control systems by performance*, Ibid., 5 (1971), pp. 349–352.
[4] M. L. J. HAUTUS AND G. J. OLSDER, *A uniqueness theorem for linear control systems with coinciding reachable set*, this Journal, 11 (1973), pp. 412–416.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

A BOUNDARY VALUE CRITERION FOR SEMIPASSIVE HILBERT PORTS*

L. P. D'AMATO† AND A. H. ZEMANIAN‡

Abstract. Necessary and sufficient conditions on the boundary values of an operator-valued analytic function Y defined on the open right half-plane are given which ensure that Y is the Laplace transform of the unit-impulse response of a linear time-invariant causal semipassive Hilbert port. Similar results are also obtained for the more restricted case where the Hilbert port is linear time-invariant and passive.

1. Introduction. The theory of linear passive systems encompasses various facets of a number of physical phenomena such as the dispersion of nuclear particles, the behavior of electrical networks and viscoelasticity. A recurrent theme in that theory is the characterization of a system function that is the Laplace transform of the unit-impulse response of such a system. See, for example, [1]–[10]. One way of doing this is by using the boundary values of the system function on the imaginary axis. Beltrami and Wohlers [1, Thm. 3.17] have established the principal result in this direction for the case where the system is a linear time-invariant passive n-port. The present work extends the result of Beltrami and Wohlers in two directions. It merely requires linearity, time-invariance, causality and semipassivity, this being a weaker set of restrictions than linearity, time-invariance and passivity. It also allows the system function to take its values in the space of continuous linear operators in an arbitrary complex Hilbert space; in other words, it encompasses Hilbert ports [10, § 4.2].

2. Some preliminary considerations. Let $\mathcal{U}$ and $\mathcal{V}$ be two topological linear spaces. $[\mathcal{U}; \mathcal{V}]$ will denote the linear space of all continuous linear mappings supplied with the bounded topology. We will also employ at one point a weaker topology, namely, the pointwise topology of $[\mathcal{U}; \mathcal{V}]$. For definitions of these topologies, see [10, p. 208]. According to this notation, if H is a complex Hilbert space, $[H; H]$ is the space of continuous linear operators in H supplied with the uniform operator topology. $(\cdot, \cdot)$ denotes the inner product for H. On the other hand, $\langle f, \phi \rangle$ denotes the value that a distribution f assigns to a testing function ϕ.

$\mathcal{D}(H)$ is the space of smooth (i.e., continuous derivatives of all orders) H-valued testing functions on the real line R; $\mathcal{D}(H)$ is assigned its customary Schwartz topology [10; p. 50]. When H happens to be the complex plane C, $\mathcal{D}(C)$ is denoted by $\mathcal{D}$. $\mathcal{S}$ is the customary testing-function space of smooth complex-valued functions of rapid descent on R; $\mathcal{S}$ too has its usual topology [10, p. 66]. Thus if A is any complex Banach space, $[\mathcal{D}; A]$ is the space of all A-valued distributions on R and $[\mathcal{S}; A]$ is the space of all A-valued tempered distributions on R.

* Received by the editors February 4, 1975. This work was supported by NSF Grant GP P033568-X001.

† Computer Services Directorate, Naval Air Test Center, Patuxent River, Maryland 20670.

‡ Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York 11794.

$\mathfrak{N}$ will denote the immittance operator of a Hilbert port; in particular, $\mathfrak{N}$ is a mapping of $\mathscr{D}(H)$ into $[\mathscr{D}; H]$. The linearity, time-invariance and causality of $\mathfrak{N}$ are defined in the customary way [10, pp. 93, 112 and 196]. As for semipassivity and passivity, we have the following definitions. $\mathfrak{N}$ is called semipassive on $\mathscr{D}(H)$ if for every $v \in \mathscr{D}(H)$ and for $u = \mathfrak{N}v$, we have that $(u(\cdot), v(\cdot))$ is Lebesgue integrable on R and

$$\operatorname{Re} \int_{-\infty}^{\infty} (u(t), v(t))\, dt \geqq 0.$$

$\mathfrak{N}$ is called passive on $\mathscr{D}(H)$ if in addition

$$\operatorname{Re} \int_{-\infty}^{T} (u(t), v(t))\, dt \geqq 0$$

for every $T \in R$. Thus the passivity of $\mathfrak{N}$ on $\mathscr{D}(H)$ implies its semipassivity on $\mathscr{D}(H)$; however, the converse is not true in general [10, p. 151].

The following are known facts about an operator $\mathfrak{N}$ from $\mathscr{D}(H)$ into $[\mathscr{D}; H]$. If $\mathfrak{N}$ is linear and semipassive, then it is continuous. If $\mathfrak{N}$ is linear and passive, then it is causal. $\mathfrak{N}$ is linear, continuous and time-invariant if and only if it is a convolution operator $y *$, where y is an $[H; H]$-valued distribution on R; that is, $\mathfrak{N}v$ is the convolution product $y * v$ for at least every $v \in \mathscr{D}(H)$. Moreover, $\mathfrak{N} = y *$ is causal if and only if the support of y is contained in the semi-infinite closed interval $[0, \infty)$. (Proofs of these results are given in [10, §§ 5.10, 5.11, 8.2 and 8.3].)

We shall make use of a representation due to Hackenbroch [2, Thm. 3.6], which can be stated as follows. (An integral sign without limits, as in (2) below, will denote an integration over all of R.)

Proposition. $\mathfrak{N}$ is a linear, time-invariant causal semipassive mapping on $\mathscr{D}(H)$ if and only if $\mathfrak{N} = y *$, where

$$y = \sum_{k=0}^{n} P_k D^k \delta + j(0) D^{p-1} \delta + j1_+ - D^p(j1_+). \tag{1}$$

Here, n is a finite nonnegative integer. $P_k \in [H; H]$ and $P_k = (-1)^{k+1} P_k'$, where the prime denotes the adjoint operator. $p = 2m$, where m is a positive odd integer. δ is the delta functional. D^k denotes kth order generalized differentiation. 1_+ is the function on $-\infty < t < \infty$ equal to 1 for $t \geqq 0$ and to 0 for $t < 0$. j is the $[H; H]$-valued function on R defined by

$$j(x) = \int dP_\eta\, e^{i\eta x}, \tag{2}$$

where $P_\eta = P$ is a positive-operator measure on the Borel subsets of R. That is, P maps each Borel subset into a positive member of $[H; H]$ and satisfies the customary axioms for an operator-valued measure (see [10, p. 26]). $j1_+$ denotes the function $t \to j(t)1_+(t)$.

Moreover, $\mathfrak{N}$ is a linear time-invariant passive mapping on $\mathscr{D}(H)$ if and only if $\mathfrak{N} = y *$, where y is given by (1) with the additional restrictions that $n = 1$, $p = 2$ and P_k is a positive operator for $k = 1$.

3. An exchange formula. $\tilde{f}$ and $\hat{f}$ will denote, respectively, the direct and inverse Fourier transforms of any tempered vector-valued or scalar-valued distribution f, where $\int g(t)\, e^{-i\omega t}\, dt$ is taken as the form for the classical Fourier transform of a function g. Our objective in this section is to establish the exchange formula

$$\tilde{j}\tilde{1}_+ = \frac{1}{2\pi}\, \tilde{j} * \tilde{1}_+, \tag{3}$$

where as before $*$ denotes convolution.

The generalized derivative DP of the positive-operator measure P is a distribution in $[\mathscr{S}; [H; H]]$ defined by

$$\langle DP, \theta \rangle = \int dP_\eta \theta(\eta), \qquad \theta \in \mathscr{S}.$$

Indeed, the integral on the right-hand side exists for every $\theta \in \mathscr{S}$ and satisfies the following inequality [10, § 2.2]:

$$\left\| \int dP_\eta \theta(\eta) \right\| \leqq P(R) \sup |\theta(\eta)|. \tag{4}$$

(Here, $\| \cdot \|$ denotes the norm in $[H; H]$.)

Therefore we may apply the definition of the distributional inverse Fourier transformation to DP to obtain

$$\langle \widehat{DP}, \theta \rangle = \langle DP, \hat{\theta} \rangle = \int dP_\eta \frac{1}{2\pi} \int \theta(t)\, e^{i\eta t}\, dt.$$

By virtue of [10, Thm. 2.5-2], we may interchange the order of integration to get

$$\frac{1}{2\pi} \int dt \int dP_\eta\, e^{i\eta t} \theta(t) = \frac{1}{2\pi} \langle j, \theta \rangle.$$

Thus $j = 2\pi \widehat{DP}$, or equivalently, $\tilde{j} = 2\pi DP \in [\mathscr{S}; [H; H]]$.

It is also a fact that

$$\tilde{1}_+(\eta) = \pi\delta(\eta) + Pv\frac{1}{i\eta},$$

where $Pv(1/i\eta)$ is the pseudofunction defined by the Cauchy principal value of $-i \int \eta^{-1} \phi(\eta)\, d\eta$, $\phi \in \mathscr{D}$. [9, p. 18]. Therefore we may formally write

$$\frac{1}{2\pi}(\tilde{j} * \tilde{1}_+)(\eta) = (DP_\eta) * \left[\pi\delta(\eta) + Pv\frac{1}{i\eta} \right]. \tag{5}$$

Obviously, $(DP) * \delta$ has a sense and is a member of $[\mathscr{D}; [H; H]]$. On the other hand, by the customary definition of a convolution,

$$\left\langle (DP_\eta) * Pv\frac{1}{i\eta}, \phi(\eta) \right\rangle = \left\langle DP_\eta, \left\langle Pv\frac{1}{i\omega}, \phi(\eta + \omega) \right\rangle \right\rangle,$$

where $\phi \in \mathscr{D}$. We wish to show that the right-hand side has a sense and that $(DP_\eta) * Pv(1/i\eta)$ is a member of $[\mathscr{D}; [H; H]]$.

Set

$$\psi(\eta) = \left\langle Pv\frac{1}{i\omega}, \phi(\eta+\omega)\right\rangle. \tag{6}$$

ψ is a regularization and is therefore a smooth function. Moreover, by the definition of $Pv(1/i\eta)$,

$$\psi(\eta) = \int_{-1}^{1} \frac{\phi(\eta+\omega)-\phi(\eta)}{i\omega}\,d\omega + \left(\int_{-\infty}^{-1} + \int_{1}^{\infty}\right)\frac{\phi(\eta+\omega)}{i\omega}\,d\omega. \tag{7}$$

Clearly, the first integral on the right-hand side has a bounded support. Also, for each η, the second integral is bounded by

$$\sup_{t\in R}|\phi(t)|\int_{I_\eta}\frac{d\omega}{|\omega|}, \tag{8}$$

where I_η is the intersection of the support of $\phi(\eta+\cdot\,)$ with the complement of the interval between -1 and 1. Therefore that second integral in (7) tends to zero as $|\eta|\to\infty$. These results show that $\langle DP_\eta, \psi(\eta)\rangle$, which is equal to $\int dP_\eta\psi(\eta)$, exists and is a member of $[H; H]$.

Next, replace ψ by ψ_k and ϕ by ϕ_k in (6) and (7), where $k = 1, 2, \cdots$. Assume that $\phi_k \to 0$ in $\mathscr{D}$. By the standard property of a regularization, the first integral in (7) tends to zero uniformly on every compact interval. Moreover, there is a single interval containing the support of the first integral for every value of k. On the other hand, the bound (8) shows that the second integral tends to zero uniformly on R. Therefore so too does ψ_k. The estimate (4) now implies that $\langle DP_\eta, \psi_k(\eta)\rangle \to 0$ under the uniform operator topology. So truly, $(DP_\eta) * Pv(1/i\eta)$ is a member of $[\mathscr{D}; [H; H]]$. Thus, by (5), $\tilde{j} * \tilde{1}_+$ also exists as a member of $[\mathscr{D}; [H; H]]$.

We next note that j is a bounded strongly continuous function on R and may therefore be multiplied by 1_+; moreover, $j1_+$ is a member of $[\mathscr{S}; [H; H]]$. (See [10, § 8.10]). Therefore we may apply the distributional Fourier transformation to $j1_+$.

We are finally ready to establish (3). For any $\theta \in \mathscr{S}$,

$$\langle \widetilde{j1}_+, \theta\rangle = \langle j1_+, \tilde{\theta}\rangle = \int_0^\infty j(x)\tilde{\theta}(x)\,dx = \int_0^\infty dx \int dP_\eta\, e^{i\eta x}\tilde{\theta}(x).$$

By [10, Thm. 2.5-2] again, we may interchange the order of integration to get

$$\begin{aligned}\int dP_\eta \int_0^\infty dx\tilde{\theta}(x)\, e^{i\eta x} &= \int dP_\eta \langle 1_+(x)\, e^{i\eta x}, \tilde{\theta}(x)\rangle \\ &= \langle DP_\eta, \langle \tilde{1}_+(\omega-\eta), \theta(\omega)\rangle\rangle = \langle DP_\eta * \tilde{1}_+(\eta), \theta(\eta)\rangle \\ &= \left\langle \frac{1}{2\pi}\tilde{j} * \tilde{1}_+, \theta\right\rangle.\end{aligned}$$

This proves (3).

4. A realizability theorem. In view of our preceding results, we may apply the Fourier transformation to (1) to obtain

$$\tilde{y}(\omega) = \sum_{k=0}^{n} (i\omega)^k P_k + (i\omega)^{p-1} j(0) + \frac{1}{2\pi}[1-(i\omega)^p](\tilde{j} * \tilde{1}_+)(\omega).$$

This may be converted into the following expression by using (5) and the facts that $j(0) = P(R)$ and p is two times an odd positive integer.

$$(9) \quad \tilde{y}(\omega) = \sum_{k=0}^{n} (i\omega)^k P_k + i\omega^{p-1} P(R) + (1+\omega^p)\left[\pi D P_\omega + (DP_\omega) * Pv\frac{1}{i\omega}\right].$$

This result coupled with the proposition yields the following.

THEOREM 1. *$\mathfrak{N}$ is a linear time-invariant causal semipassive mapping on $\mathcal{D}(H)$ if and only if $\mathfrak{N} = y*$, where the Fourier transform $\tilde{y}$ of y is given by* (9). *Furthermore, $\mathfrak{N}$ is a linear time-invariant passive mapping on $\mathcal{D}(H)$ if and only if $\mathfrak{N} = y*$, where the Fourier transform $\tilde{y}$ of y is given by* (9) *with $n = 1$, $p = 2$ and P_1 being a positive operator.*

Other realizability theorems can be obtained by taking the Laplace transform Y of (1). See [10, Thms. 8.10-2 and 8.12-1]. In this case, Y is an $[H; H]$-valued analytic function on H and has the following representation for each ζ in the open right half complex plane C_+.

$$(10) \qquad Y(\zeta) = \sum_{k=0}^{n} P_k \zeta^k + \int dP_\eta \frac{1 - i\eta\zeta^{p-1}}{\zeta - i\eta}, \qquad \zeta \in C_+.$$

Every $[H; H]$-valued analytic function Y on C_+ having the representation (10) will be called semipositive*, and it will be called positive* in the special case where $n = 1$, $p = 2$ and P_1 is positive. Thus $\mathfrak{N}$ has the properties indicated in the first or second sentences of Theorem 1 if and only if $\mathfrak{N} = y*$, where the Laplace transform of y is either semipositive* or positive*, respectively.

5. A boundary value criterion. We can in turn obtain necessary and sufficient conditions for Y to be semipositive* or positive* by characterizing the boundary values of Y on the imaginary axis in conjunction with a growth condition on the half-plane C_+.

An $[H; H]$-valued analytic function F on C_+ will be said to be of polynomial growth if, for every closed half-plane $C_a = \{\zeta : \operatorname{Re}\zeta \geqq a > 0\}$, there exists a polynomial P_a (depending in general on a) such that

$$\|F(\zeta)\| \geqq P_a(|\zeta|)$$

for all $\zeta \in C_a$. It is a fact that every such F is a Laplace transform with a region of definition that contains C_+, and, conversely, every Laplace transform whose region of definition contains C_+ is of polynomial growth [10, § 6.5].

We will need two lemmas. Their proofs are much like that of the first theorem in [1] and are therefore omitted.

LEMMA 1. *Let $f \in [\mathcal{S}; [H; H]]$ and assume that the support of f is contained in $[0, \infty)$. Then f has a Laplace transform, whose region of definition contains C_+. Moreover, as $\sigma \to 0+$, $F(\sigma + i\cdot) \to \tilde{f}$ in the bounded topology of $[\mathcal{S}; [H; H]]$.*

LEMMA 2. *Let F be an $[H; H]$-valued analytic function on C_+ of polynomial growth. Assume that, as $\sigma \to 0+$, $F(\sigma + i \cdot)$ converges in the pointwise topology of $[\mathcal{S}; [H; H]]$. Then the limit is $\tilde{f}$, the Fourier transform of the inverse Laplace transform of F.*

Now, assume that Y is semipositive*. Then Y is a Laplace transform on C_+ and is therefore of polynomial growth. Moreover, the inverse Laplace transform y or Y has the representation (1) and therefore satisfies the hypothesis of Lemma 1. So, as $\sigma \to 0+$, $Y(\sigma + i \cdot)$ converges in the bounded topology of $[\mathcal{S}; [H; H]]$ to the Fourier transform (9) of y. Conversely, assume Y satisfies the hypothesis of Lemma 2. Assume furthermore that its boundary value on the imaginary axis has the form (9). Then by Lemma 2, that boundary value is the Fourier transform of the inverse Laplace transform y of Y. Consequently, y has the representation (1), and Y is semipositive*. Similar arguments hold in the special case where Y is positive*. Thus we have proven our second realizability theorem.

THEOREM 2. *Y is a semipositive* $[H; H]$-valued function if and only if Y is an $[H; H]$-valued analytic function on C_+ of polynomial growth and, as $\sigma \to 0+$, $Y(\sigma + i \cdot)$ converges in the pointwise topology of $[\mathcal{S}; [H; H]]$ to a limit having the representation* (9). *This theorem remains true when we replace "semipositive*" by "positive*" and let $n = 1$, $p = 2$ and P_1 be positive in the representation* (9). *It also remains true when "pointwise topology" is replaced by "bounded topology".*

REFERENCES

[1] E. J. BELTRAMI AND M. R. WOHLERS, *Distributions and the Boundary Values of Analytic Functions*, Academic Press, New York, 1966.

[2] W. HACKENBROCH, *Integraldarstellung einer Klasse dissipativer linearer Operatoren*, Math. Z., 109 (1969), pp. 273–287.

[3] H. KÖNIG AND J. MEIXNER, *Lineare Systeme und lineare Transformationen*, Math. Nachr., 19 (1958), pp. 265–322.

[4] R. W. NEWCOMB, *Linear Multiport Synthesis*, McGraw-Hill, New York, 1966.

[5] J. S. TOLL, *Causality and the dispersion relation: Logical foundations*, Phys. Rev. A., 104 (1956), pp. 1760–1770.

[6] V. S. VLADIMIROV, *On the theory of linear passive systems*, Soviet Math. Dokl., 10 (1969), pp. 733–736.

[7] ———, *Linear passive systems*, Theoretical and Mathematical Physics (Academy of Sciences of the U.S.S.R.), 1 (1969), pp. 67–94.

[8] D. C. YOULA, L. J. CASTRIOTA AND H. J. CARLIN, *Bounded real scattering matrices and the foundations of linear passive network theory*, IEEE Trans. Circuit Theory, CT-6 (1959), pp. 102–124.

[9] A. H. ZEMANIAN, *Distribution Theory and Transform Analysis*, McGraw-Hill, New York, 1965.

[10] ———, *Realizability Theory for Continuous Linear Systems*, Academic Press, New York, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 4, July 1976

CONTROLLABILITY SUBSPACES AND FEEDBACK SIMULATION*

MICHAEL HEYMANN†

Abstract. The concepts of input chain and controllability chain are introduced, and the structure of controllability subspaces of a linear system is investigated. It is shown that the input and controllability chains are the fundamental feedback invariants of a linear system.

The feedback simulation problem (a generalization of the feedback equivalence problem) is defined and solved.

1. Introduction.

Consider a time-invariant, linear multivariable system

$$\Sigma \qquad \dot{x} = Ax + Bu,$$

where $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ ($m \leqq n$) are, respectively, the state and control vectors, and A and B are constant real matrices with B of full rank. Assume Σ is controllable and let

$$\Sigma' \qquad \dot{z} = A'z + B'w$$

be another controllable linear system where $z \in \mathbb{R}^{n'}$ and $w \in \mathbb{R}^{m'}$ ($m' \leqq n'$) are the state and control vectors for Σ', and A' and B' are also constant and real matrices with B' of full rank. Suppose there exists a triple (F, G, T) of real matrices, where F is $m \times n$, G is $m \times m'$ and T is $n' \times n$ such that the system

$$\dot{x} = (A + BF)x + BGw,$$
$$y = Tx$$

has exactly the same input–output behavior as the system Σ'; that is, given that $x(0) = 0$ and $z(0) = 0$ and given any input function $w(t)$, then the responses $z(t)$ and $y(t)$ are identical for all $t \geqq 0$. It is then said that Σ' can be (*feedback*) *simulated* by Σ. The class of all controllable systems Σ' which can be simulated by a given controllable system Σ is called the *simulation orbit* of Σ and is denoted $O\{\Sigma\}$. In view of the controllability condition imposed on elements in $O\{\Sigma\}$ and the rank condition on B (and B') it is readily verified that $\Sigma' \in O\{\Sigma\}$, if and only if there exists a triple (F, G, T) with G of full column rank and T of full row rank such that

$$\begin{aligned} T(A + BF) &= A'T, \\ TBG &= B'. \end{aligned} \tag{1.1}$$

Let $E\{\Sigma\}$ denote the subset of $O\{\Sigma\}$ consisting of all elements $\Sigma' \in O\{\Sigma\}$ for which $n' = n$ and $m' = m$. The relation $\Sigma' \in E\{\Sigma\}$ is then an equivalence relation (that is, $\Sigma \in E\{\Sigma\}$, $\Sigma' \in E\{\Sigma\}$ implies $\Sigma \in E\{\Sigma'\}$, and $\Sigma' \in E\{\Sigma\}$ and $\Sigma'' \in E\{\Sigma'\}$ implies $\Sigma'' \in E\{\Sigma\}$) and $E\{\Sigma\}$ is an equivalence class. Accordingly,

* Received by the editors May 12, 1975, and in revised form October 20, 1975.

† Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. This research was supported in part by the National Research Council of Canada Grant A-7399. This work was done while the author was on leave with the Control Group, Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada.

two systems Σ and Σ' have been called *feedback equivalent* if and only if $\Sigma' \in E\{\Sigma\}$. The problem of characterizing $E\{\Sigma\}$ for any given controllable linear system Σ was first solved by Brunovsky [1] who showed that $E\{\Sigma\}$ is uniquely and completely determined by a list of positive integers called the *controllability indices* of Σ. Subsequently, this problem received a great deal of attention in the literature and various authors studied the relation of the controllability indices to other concepts in linear system theory (see, e.g., [2], [3], [4] and [5]).

An interesting and more difficult problem than that of characterizing $E\{\Sigma\}$, is that of characterizing $O\{\Sigma\}$. This problem has not been treated in the literature to date and is one of the main objects of the present paper.

In our initial examination of the feedback simulation problem, it became apparent that a major role is played by certain "chains" of subspaces which are uniquely specified by the given parent system Σ. Central therein is the concept of controllability subspace, first introduced by Wonham and Morse in [6]. While this concept was originally defined in connection with a specific regulator synthesis problem (that of decoupling with pole assignment), it has since become increasingly evident that controllability subspaces are major building blocks in a variety of synthesis problems of linear feedback systems and, in fact, have a direct bearing on the "structural modifiability" of a given linear system (see, e.g., [4], [7]). Nevertheless, it seems that many questions pertaining to the structure of controllability subspaces remain as yet unanswered, and their properties are still not very well understood.

In [2], it was shown that there is a close relation between the controllability indices and the controllability subspaces of a given system Σ. Specifically, it was shown that given a controllable system Σ, the state space can be decomposed into a direct sum of singly-generated controllability subspaces whose (ordered) list of dimensions is precisely equal to the (ordered) list of the controllability indices of Σ. However, although this list of dimensions has been shown to be uniquely specified by the system (and is actually a complete invariant for $E\{\Sigma\}$), the decomposition itself has no uniqueness properties and many decompositions with the same dimension list can, in general, be exhibited. This fact strengthens the suspicion that in the study of feedback (in contrast to realization theory), direct sum decompositions of the state space into singly generated controllability subspaces are *not* the "right way" to split the system into its elementary structural components. Still, one might expect that some kind of decomposition (or partial decomposition) of the state space into controllability subspaces would be unique and invariant under feedback.

It will be shown in the present paper that there exists a natural "chain" of controllability subspaces, called the *controllability chain*, which is uniquely specified by the underlying system and is invariant under feedback. It will also be shown that the controllability chain is strongly related to another natural chain of subspaces called the *input chain*, which is also unique and feedback invariant. Moreover, the two lists of dimensions of the subspaces in these chains are in one-to-one correspondence and each is derivable from the other. These dimension lists are also in one-to-one correspondence with the list of controllability indices of the system and hence are each a complete invariant for $E(\Sigma)$. Much of the paper is devoted to a study of the structure of the input and controllability chains and

their properties. The results are then applied to the solution of the feedback simulation problem.

2. Notation. The field of real numbers will be denoted $\mathbb{R}$. Script capital letters $\mathscr{B}, \mathscr{X}, \mathscr{Y}, \cdots$ will denote finite-dimensional vector spaces over $\mathbb{R}$ with elements $b, x, y, \cdots$. The dimension of $\mathscr{X}$ will be denoted dim $(\mathscr{X})$. The vector space spanned by a set of vectors $\{x_1, x_2, \cdots\}$ will be denoted sp $\{x_1, x_2, \cdots\}$. The zero vector, zero space, $\cdots$ will be denoted 0. Capital Roman letters $A, B, \cdots$ will be used to denote linear maps. (While no notational distinction will be made between a map and its matrix, it will be clear from the context whenever reference is made to a matrix.) If $M: \mathscr{X} \to \mathscr{Y}$ is a linear map, we denote the image of M by Im (M), the nullspace of M by ker (M), and if $\mathscr{U} \subset \mathscr{X}$ is a subspace, then $M\mathscr{U} \triangleq \text{sp}\{y \in \mathscr{Y} | y = Mu, u \in \mathscr{U}\}$. The restriction of M to $\mathscr{U}$ will be denoted $M|\mathscr{U}$, the codomain being taken as $M\mathscr{U}$. If $\mathscr{Y} = \mathscr{X}$ and $M\mathscr{U} \subset \mathscr{U}$, then the codomain of $M|\mathscr{U}$ will be understood to be $\mathscr{U}$. A sequence of subspaces $\{\mathscr{U}_i\}_{i=1}^{\infty}$, $\mathscr{U}_i \subset \mathscr{X}$, is called a chain in case $U_i \subset U_{i+1}$ for all $i \geqq 1$ and is denoted $[\mathscr{U}_i]$. The least integer k for which $\mathscr{U}_{k+j} = \mathscr{U}_k$ for all $j > 0$ is called the *length* of the chain, and the subspace $\mathscr{U}_k$ is called the *limit* of the chain. Sometimes we will denote a chain $[\mathscr{U}_i]_1^k$ to emphasize that its length is k. If $\mathscr{B} \subset \mathscr{X}$ is a subspace, we denote by $\mathscr{M}(\mathscr{B})$ the class of all linear maps $M: \mathscr{X} \to \mathscr{X}$ for which Im $(M) \subset \mathscr{B}$. Finally, for an integer $k > 0$, we write $\underline{k} \triangleq \{1, 2, \cdots, k\}$.

3. Preliminaries. Let Σ be a given linear system and let $\mathscr{X} = \mathbb{R}^n$ and $\mathscr{U} = \mathbb{R}^m$ denote the state and input spaces, respectively. Henceforth A and B will be generally regarded as linear maps rather than matrices, and denote $\mathscr{B} = \text{Im}\,(B)$. Write

$$\langle A|\mathscr{B}\rangle_i = \mathscr{B} + A\mathscr{B} + \cdots + A^{i-1}\mathscr{B}, \qquad i = 1, 2, \cdots. \tag{3.1}$$

It is readily noted that the following simple relations hold:

$$\langle A|\mathscr{B}\rangle_i = \mathscr{B} + A\langle A|\mathscr{B}\rangle_{i-1}, \qquad i = 2, 3, \cdots, \tag{3.2}$$

$$\langle A|\mathscr{B}\rangle_i = A^{i-1}\mathscr{B} + \langle A|\mathscr{B}\rangle_{i-1}, \qquad i = 2, 3, \cdots, \tag{3.3}$$

$$\langle A|\mathscr{B}\rangle_{n+j} = \langle A|\mathscr{B}\rangle_n, \qquad j = 1, 2, \cdots. \tag{3.4}$$

Hence, the sequence $\{\langle A|\mathscr{B}\rangle_i\}_1^{\infty}$ is a chain which we call the *fundamental chain* of Σ. The length k of the fundamental chain then satisfies $k \leqq n$, and we denote by $\langle A|\mathscr{B}\rangle$ its limit which is simply the controllable subspace of Σ. The following lemma summarizes some immediate consequences of (3.2) and (3.3).

LEMMA 3.1. *Let $A: \mathscr{X} \to \mathscr{X}$ be a linear map and let $\mathscr{V} \subset \mathscr{X}$ and $\mathscr{B} \subset \mathscr{X}$ be any subspaces. Then for any $\hat{A} \in \mathscr{M}(\mathscr{B})$ the following hold:*

(i) $$\langle A|\mathscr{B}\rangle_i = \langle A + \hat{A}|\mathscr{B}\rangle_i, \qquad i = 1, 2, \cdots,$$

(ii) $$\langle A + \hat{A}|\mathscr{V}\rangle_{i+1} \subset \langle A|\mathscr{V}\rangle_{i+1} + \langle A|\mathscr{B}\rangle_i, \qquad i = 1, 2, \cdots,$$

(iii) $$\langle A + \hat{A}|\mathscr{V}\rangle_{i+1} \subset \langle A|\mathscr{B}\rangle_i, \quad \textit{if and only if}$$
$$\langle A|\mathscr{V}\rangle_{i+1} \subset \langle A|\mathscr{B}\rangle_i.$$

By definition [6], a subspace $\mathscr{R} \subset \mathscr{X}$ is a *controllability* subspace of Σ, provided there exists a linear map $F: \mathscr{X} \to \mathscr{U}$ such that $\mathscr{R} = \langle A + BF|\mathscr{R} \cap \mathscr{B}\rangle$. Clearly, for any map F, Im $(BF) \subset \mathscr{B}$, and it is also well known (see, e.g., [8])

that if $\hat{A}: \mathscr{X} \to \mathscr{X}$ is any linear map such that $\hat{A} \in \mathscr{M}(\mathscr{B})$, then there exists a map $F: \mathscr{X} \to \mathscr{U}$ such that $\hat{A} = BF$. Hence, the system Σ is completely specified (in so far as controllability subspaces are concerned) by the pair $(A, \mathscr{B})$, and from now on we regard a linear system as such a pair. A subspace $\mathscr{R} \subset \mathscr{X}$ is then a controllability subspace of $(A, \mathscr{B})$, whenever there exists a map $\hat{A} \in \mathscr{M}(\mathscr{B})$ such that $\langle A + \hat{A} | \mathscr{R} \cap \mathscr{B} \rangle = \mathscr{R}$.

The following lemma is elementary.

LEMMA 3.2. *Let $A: \mathscr{X} \to \mathscr{X}$ be a linear map and let $\mathscr{V} \subset \mathscr{X}$ and $\mathscr{B} \subset \mathscr{X}$ be subspaces. Then there exists $\hat{A} \in \mathscr{M}(\mathscr{B})$ such that $\mathscr{V} \subset \ker (A + \hat{A})$ if and only if $A\mathscr{V} \subset \mathscr{B}$.*

Proof. If $\hat{A}$ exists, then $A\mathscr{V} \subset \mathscr{B}$ is immediate. Conversely, assume $A\mathscr{V} \subset \mathscr{B}$, and write $\mathscr{X} = \mathscr{B} \oplus \mathscr{W}$ for some subspace $\mathscr{W} \subset \mathscr{X}$. Let P be the projection of $\mathscr{X}$ onto $\mathscr{B}$ along $\mathscr{W}$. Then $PA \in \mathscr{M}(\mathscr{B})$ and $\mathscr{V} \subset \ker (A - PA)$. □

THEOREM 3.3. *Let $(A, \mathscr{B})$ be given and let $\mathscr{R} \subset \mathscr{X}$ be a subspace. Then $\mathscr{R}$ is a controllability subspace of (A, B) if and only if given any $\mathscr{W} \subset \mathscr{B}$ such that $\mathscr{W} \oplus (\mathscr{R} \cap \mathscr{B}) = \mathscr{B}$, there exists $\hat{A} \in \mathscr{M}(\mathscr{W})$ so that $\mathscr{R} = \langle A + \hat{A} | \mathscr{R} \cap \mathscr{B} \rangle$.*

Proof. The "if" part is immediate since $\mathscr{M}(\mathscr{W}) \subset \mathscr{M}(\mathscr{B})$ for any $\mathscr{W} \subset \mathscr{B}$. To prove the "only if" part, assume $\mathscr{R}$ is a controllability subspace. Then there exists $\hat{A} \in \mathscr{M}(\mathscr{B})$ such that $\mathscr{R} = \langle A + \hat{A} | \mathscr{R} \cap \mathscr{B} \rangle$. Let $\mathscr{W}$ satisfy $\mathscr{W} \oplus (\mathscr{R} \cap \mathscr{B}) = \mathscr{B}$ and write $\mathscr{X} = \mathscr{B} \oplus \mathscr{S}$ for some subspace $\mathscr{S} \subset \mathscr{X}$. Let P be the projection of $\mathscr{X}$ onto $\mathscr{R} \cap \mathscr{B}$ along $\mathscr{W} \oplus \mathscr{S}$. Then $P\hat{A} \in \mathscr{M}(\mathscr{R} \cap \mathscr{B})$ so that by (i) of Lemma 3.1,

$$\mathscr{R} = \langle A + \hat{A} - P\hat{A} | \mathscr{R} \cap \mathscr{B} \rangle = \langle A + (I - P)\hat{A} | \mathscr{R} \cap \mathscr{B} \rangle$$

and it is readily noted that $\operatorname{Im} (I - P)\hat{A} \subset \mathscr{W}$. □

We conclude this section with two well-known and important facts, the proof of which can be found, e.g., in [7].

THEOREM 3.4. *Let $(A, \mathscr{B})$ be given and let $b \in \mathscr{B}$ be any vector. Then for an integer $k > 0$ there exists $\hat{A} \in \mathscr{M}(\mathscr{B})$ such that $(A + \hat{A})^k b = 0$, if and only if $A^k b \in \langle A | \mathscr{B} \rangle_k$.*

THEOREM 3.5. *Let $(A, \mathscr{B})$ be given. Then the class $\mathscr{C}(A, \mathscr{B})$ of controllability subspaces in $\langle A | \mathscr{B} \rangle$ is closed under subspace addition. Hence any subspace $\mathscr{S} \subset \langle A | \mathscr{B} \rangle$ contains a unique supremal controllability subspace.*

4. The input and controllability chains. In the present section, we will investigate some basic properties of the fundamental chain of a linear system $(A, \mathscr{B})$.

Fix an integer $k \geqq 1$, and consider the class of all subspaces $\mathscr{V} \subset \mathscr{B}$ that satisfy

$$\langle A | \mathscr{V} \rangle_{k+1} \subset \langle A | \mathscr{B} \rangle_k. \tag{4.1}$$

It is readily verified that this class is closed under subspace addition, and hence for each $k \geqq 1$, there exists a unique supremal subspace of $\mathscr{B}$ for which (4.1) holds. Accordingly denote

$$\mathscr{B}_k \triangleq \sup \{\mathscr{V} \subset \mathscr{B} | \langle A | \mathscr{V} \rangle_{k+1} \subset \langle A | \mathscr{B} \rangle_k\}, \tag{4.2}$$

and consider the sequence $\{\mathscr{B}_k\}$. This sequence is clearly a chain, since

$$\langle A|\mathscr{B}_k\rangle_{k+2} = \mathscr{B}_k + A\langle A|\mathscr{B}_k\rangle_{k+1} \subset \mathscr{B}_k + A\langle A|\mathscr{B}\rangle_k \subset \langle A|\mathscr{B}\rangle_{k+1},$$

and hence $\mathscr{B}_k \subset \mathscr{B}_{k+1}$ for each $k \geqq 1$. We call this chain of subspaces the *input chain* of $(A, \mathscr{B})$. The following theorem summarizes several interesting properties of the input chain.

THEOREM 4.1. *Let $(A, \mathscr{B})$ be a given linear system. Then there exists a unique chain of subspaces $[\mathscr{B}_k]$, $\mathscr{B}_k \subset \mathscr{B}$, called the input chain of $(A, \mathscr{B})$, with the following properties*:

(i) *for each $k \geqq 1$, $\mathscr{B}_k$ satisfies* (4.2);

(ii) *the input chain is feedback invariant, i.e., it is the same for all systems $(A + \hat{A}, \mathscr{B})$, $\hat{A} \in \mathscr{M}(\mathscr{B})$;*

(iii) *the length of the input chain is the same as the length of the fundamental chain, and $\lim [\mathscr{B}_k] = \mathscr{B}$;*

(iv) *set $\gamma_0 = 0$ and for each $k \geqq 1$, let $\gamma_k \triangleq \dim(\mathscr{B}_k)$. Then for $k \geqq 1$,*

$$\dim(\langle A|\mathscr{B}\rangle_k) = k \cdot \dim(\mathscr{B}) - \sum_{i=0}^{k-1} \gamma_i. \tag{4.3}$$

Proof. (i) follows from the definition, and (ii) is an immediate consequence of Lemma 3.1. To see (iii), let k^* be the length of the fundamental chain. Then $\langle A|\mathscr{B}\rangle_{k^*+1} = \langle A|\mathscr{B}\rangle_{k^*}$, so that $\mathscr{B}_{k^*} = \mathscr{B}$ and consequently, $\lim [\mathscr{B}_k] = \mathscr{B}$ and $s \leqq k^*$, where s is the length of $[\mathscr{B}_k]$. That $s = k^*$ is an immediate consequence of the definition of k^* and (4.1). To see (iv), first note that it trivially holds for $k = 1$. Assume now that (iv) holds for all $k = 1, 2, \cdots, t-1$, and write $\mathscr{B} = \mathscr{B}_t \oplus \mathscr{B}^t$ for some subspace $\mathscr{B}^t \subset \mathscr{B}$. Then

$$\langle A|\mathscr{B}\rangle_t = A^{t-1}\mathscr{B} + \langle A|\mathscr{B}\rangle_{t-1} = A^{t-1}\mathscr{B}^t \oplus \langle A|\mathscr{B}\rangle_{t-1},$$

and hence

$$\begin{aligned}\dim(\langle A|\mathscr{B}\rangle_t) &= \dim(A^{t-1}\mathscr{B}^t) + \dim(\langle A|\mathscr{B}\rangle_{t-1}) \\ &= \dim(\mathscr{B}^t) + \dim(\langle A|\mathscr{B}\rangle_{t-1}) = t\dim(\mathscr{B}) - \sum_{i=0}^{t-1} \gamma_i,\end{aligned}$$

as required. □

In view of the feedback invariance of the input chain of $(A, \mathscr{B})$, it is clear that the sequence $\{\gamma_i\}$, where the γ_i's are defined as in Theorem 4.1 (iv), is also feedback invariant. We call this sequence the *input list* of $(A, \mathscr{B})$ and denote it sometimes by $\{\gamma_i\}_s$ to indicate that we refer to the truncated sequence $\{\gamma_1, \cdots, \gamma_s\}$.

Let $(A, \mathscr{B})$ be a given linear system, and let $\mathscr{V} \subset \mathscr{B}$ be any subspace. For each integer $k > 0$, define

$$\mathscr{V}_k(\mathscr{B}) \triangleq \sup\{\mathscr{U} \subset \mathscr{V} | \langle A|\mathscr{U}\rangle_{k+1} \subset \langle A|\mathscr{B}\rangle_k\}. \tag{4.4}$$

Clearly, $\mathscr{V}_k(\mathscr{B})$ is uniquely determined for each k and the sequence $\{\mathscr{V}_k(\mathscr{B})\}$ is a chain. We call this chain the $(A, \mathscr{B})$-*induced chain* of $\mathscr{V}$. It is readily noted that $\mathscr{V}_k(\mathscr{B}) = \mathscr{V} \cap \mathscr{B}_k$, and hence the length $s_{\mathscr{B}}(\mathscr{V})$ of the $(A, \mathscr{B})$-induced chain of $\mathscr{V}$ is less than or equal to the length of the input chain of $(A, \mathscr{B})$. Define $\gamma_0(\mathscr{V}) = 0$, and for each $k > 0$, let $\gamma_k(\mathscr{V}) \triangleq \dim(\mathscr{V}_k(\mathscr{B}))$. The list $\{\gamma_k(\mathscr{V})\}$ is called the $(A, \mathscr{B})$-

induced list of $\mathscr{V}$, and the integer $L_{\mathscr{B}}(\mathscr{V})$ defined by

$$L_{\mathscr{B}}(\mathscr{V}) \triangleq \sum_{i=1}^{s_{\mathscr{B}}(\mathscr{V})} i[\gamma_i(\mathscr{V}) - \gamma_{i-1}(\mathscr{V})]$$

is called the $(A, \mathscr{B})$-*cover dimension* of $\mathscr{V}$. We will see that the cover dimension is the least dimension which a controllability subspace $\mathscr{R}$ of $(A, \mathscr{B})$ must possess in order for $\mathscr{V}$ to be contained in $\mathscr{R}$. Moreover, we will see that given any $\mathscr{V} \subset \mathscr{B}$, there always exist controllability subspaces containing $\mathscr{V}$ whose dimension equals the cover dimension. However, while the cover dimension is uniquely determined by $\mathscr{V}$, it is not true, in general, that there exists a unique controllability subspace containing $\mathscr{V}$ with this dimension. Subspaces with this property are quite special and we will return to this question later.

An immediate consequence of the definition of cover dimension is the following.

LEMMA 4.2. *Let $(A, \mathscr{B})$ be given and let $\mathscr{V}$ and $\mathscr{W}$ be subspaces of $\mathscr{B}$. If $\mathscr{V} \subset \mathscr{W}$ properly, then $L_{\mathscr{B}}(\mathscr{V}) < L_{\mathscr{B}}(\mathscr{W})$.*

We now have the following.

THEOREM 4.3. *Let $(A, \mathscr{B})$ be given, let $\mathscr{V} \subset \mathscr{B}$ be a subspace, and let $\mathscr{R}$ be a controllability subspace of $(A, \mathscr{B})$ such that $\mathscr{V} \subset \mathscr{R}$. Then* $\dim(\mathscr{R}) \geqq L_{\mathscr{B}}(\mathscr{V})$. *Moreover, if $\hat{A} \in \mathscr{M}(\mathscr{B})$ is any map such that $\mathscr{R} = \langle A + \hat{A} | \mathscr{R} \cap \mathscr{B} \rangle$, then* $\dim(\mathscr{R}) = L_{\mathscr{B}}(\mathscr{V})$ *if and only if the following two conditions hold*:

(i) $\mathscr{V} = \mathscr{R} \cap \mathscr{B}$,

(ii) $\langle A + \hat{A} | \mathscr{V}_k(\mathscr{B}) \rangle_{k+1} \subset \langle A + \hat{A} | \mathscr{V} \rangle_k$ *for all* $k \in s_{\mathscr{B}}(\mathscr{V})$, *where* $[\mathscr{V}_k(\mathscr{B})]$ *is the* $(A, \mathscr{B})$*-induced chain of* $\mathscr{V}$.

Proof. First observe that $L_{\mathscr{B}}(\mathscr{V})$ can also be expressed as

$$L_{\mathscr{B}}(\mathscr{V}) = s_{\mathscr{B}}(\mathscr{V}) \cdot \dim(\mathscr{V}) - \sum_{i=0}^{s_{\mathscr{B}}(\mathscr{V})-1} \gamma_i(\mathscr{V}).$$

Let $\hat{A} \in \mathscr{M}(\mathscr{B})$ be any map such that $\mathscr{R} = \langle A + \hat{A} | \mathscr{R} \cap \mathscr{B} \rangle$, and let $[\mathscr{V}_k^*]_t$ be the input chain of $(A + \hat{A}, \mathscr{V})$. Then by Lemma 3.1, $\mathscr{V}_k^* \subset \mathscr{V}_k(\mathscr{B})$, and hence $\dim(\mathscr{V}_k^*) \leqq \dim(\mathscr{V}_k(\mathscr{B}))$ for all $k > 0$. Also, $\mathscr{V} = \mathscr{V}_t^* = \mathscr{V}_{s_{\mathscr{B}}(\mathscr{V})}(\mathscr{B})$, so that $t \geqq s_{\mathscr{B}}(\mathscr{V})$. Consequently,

$$\begin{aligned} \dim(\mathscr{R}) &= \dim(\langle A + \hat{A} | \mathscr{R} \cap \mathscr{B} \rangle) \geqq \dim(\langle A + \hat{A} | \mathscr{V} \rangle) \\ &= t \cdot \dim(\mathscr{V}) - \sum_{k=1}^{t-1} \dim(\mathscr{V}_k^*) \geqq t \cdot \dim(\mathscr{V}) - \sum_{k=1}^{t-1} \dim(\mathscr{V}_k(\mathscr{B})) \\ &= s_{\mathscr{B}}(\mathscr{V}) \cdot \dim(\mathscr{V}) - \sum_{k=0}^{s_{\mathscr{B}}(\mathscr{V})-1} \gamma_k(\mathscr{V}) = L_{\mathscr{B}}(\mathscr{V}). \end{aligned} \tag{4.5}$$

Assume now that $\dim(\mathscr{R}) = L_{\mathscr{B}}(\mathscr{V})$. Then also $L_{\mathscr{B}}(\mathscr{R} \cap \mathscr{B}) = \dim(\mathscr{R})$, and hence by Lemma 4.2, (i) must hold. By (4.5), $\dim(\mathscr{R}) = L_{\mathscr{B}}(\mathscr{V})$ also implies that $\dim(\mathscr{V}_k^*) = \dim(\mathscr{V}_k(\mathscr{B}))$ for all k, and hence $\mathscr{V}_k^* = \mathscr{V}_k(\mathscr{B})$ for all k. But this together with (i) implies that (ii) holds. Conversely, observe that (i) and (ii) imply that $\dim(\mathscr{R}) \leqq L_{\mathscr{B}}(\mathscr{V})$, so that equality must hold. □

THEOREM 4.4. *Let $(A, \mathscr{B})$ be given, and let $\mathscr{V} \subset \mathscr{B}$ be any subspace. Then there exists a controllability subspace $\mathscr{R}$ such that $\mathscr{V} \subset \mathscr{R}$ and* $\dim(\mathscr{R}) = L_{\mathscr{B}}(\mathscr{V})$.

Proof. Let $[\mathscr{V}_k(\mathscr{B})]$ be the $(A, \mathscr{B})$-induced chain of $\mathscr{V}$. First note that $\langle A|\mathscr{V}_1(\mathscr{B})\rangle_2 \subset \mathscr{B}$ and hence $A\mathscr{V}_1(\mathscr{B}) \subset \mathscr{B}$. By Lemma 3.2, there exists $\hat{A} \in \mathscr{M}(\mathscr{B})$ such that $\mathscr{V}_1(\mathscr{B}) \subset \ker(A + \hat{A})$. Setting $\mathscr{R}_1 = \langle A + \hat{A}|\mathscr{V}_1(\mathscr{B})\rangle$, it follows that $\dim(\mathscr{R}_1) = L_{\mathscr{B}}(\mathscr{V}_1(\mathscr{B})) = \gamma_1(\mathscr{V}) - \gamma_0(\mathscr{V})$, and by Theorem 4.3, $\mathscr{R}_1 \cap \mathscr{B} = \mathscr{V}_1(\mathscr{B})$. We proceed by induction on k and assume there exists $\hat{A} \in \mathscr{M}(\mathscr{B})$ such that $\mathscr{R}_{k-1} \triangleq \langle A + \hat{A}|\mathscr{V}_{k-1}(\mathscr{B})\rangle$ satisfies $\dim(\mathscr{R}_{k-1}) = L_{\mathscr{B}}(\mathscr{V}_{k-1}(\mathscr{B})) = \sum_{i=1}^{k-1} i(\gamma_i(\mathscr{V}) - \gamma_{i-1}(\mathscr{V}))$. Write $\mathscr{V}_k(\mathscr{B}) = \mathscr{V}_{k-1}(\mathscr{B}) \oplus \mathscr{V}^*$ for some $\mathscr{V}^* \subset \mathscr{B}$, and let $v_1^*, \cdots, v_q^*$ be a basis for $\mathscr{V}^*$, where $q = \gamma_k(\mathscr{V}) - \gamma_{k-1}(\mathscr{V})$. Let $\hat{A}_i \in \mathscr{M}(\mathscr{B})$ be chosen according to Theorem 3.4, and denote $\mathscr{E}_i = \langle A + \hat{A}_i|\operatorname{sp}\{v_i^*\}\rangle$, $i = 1, \cdots, q$. Then $\dim(\mathscr{E}_i) \leqq k$, and by Theorem 3.5, the subspace $\mathscr{R}_k \triangleq \mathscr{R}_{k-1} + \mathscr{E}_1 + \cdots + \mathscr{E}_q$ is also a controllability subspace of $(A, \mathscr{B})$. Moreover, $\mathscr{V}_k(\mathscr{B}) \subset \mathscr{R}_k$ and

$$\begin{aligned}\dim(\mathscr{R}_k) &\leqq \dim(\mathscr{R}_{k-1}) + \sum_{i=1}^{q} \dim(\mathscr{E}_i) \\ &\leqq L_{\mathscr{B}}(\mathscr{V}_{k-1}(\mathscr{B})) + k(\gamma_k(\mathscr{V}) - \gamma_{k-1}(\mathscr{V})) = L_{\mathscr{B}}(\mathscr{V}_k(\mathscr{B})).\end{aligned}$$

Combining the last inequality with Theorem (4.3) completes the proof. □

Consider now the fundamental chain $[\langle A|\mathscr{B}\rangle_k]$ of a linear system $(A, \mathscr{B})$. By Theorem 3.5, each subspace $\langle A|\mathscr{B}\rangle_k$ contains a unique supremal controllability subspace $\mathscr{S}_k$, and it is clear that the sequence $\{\mathscr{S}_k\}$ constitutes a chain. We call this chain of controllability subspaces the *controllability chain* of $(A, \mathscr{B})$. It is quite natural to expect that there must be a fundamental connection between the input chain and the controllability chain. This is indeed the case, and we devote the remainder of this section to an investigation of the controllability chain.

LEMMA 4.5. *Let $(A, \mathscr{B})$ be given and let $t > 0$ be any integer. Assume there exists $\hat{A} \in \mathscr{M}(\mathscr{B})$ such that $\langle A + \hat{A}|\mathscr{B}_t\rangle_{t+1} = \langle A + \hat{A}|\mathscr{B}_t\rangle_t$. Let $[\mathscr{B}_k^*]$ be the input chain of $(A + \hat{A}, \mathscr{B}_t)$. Then $\mathscr{B}_k^* = \mathscr{B}_k$ for all $k \in \underline{t}$, and hence $\dim(\langle A + \hat{A}|\mathscr{B}_t\rangle) = L_{\mathscr{B}}(\mathscr{B}_t)$.*

Proof. That for all $k \in \underline{t}$, $\mathscr{B}_k^* \subset \mathscr{B}_k$, is an immediate consequence of Lemma 3.1. Also, $\langle A + \hat{A}|\mathscr{B}_t\rangle_{t+1} = \langle A + \hat{A}|\mathscr{B}_t\rangle_t$ implies that $\langle A + \hat{A}|\mathscr{B}_k\rangle_{k+1} \subset \langle A + \hat{A}|\mathscr{B}_t\rangle_k$ for all $k < t$. Hence $\mathscr{B}_k \subset \mathscr{B}_k^*$, and we have $\mathscr{B}_k^* = \mathscr{B}_k$ for all $k \in \underline{t}$. Combining this fact with Theorem 4.3(ii) yields that $\dim(\langle A + \hat{A}|\mathscr{B}_t\rangle) = L_{\mathscr{B}}(\mathscr{B}_t)$, and the proof is complete. □

LEMMA 4.6. *Let $(A, \mathscr{B})$ be given and fix $k > 0$. Let $\mathscr{R}$ be any controllability subspace in $\langle A|\mathscr{B}\rangle_k$. Then $\mathscr{R} \cap \mathscr{B} \subset \mathscr{B}_k$, and hence $\dim(\mathscr{R}) \leqq L_{\mathscr{B}}(\mathscr{B}_k)$.*

Proof. If $\mathscr{R}$ is a controllability subspace, then $A\mathscr{R} \subset \mathscr{R} + \mathscr{B}$, and hence $\langle A|\mathscr{R}\rangle_{j+1} \subset \mathscr{R} + \langle A|\mathscr{B}\rangle_j$ for all $j > 0$. Consequently, $\mathscr{R} \subset \langle A|\mathscr{B}\rangle_k$ implies

$$\langle A|\mathscr{R} \cap \mathscr{B}\rangle_{k+1} \subset \langle A|\mathscr{R}\rangle_{k+1} \subset \mathscr{R} + \langle A|\mathscr{B}\rangle_k = \langle A|\mathscr{B}\rangle_k,$$

so that $\mathscr{R} \cap \mathscr{B} \subset \mathscr{B}_k$.

If $\mathscr{R} \cap \mathscr{B} \neq \mathscr{B}_k$, let $\mathscr{V} \subset \mathscr{B}_k$ satisfy $\mathscr{R} \cap \mathscr{B}_k \oplus \mathscr{V} = \mathscr{B}_k$, and let $v_1, \cdots, v_q$ be a basis for $\mathscr{V}$. For each v_i, let $\hat{A}_i$ satisfy Theorem 3.4 and let $\mathscr{E}_i \triangleq \langle A + \hat{A}_i|\operatorname{sp}\{v_i\}\rangle$. Then $\mathscr{E}_i \subset \langle A|\mathscr{B}_k\rangle$ for all $i \in \underline{q}$, and the controllability subspace $\mathscr{R}^* = \mathscr{R} + \mathscr{E}_1 + \cdots + \mathscr{E}_q$ is also in $\langle A|\mathscr{B}\rangle_k$. Moreover, $\mathscr{R}^* \cap \mathscr{B} = \mathscr{B}_k$. The proof will be completed by showing that $\dim(\mathscr{R}^*) = L_{\mathscr{B}}(\mathscr{B}_k)$, and then since $\mathscr{R} \subset \mathscr{R}^*$, it will follow that $\dim(\mathscr{R}) \leqq L_{\mathscr{B}}(\mathscr{B}_k)$. Let $\hat{A} \in \mathscr{M}(\mathscr{B})$ satisfy $\mathscr{R}^* = \langle A + \hat{A}|\mathscr{B}_k\rangle$. By Lemma 4.5, we will be done if we can show that $\langle A + \hat{A}|\mathscr{B}_k\rangle_{k+1} =$

$\langle A + \hat{A}|\mathscr{B}_k\rangle_k = \mathscr{R}^*$. Write $\mathscr{B} = \mathscr{B}_k \oplus \mathscr{B}^k$ for some subspace $\mathscr{B}^k \subset \mathscr{B}$. Then surely $\langle A + \hat{A}|\mathscr{B}_k\rangle_{k+1} \subset \langle A + \hat{A}|\mathscr{B}_k\rangle_k + \langle A + \hat{A}|\mathscr{B}^k\rangle_k$. Assume there exist vectors $u \in \langle A + \hat{A}|\mathscr{B}_k\rangle_{k+1}$, $v \in \langle A + \hat{A}|\mathscr{B}_k\rangle_k$ and $w \in \langle A + \hat{A}|\mathscr{B}^k\rangle_k$ such that $u = v + w$. Clearly both u and v are in $\mathscr{R}^*$, so that also $w \in \mathscr{R}^*$. But, if $w \neq 0$, then for some integer $j > 0$, $(A + \hat{A})^j w \notin \langle A|\mathscr{B}\rangle_k$, a contradiction. Hence w must be zero, and the proof is complete. □

For a given system $(A, \mathscr{B})$, a collection of controllability subspaces $\mathscr{R}_1, \cdots, \mathscr{R}_k$ is called *compatible* [7], if and only if there exists $\hat{A} \in \mathscr{M}(\mathscr{B})$ such that $\mathscr{R}_i = \langle A + \hat{A}|\mathscr{R}_i \cap \mathscr{B}\rangle$ for all $i \in \underline{k}$.

We now state the main result of this section.

THEOREM 4.7. *Let $(A, \mathscr{B})$ be a given system and let $[\mathscr{S}_k]$ be its controllability chain. Then*

(i) *for each $k > 0$, $\mathscr{S}_k \cap \mathscr{B} = \mathscr{B}_k$;*

(ii) *for each $k > 0$, $\mathscr{S}_k$ is the unique controllability subspace which contains $\mathscr{B}_k$ and has the (minimal) cover dimension $L_{\mathscr{B}}(\mathscr{B}_k)$;*

(iii) *the controllability subspaces $\mathscr{S}_1, \mathscr{S}_2, \cdots$ of the controllability chain are compatible and hence induce a unique partial decomposition in $(A, \mathscr{B})$.*

Proof. By Theorem 4.3, a necessary condition for a controllability subspace $\mathscr{R}$ containing $\mathscr{B}_k$ to have dimension $L_{\mathscr{B}}(\mathscr{B}_k)$ is that

$$\mathscr{R} \cap \mathscr{B} = \mathscr{B}_k$$

and

$$\mathscr{R} = \langle A + \hat{A}|\mathscr{B}_k\rangle_{k+1} = \langle A + \hat{A}|\mathscr{B}_k\rangle_k \subset \langle A|\mathscr{B}\rangle_k$$

for some $\hat{A} \in \mathscr{M}(\mathscr{B})$. By Theorem 4.4, such a controllability subspace indeed exists. By Lemma 4.6, $L_{\mathscr{B}}(\mathscr{B}_k)$ is an upper bound on the dimension of controllability subspaces in $\langle A|\mathscr{B}\rangle_k$. Hence by the supremality of $\mathscr{S}_k$, (i) and (ii) follow.

While (iii) is a consequence of results in [7], in view of the fact that the $\mathscr{S}_i$ form a chain, (see, in particular, p. 245 therein), we shall give a detailed proof for the sake of completeness.

Let s be the length of the input chain $[\mathscr{B}_k]$. Then for any $\hat{A} \in \mathscr{M}(\mathscr{B})$, $\langle A + \hat{A}|\mathscr{B}_s\rangle_{s+1} = \langle A + \hat{A}|\mathscr{B}_s\rangle_s$ by Lemma 3.1. Hence $\hat{A} \in \mathscr{M}(\mathscr{B})$ can surely be chosen such that $\mathscr{S}_s = \langle A + \hat{A}|\mathscr{B}_s\rangle_s$ and $\mathscr{S}_{s-1} = \langle A + \hat{A}|\mathscr{B}_{s-1}\rangle_{s-1}$. We proceed by induction and show that if $\mathscr{S}_s, \cdots, \mathscr{S}_{k+1}$ are compatible, then so are $\mathscr{S}_s, \cdots, \mathscr{S}_k$. Accordingly, assume that $\hat{A} \in \mathscr{M}(\mathscr{B})$ satisfies

$$\mathscr{S}_j = \langle A + \hat{A}|\mathscr{B}_j\rangle_j \quad \text{for } j = s, s-1, \cdots, k+1,$$

and define

$$\mathscr{B}_k^* = \sup\{\mathscr{V} \subset \mathscr{B}_{k+1} | \langle A + \hat{A}|\mathscr{V}\rangle_{k+1} \subset \langle A + \hat{A}|\mathscr{B}_{k+1}\rangle_k\}.$$

If $\mathscr{R}$ is the supremal controllability subspace of $(A + \hat{A}, \mathscr{B}_{k+1})$ which is contained in $\langle A + \hat{A}|\mathscr{B}_{k+1}\rangle_k$, then $\mathscr{R} \cap \mathscr{B}_{k+1} = \mathscr{B}_k^*$ and there exists $\hat{\hat{A}} \in \mathscr{M}(\mathscr{B}_{k+1})$ such that $\mathscr{R} = \langle A + \hat{A} + \hat{\hat{A}}|\mathscr{B}_k^*\rangle_{k+1} = \langle A + \hat{A} + \hat{\hat{A}}|\mathscr{B}_k^*\rangle_k$. Now by Lemma 4.5, $\mathscr{B}_k^* = \mathscr{B}_k$, and hence $\mathscr{R} = \mathscr{S}_k$. Since $\hat{\hat{A}} \in \mathscr{M}(\mathscr{B}_{k+1}) \subset \mathscr{M}(\mathscr{B}_j)$ for all $j > k+1$, the feedback invariance implies the desired compatibility result, and the proof is complete. □

Remark 4.8. The dimensions μ_k of the controllability subspaces $\mathscr{S}_k$ are given by

$$\mu_k = \sum_{j=1}^{k} j(\gamma_j - \gamma_{j-1}), \tag{4.6}$$

where $[\gamma_j]$ is the input list of $(A, \mathscr{B})$. Hence, the list $[\mu_j]$ is in one-to-one correspondence with the input list, and each can be computed from the other. We call the list $[\mu_j]$ the *controllability list* of $(A, \mathscr{B})$. As can be expected (and will be seen in the next section), this list is also in one-to-one correspondence with the list of controllability indices of $(A, \mathscr{B})$ which so far have been considered to be *the* fundamental feedback invariants of a system $(A, \mathscr{B})$. Yet, while the controllability list is tied with a unique and naturally defined chain of controllability subspaces, the list of controllability indices is associated with a (somewhat) arbitrary direct sum decomposition of $\langle A|\mathscr{B}\rangle$ into singly generated controllability subspaces. As will be seen later, it seems much more natural to regard the input list or the controllability list as the *fundamental* feedback invariants of which the list of controllability indices is a derived quantity.

Remark 4.9. Property (ii) of Theorem 4.7 deserves some special attention. Since the class of controllability subspaces $(A, \mathscr{B})$ is not closed under subspace intersection, there do not, in general, exist infimal controllability subspaces which contain given subspaces. Indeed, even the $\mathscr{S}_i$ are not infimal elements in the class of controllability subspaces which contain the $\mathscr{B}_i$.

Theorem 4.7 yields a number of immediate corollaries which were some of the results of the recent paper by Warren and Eckberg [9].

COROLLARY 4.10. *Let $\mathscr{R}$ be any controllability subspace of a given system $(A, \mathscr{B})$, and let $k = k(\mathscr{R})$ be the least integer such that $\mathscr{R} \subset \mathscr{S}_k$. Then*

$$k \leqq \dim(\mathscr{R}) \leqq \mu_k.$$

Proof. That $\dim(\mathscr{R}) \leqq \mu_k$ is immediate since $\mathscr{R} \subset \mathscr{S}_k$ and $\dim(\mathscr{S}_k) = \mu_k$. If $\dim(\mathscr{R}) = t < k$, then

$$\mathscr{R} = \langle A + \hat{A}|\mathscr{R} \cap \mathscr{B}\rangle_t \subset \langle A|\mathscr{B}\rangle_t$$

for some $\hat{A} \in \mathscr{M}(\mathscr{B})$ and hence $\mathscr{R} \subset \mathscr{S}_t$, a contradiction with the minimality of k. □

COROLLARY 4.11. *Let $(A, \mathscr{B})$ be given. If for some k, $\dim(\mathscr{S}_k) < k - 1$, there exists no controllability subspace $\mathscr{R}$ such that $\dim(\mathscr{S}_{k-1}) < \dim(\mathscr{R}) < k$.*

Proof. If $\dim(\mathscr{R}) < k$, then $\mathscr{R} \subset \mathscr{S}_{k-1}$. Hence $\dim(\mathscr{R}) \leqq \dim(\mathscr{S}_{k-1})$. □

COROLLARY 4.12. *Let $(A, \mathscr{B})$ be given. If for some k, $\dim(\mathscr{S}_{k-1}) < k$, then $\mathscr{S}_{k-1}$ is the unique controllability subspace $\mathscr{R}$ of $(A, \mathscr{B})$ that satisfies $\dim(\mathscr{R}) = L_{\mathscr{B}}(\mathscr{B}_{k-1})$.*

Proof. If for any controllability subspace $\mathscr{R}$ of $(A, \mathscr{B})$, $\dim(\mathscr{R}) < k$, then $\mathscr{R} \subset \mathscr{S}_{k-1} \subset \langle A|\mathscr{B}\rangle_{k-1}$. By the supremality of $\mathscr{S}_{k-1}$ in $\langle A|\mathscr{B}\rangle_{k-1}$, uniqueness of dimension follows. □

5. Splitting decompositions. It is well known [7] and is also an easy consequence of Theorem 3.3, that if $\mathscr{R}_1, \cdots, \mathscr{R}_k$ is any collection of linearly independent controllability subspaces of a system $(A, \mathscr{B})$, then this collection is compatible.

If $\mathscr{R}_1 \oplus \cdots \oplus \mathscr{R}_k = \langle A|\mathscr{B}\rangle$ is a direct sum decomposition of $\langle A|\mathscr{B}\rangle$ into linearly independent controllability subspaces, it does not necessarily follow, however, that $(\mathscr{R}_1 \cap \mathscr{B}) \oplus \cdots \oplus (\mathscr{R}_k \cap \mathscr{B}) = \mathscr{B}$. Conversely, if $\mathscr{V}_1 \oplus \cdots \oplus \mathscr{V}_k = \mathscr{B}$ is a direct sum decomposition of $\mathscr{B}$, it does not necessarily follow that there exists a direct sum decomposition of $\langle A|\mathscr{B}\rangle$ into controllability subspaces $\mathscr{R}_1, \cdots, \mathscr{R}_k$ such that $\mathscr{R}_i \cap \mathscr{B} = \mathscr{V}_i$ for all $i \in \underline{k}$. Accordingly, we have the following:

DEFINITION 5.1. A direct sum decomposition $\mathscr{R}_1 \oplus \cdots \oplus \mathscr{R}_k = \langle A|\mathscr{B}\rangle$ of $\langle A|\mathscr{B}\rangle$ into controllability subspaces is called a *split* (or *splitting decomposition*) of $\langle A|\mathscr{B}\rangle$, if and only if $(\mathscr{R}_1 \cap \mathscr{B}) \oplus \cdots \oplus (\mathscr{R}_k \cap \mathscr{B}) = \mathscr{B}$. A direct sum decomposition $\mathscr{V}_1 \oplus \cdots \oplus \mathscr{V}_k = \mathscr{B}$ of $\mathscr{B}$ is called a *split* of $\mathscr{B}$, if and only if there exist linearly independent controllability subspaces $\mathscr{R}_1, \cdots, \mathscr{R}_k$ such that $\mathscr{R}_i \cap \mathscr{B} = \mathscr{V}_i$ for all $i \in \underline{k}$.

The interest in splitting decompositions is self-evident since, in effect, a split corresponds to a decomposition of the system into a set of completely independent subsystems. Hence, splits should be of interest in connection with "decentralization" problems. For one instance in which splitting decompositions (into two subsystems) received some attention, the reader is referred to [10].

In the present section, we shall characterize splitting decompositions and investigate some of the consequences of their existence. In particular, we shall examine the connection between the controllability chain and the corresponding controllability list, and the decomposition of the system into a direct sum of singly generated controllability subspaces and the controllability indices.

THEOREM 5.2. *Let $(A, \mathscr{B})$ be given. Then a direct sum decomposition*

$$\mathscr{B} = \mathscr{V}_1 \oplus \cdots \oplus \mathscr{V}_k$$

of $\mathscr{B}$ is a split, if and only if

$$\sum_{i=1}^{k} L_{\mathscr{B}}(\mathscr{V}_i) = \dim(\langle A|\mathscr{B}\rangle).$$

Proof. First note that since $L_{\mathscr{B}}(\mathscr{V}_i)$ is the cover dimension of $\mathscr{V}_i$,

$$\sum_{i=1}^{k} L_{\mathscr{B}}(\mathscr{V}_i) \geqq \dim(\langle A|\mathscr{B}\rangle)$$

for any direct sum decomposition of $\mathscr{B}$. For each $i \in \underline{k}$, let $\mathscr{R}_i$ be a controllability subspace of dimension $L_{\mathscr{B}}(\mathscr{V}_i)$ such that $\mathscr{V}_i \subset \mathscr{R}_i$. Then surely (by Lemma 3.1) $\mathscr{R}_1 + \cdots + \mathscr{R}_k = \langle A|\mathscr{B}\rangle$, and hence if $\sum_{i=1}^{k} L_{\mathscr{B}}(\mathscr{V}_i) = \dim(\langle A|\mathscr{B}\rangle)$, the $\mathscr{R}_i$ must be linearly independent, and the decomposition of $\mathscr{B}$ is a split. Conversely, assume that the decomposition is a split. Let $\mathscr{R}_1 \oplus \cdots \oplus \mathscr{R}_k = \langle A|\mathscr{B}\rangle$ be a corresponding direct sum decomposition of $\langle A|\mathscr{B}\rangle$. Then by Theorem 4.3, $\dim(\mathscr{R}_i) \geqq L_{\mathscr{B}}(\mathscr{V}_i)$ for all $i \in \underline{k}$, and it follows that $\sum_{i=1}^{k} L_{\mathscr{B}}(\mathscr{V}_i) \leqq \sum_{i=1}^{k} \dim(\mathscr{R}_i) = \dim\langle A|\mathscr{B}\rangle$. Combining this inequality with the reversed inequality above, implies that equality must hold and the proof is complete. □

The following examples illustrate the fact that a direct sum decomposition of $\mathscr{B}$ may or may not be a split.

Example 5.3. Consider the system $(A, \mathscr{B})$, where $A: \mathbb{R}^5 \to \mathbb{R}^5$ and $\dim(\mathscr{B}) = 2$ as follows. Let $e_1, \cdots, e_5$ be a basis for $\mathbb{R}^5$ and define A by $Ae_1 = e_2$, $Ae_2 = e_3$,

$Ae_3 = e_4$, $Ae_4 = 0$, $Ae_5 = e_3$. Let $\mathscr{B} = \text{sp}\{b_1, b_2\}$, where $b_1 = e_1$ and $b_2 = e_5$. Then $\text{sp}\{b_1\} \oplus \text{sp}\{b_2\} = \mathscr{B}$ is a direct sum decomposition of $\mathscr{B}$, and it is easily seen that this decomposition is a split since $L_{\mathscr{B}}(\text{sp}\{b_1\}) = 3$, $L_{\mathscr{B}}(\text{sp}\{b_2\}) = 2$ and $\dim \langle A|\mathscr{B}\rangle = 5$.

Example 5.4. Consider the system $(A, \mathscr{B})$, where $A : \mathbb{R}^6 \to \mathbb{R}^6$ and $\dim(\mathscr{B}) = 2$ as follows. Let $e_1, \cdots, e_6$ be a basis for $\mathbb{R}^6$, and define A by $Ae_1 = e_2$, $Ae_2 = e_3$, $Ae_3 = e_4$, $Ae_4 = 0$, $Ae_5 = e_3$, $Ae_6 = e_5$. Let $\mathscr{B} = \text{sp}\{b_1, b_2\}$, where $b_1 = e_1$ and $b_2 = e_6$. In this case, $L_{\mathscr{B}}(\text{sp}\{b_1\}) = L_{\mathscr{B}}(\text{sp}\{b_2\}) = 4$, but $\dim \langle A|\mathscr{B}\rangle = 6$. Surely the decomposition $\text{sp}\{b_1\} \oplus \text{sp}\{b_2\} = \mathscr{B}$ is *not* a split.

In the next theorem, we give a necessary and sufficient condition for a decomposition of $\langle A|\mathscr{B}\rangle$ into a direct sum of controllability subspaces to be a split.

THEOREM 5.5. *Let $(A, \mathscr{B})$ be given and let $\mathscr{R}_1 \oplus \cdots \oplus \mathscr{R}_k = \langle A|\mathscr{B}\rangle$ be a direct sum decomposition of $\langle A|\mathscr{B}\rangle$ into controllability subspaces $\mathscr{R}_1, \cdots, \mathscr{R}_k$. Then the decomposition is a split if and only if $\dim(\mathscr{R}_i) = L_{\mathscr{B}}(\mathscr{R}_i \cap \mathscr{B})$ for all $i \in \underline{k}$.*

Proof. Assume first that the decomposition is a split. Then

$$(\mathscr{R}_1 \cap \mathscr{B}) \oplus \cdots \oplus (\mathscr{R}_k \cap \mathscr{B}) = \mathscr{B}, \tag{5.1}$$

and hence

$$\dim(\langle A|\mathscr{B}\rangle) \leqq \sum_{i=1}^{k} L_{\mathscr{B}}(\mathscr{R}_i \cap \mathscr{B}). \tag{5.2}$$

By Theorem 4.3, however, we have

$$\dim(\mathscr{R}_i) \geqq L_{\mathscr{B}}(\mathscr{R}_i \cap \mathscr{B}), \qquad i \in \underline{k} \tag{5.3}$$

and hence

$$\dim(\langle A|\mathscr{B}\rangle) = \sum_{i=1}^{k} \dim(\mathscr{R}_i) \geqq \sum_{i=1}^{k} L_{\mathscr{B}}(\mathscr{R}_i \cap \mathscr{B}). \tag{5.4}$$

Combining (5.2) with (5.3) and (5.4) we obtain that $\dim(\mathscr{R}_i) = L_{\mathscr{B}}(\mathscr{R}_i \cap \mathscr{B})$ for all $i \in \underline{k}$. Conversely, assume that $\mathscr{V} \triangleq (\mathscr{R}_1 \cap \mathscr{B}) \oplus \cdots \oplus (\mathscr{R}_k \cap \mathscr{B}) \neq \mathscr{B}$. Then by Lemma 4.2, $L_{\mathscr{B}}(\mathscr{V}) < L_{\mathscr{B}}(\mathscr{B}) = \dim(\langle A|\mathscr{B}\rangle) = \sum_{i=1}^{k} \dim(\mathscr{R}_i)$. But then by Theorem 5.2 at least for some of the $\mathscr{R}_i$, $\dim(\mathscr{R}_i) > L_{\mathscr{B}}(\mathscr{R}_i \cap \mathscr{B})$. □

An interesting question is to what extent do splitting decompositions of $\mathscr{B}$ actually exist. The key to the answer is provided by the following.

THEOREM 5.6. *Let $(A, \mathscr{B})$ be given, and let $\mathscr{V} \subset \mathscr{B}$ be any subspace. Then there exists a subspace $\mathscr{W} \subset \mathscr{B}$ such that the decomposition $\mathscr{B} = \mathscr{V} \oplus \mathscr{W}$ is a split.*

Proof. We need to show that there exists $\mathscr{W}$ such that $\mathscr{B} = \mathscr{V} \oplus \mathscr{W}$ and $L_{\mathscr{B}}(\mathscr{V}) + L_{\mathscr{B}}(\mathscr{W}) = \dim(\langle A|\mathscr{B}\rangle)$. For each $i > 0$, denote $\mathscr{V}_i \triangleq \mathscr{V}_i(\mathscr{B})$ and $t \triangleq s_{\mathscr{B}}(\mathscr{V})$, and let $[\mathscr{B}_i]_s$ and $[\mathscr{V}_i]_t$ $(t \leqq s)$ be, respectively, the input chain of $(A, \mathscr{B})$ and the $(A, \mathscr{B})$-induced chain of $\mathscr{V}$. If $\mathscr{W} \subset \mathscr{B}$ satisfies $\mathscr{W} \cap \mathscr{V} = 0$ and $\mathscr{W}_i \oplus \mathscr{V}_i = \mathscr{B}_i$ for all $i \in \underline{s}$, where $\mathscr{W}_i \triangleq \mathscr{W}_i(\mathscr{B})$, (i.e., $[\mathscr{W}_i]_r$ $(r \leqq s)$ is the

$(A, \mathscr{B})$-induced chain of $\mathscr{W}$), then

$$\begin{aligned}\dim(\langle A|\mathscr{B}\rangle) - L_{\mathscr{B}}(\mathscr{V}) &= s\cdot\dim(\mathscr{B}) - \sum_{i=1}^{s-1}\dim(\mathscr{B}_i) - \left[t\cdot\dim(\mathscr{V}) - \sum_{i=1}^{t-1}\dim(\mathscr{V}_i)\right] \\ &= s[\dim(\mathscr{B}) - \dim(\mathscr{V})] - \sum_{i=1}^{s-1}[\dim(\mathscr{B}_i) - \dim(\mathscr{V}_i)] \\ &= s\cdot\dim(\mathscr{W}) - \sum_{i=1}^{s-1}\dim(\mathscr{W}_i) \\ &= r\cdot\dim(\mathscr{W}) - \sum_{i=1}^{r-1}\dim(\mathscr{W}_i) = L_{\mathscr{B}}(\mathscr{W}).\end{aligned}$$

Hence the proof will be complete if we demonstrate the existence of $\mathscr{W} \subset \mathscr{B}$ such that

(i) $$\mathscr{W} \cap \mathscr{V} = 0,$$

(ii) $$(\mathscr{W} \cap \mathscr{B}_i) \oplus (\mathscr{V} \cap \mathscr{B}_i) = \mathscr{B}_i, \qquad i \in \underline{s}.$$

For $i = 1, 2, \cdots$, define $\mathscr{V}^i \subset \mathscr{V}$ such that $\mathscr{V}_i \oplus \mathscr{V}^i = \mathscr{V}_{i+1}$. Clearly, $\mathscr{B}_i \cap \mathscr{V}^i = 0$ for all i, and for $i = 1, 2, \cdots$, define $\mathscr{W}^i$ such that $\mathscr{W}^i \oplus \mathscr{V}^i \oplus \mathscr{B}_i = \mathscr{B}_{i+1}$. Choose $\mathscr{W}_1$ such that $\mathscr{W}_1 \oplus \mathscr{V}_1 = \mathscr{B}_1$ and for each $i = 1, 2, \cdots$, let $\mathscr{W}_{i+1} = \mathscr{W}_i \oplus \mathscr{W}^i$. Setting $\mathscr{W} = \lim\{\mathscr{W}_i\}$, it is immediately seen that $\mathscr{W} \cap \mathscr{V} = 0$, and hence (i) holds. Also, from the definition of the $\mathscr{W}_i$, it is clear that $\mathscr{W}_i + \mathscr{V}_i = \mathscr{B}_i$ for all $i = 1, 2, \cdots$. That also $\mathscr{W}_i \cap \mathscr{V}_i = 0$ is seen by induction on i as follows. It surely holds for $i = 1$ by construction. Suppose it holds for all $i \leqq k-1$ and let $v \in \mathscr{W}_k \cap \mathscr{V}_k$ be any vector. Then $v \in [\mathscr{W}_{k-1} \oplus \mathscr{W}^{k-1}] \cap [\mathscr{V}_{k-1} \oplus \mathscr{V}^{k-1}]$ and hence $v = w_{k-1} + w^{k-1} = v_{k-1} + v^{k-1}$ for vectors in the respective subspaces. Consequently, $w_{k-1} - v_{k-1} = v^{k-1} - w^{k-1}$ and $w_{k-1} - v_{k-1} \in \mathscr{B}_{k-1}$ and $v^{k-1} - w^{k-1} \in \mathscr{V}^{k-1} \oplus \mathscr{W}^{k-1}$. Since $\mathscr{B}_{k-1} \cap [\mathscr{V}^{k-1} \oplus \mathscr{W}^{k-1}] = 0$, it follows that $w_{k-1} = v_{k-1}$ and $v^{k-1} = w^{k-1}$. Since the assertion is assumed to hold for $k-1$, we conclude that $w_{k-1} = v_{k-1} = 0$, and since $\mathscr{W}^{k-1} \cap \mathscr{V}^{k-1} = 0$, we also have $w^{k-1} = v^{k-1} = 0$. Hence $v = 0$ and the assertion also holds for $i = k$. To complete the proof we only need to observe that $\mathscr{W} \cap \mathscr{B}_i = \mathscr{W}_i$ for all i. Indeed, $\mathscr{W} \cap \mathscr{B}_i = \mathscr{W} \cap [\mathscr{V}_i \oplus \mathscr{W}_i] = \mathscr{W} \cap \mathscr{V}_i \oplus \mathscr{W}_i = \mathscr{W}_i$, since $\mathscr{W}_i \subset \mathscr{W}$ and since $\mathscr{W} \cap \mathscr{V}_i \subset \mathscr{W} \cap \mathscr{V} = 0$. □

Consider now a system $(A, \mathscr{B})$ and let $m = \dim(\mathscr{B})$. Let $\mathscr{V}_1 \subset \mathscr{B}$ be any one-dimensional subspace and denote $\sigma_1 = L_{\mathscr{B}}(\mathscr{V}_1)$. By Theorem 5.6, there exists a subspace $\mathscr{V}^1 \subset \mathscr{B}$ such that the direct sum decomposition $\mathscr{V}_1 \oplus \mathscr{V}^1 = \mathscr{B}$ is a split, and hence $\sigma_1 + L_{\mathscr{B}}(\mathscr{V}^1) = \dim(\langle A|\mathscr{B}\rangle)$. Let $\hat{A}_1 \in \mathscr{M}(\mathscr{B})$ be such that $\mathscr{R}_1 = \langle A + \hat{A}_1|\mathscr{V}_1\rangle$ and $\mathscr{R}^1 = \langle A + \hat{A}_1|\mathscr{V}^1\rangle$ satisfy $\mathscr{R}_1 \oplus \mathscr{R}^1 = \langle A|\mathscr{B}\rangle$. Next, consider the system $(A + \hat{A}_1, \mathscr{V}^1)$. By repeating the above procedure for $(A + \hat{A}, \mathscr{V}^1)$ (in the quotient space $\langle A|\mathscr{B}\rangle \bmod (\mathscr{R}_1)$) we find a one-dimensional subspace $\mathscr{V}_2 \subset \mathscr{V}^1$ and a subspace $\mathscr{V}^2 \subset \mathscr{V}^1$ such that $\mathscr{V}_2 \oplus \mathscr{V}^2 = \mathscr{V}^1$ and $\sigma_2 + L_{\mathscr{B}}(\mathscr{V}^2) = L_{\mathscr{B}}(\mathscr{V}^1)$, where $\sigma_2 = L_{\mathscr{B}}(\mathscr{V}_2)$. Hence there exists a map $\hat{A}_2 \in \mathscr{M}(\mathscr{B})$ such that $\mathscr{R}_i = \langle A + \hat{A}_2|\mathscr{V}_i\rangle$, $i = 1, 2$, and $\mathscr{R}^2 = \langle A + \hat{A}_2|\mathscr{V}^2\rangle$ satisfy $\mathscr{R}_1 \oplus \mathscr{R}_2 \oplus \mathscr{R}^2 = \langle A|\mathscr{B}\rangle$. The above decomposition can be applied repeatedly until finally we clearly obtain a splitting direct sum decomposition $\mathscr{B} = \mathscr{V}_1 \oplus$

$\cdots \oplus \mathscr{V}_m$ of $\mathscr{B}$ into one-dimensional subspace $\mathscr{V}_i$. Correspondingly, we also have a splitting direct sum decomposition $\langle A|\mathscr{B}\rangle = \mathscr{R}_1 \oplus \cdots \oplus \mathscr{R}_m$, where $R_i \cap \mathscr{B} = \mathscr{V}_i$ and $\dim(\hat{\mathscr{R}}_i) = L_{\mathscr{B}}(\mathscr{V}_i) = \sigma_i$ for each controllability subspace $\mathscr{R}_i$, $i \in \underline{m}$.

It is worth observing that the splitting decomposition of $\mathscr{B}$ that was demonstrated above is by no means unique, and that the splitting decomposition of $\langle A|\mathscr{B}\rangle$ is not even unique relative to a *given* corresponding splitting decomposition of $\mathscr{B}$. We will see however that, regardless of the specific decomposition, the list $\{\sigma_1, \cdots, \sigma_m\}$ of dimensions is unique up to ordering. Without loss of generality, let us assume that $\sigma_i \geqq \sigma_{i+1}$ for all $i \in \underline{m-1}$ (otherwise we could simply relabel the subspaces). For any integer $k > 0$, define the set $I(k)$ as $I(k) = \{i \in \underline{m} | \sigma_i \leqq k\}$. In view of the definition of the input chain $[\mathscr{B}_i]$ and the controllability chain $[\mathscr{S}_i]$, it is easily verified that $\sum_{i \in I(k)} \mathscr{V}_i = \mathscr{B}_k$ and $\sum_{i \in I(k)} \mathscr{R}_i = \mathscr{S}_k$. Hence if $\{\gamma_i\}$ is the input list of $(A, \mathscr{B})$, then for each integer $k > 0$, there exist exactly $(\gamma_k - \gamma_{k-1})$ elements in the list $\{\sigma_1, \cdots, \sigma_m\}$ whose value is k (here as before we define $\gamma_0 = 0$). From the uniqueness of the input list, it is now a direct consequence that the (ordered) list $\{\sigma_1, \cdots, \sigma_m\}$ is also unique and independent of the specific splitting decomposition. However, as was noted, a splitting decomposition cannot be quite arbitrary, since for each $k > 0$, $\sum_{i \in I(k)} \mathscr{V}_i = \mathscr{B}_k$ and $\sum_{i \in I(k)} \mathscr{R}_i = \mathscr{S}_k$, and the subspaces $\mathscr{B}_k$ and $\mathscr{S}_k$ are unique. We summarize the above observations in the following.

THEOREM 5.7. *Let $(A, \mathscr{B})$ be given and let $m = \dim(\mathscr{B})$. Then there exists a splitting direct sum decomposition $\mathscr{B} = \mathscr{V}_1 \oplus \cdots \oplus \mathscr{V}_m$ of $\mathscr{B}$ into one-dimensional subspaces $\mathscr{V}_i$. Correspondingly, there also exists a splitting direct sum decomposition $\langle A|\mathscr{B}\rangle = \mathscr{R}_1 \oplus \cdots \oplus \mathscr{R}_m$, where $\mathscr{R}_i \cap \mathscr{B} = \mathscr{V}_i$ for each controllability subspace $\mathscr{R}_i$, $i \in \underline{m}$. Moreover, while these decompositions are not unique, the following always hold:*

(i) *the list $\{\sigma_1, \cdots, \sigma_m\}$, where for each i, $\sigma_i = L_{\mathscr{B}}(\mathscr{V}_i)$, is unique up to ordering of elements;*

(ii) *for each integer $k > 0$, the list $\{\sigma_1, \cdots, \sigma_m\}$ contains exactly $(\gamma_k - \gamma_{k-1})$ elements whose value is k, where $\{\gamma_i\}$ is the input list of $(A, \mathscr{B})$;*

(iii) *for each integer $k > 0$,*

$$\sum_{i \in I(k)} \mathscr{V}_i = \mathscr{B}_k,$$

$$\sum_{i \in I(k)} \mathscr{R}_i = \mathscr{S}_k,$$

where $I(k) \triangleq \{i \in \underline{m} | \sigma_i \leqq k\}$.

Remark 5.8. Theorem 5.7 is essentially a generalization of several previous results by various authors. The list of integers $\{\sigma_1, \cdots, \sigma_m\}$ is precisely the list of controllability indices whose feedback invariance properties were already proved by Brunovsky in [1]. The fact that there exists a splitting direct sum decomposition of $\mathscr{B}$ whose list of cover dimensions are the controllability indices, was shown by Wonham and Morse in [2]. The uniqueness of the subspaces $\sum_{i \in I(k)} \mathscr{R}_i$ was also proved by Warren and Eckberg in [9] using different considerations. They also relate these subspaces to the fundamental chain of $\langle A|\mathscr{B}\rangle$ (although in a different way, see [9, Prop. 3]). The connection between the corresponding splitting decompositions of $\langle A|\mathscr{B}\rangle$ and of $\mathscr{B}$ is new.

We now introduce the following.

DEFINITION 5.9. A system $(A, \mathscr{B})$ is said to be in *normal form*, if and only if for each vector $v \in \mathscr{B}$, $v \in \mathscr{B}_i$ if and only if $A^i v = 0$.

Upon combining Theorem 3.4 and Theorem 5.7 we immediately have the following.

THEOREM 5.10. *For every system* $(A, \mathscr{B})$ *there exists* $\hat{A} \in \mathscr{M}(\mathscr{B})$ *such that* $(A + \hat{A}, \mathscr{B})$ *is in normal form.*

The normal form is clearly a "basis independent" version of the so-called "Brunovsky canonical form" in which

$$A = \operatorname{diag}(A_1, \cdots, A_m), \qquad B = \operatorname{diag}(B_1, \cdots, B_m),$$

where for each i,

$$A_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \qquad B_i = \begin{bmatrix} 0 \\ \cdot \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Indeed, it is easy to see that if a system $(A, \mathscr{B})$ is in normal form and $\langle A|\mathscr{B}\rangle = \mathscr{X}$ (where $\mathscr{X}$ is the state space), i.e., $(A, \mathscr{B})$ is controllable, then one can find bases for $\mathscr{X}$ and $\mathscr{U}$ ($\mathscr{U}$ is the input space) such that the matrices (A, B) are in the Brunovsky canonical form. Theorem 5.10 then simply says that given a controllable pair (of matrices) (A, B) there exist matrices (T, F, G), where T and G are nonsingular, such that the pair $(T(A + BF)T^{-1}, TBG)$ is in Brunovsky canonical form, a fact which is well known. However, it is *not* the direct sum decomposition of the system, but rather the partial (chain) decomposition which exhibits the system's essential structural features. This important fact has so far been overlooked.

6. Proper lists. Let $(A, \mathscr{B})$ be a given linear system, and denote by $G\{(A, \mathscr{B})\}$ the class of all systems of the form $(A + \hat{A}, \mathscr{V})$, where $\hat{A} \in \mathscr{M}(\mathscr{B})$ and $\mathscr{V} \subset \mathscr{B}$. Let $I\{(A, \mathscr{B})\}$ denote the collection of all input lists for systems in $G\{(A, \mathscr{B})\}$, and by $J\{(A, \mathscr{B})\}$ the collection of all lists of controllability indices for systems in $G\{(A, \mathscr{B})\}$. In the present section, we shall characterize these classes and will use this characterization in the next section to solve the feedback simulation problem. We begin with the following.

DEFINITION 6.1. Let $(A, \mathscr{B})$ be a given linear system with (infinite) input list $\{\gamma_i\}$. Then a nondecreasing sequence of positive integers $\{\beta_i\}$ is called *proper* with respect to $\{\gamma_i\}$, if and only if the following relations hold:

(i) $\beta_i \leqq \gamma_i$ for all $i \geqq 1$,

(ii) $\sum_{i=1}^{k} i(\beta_i - \beta_{i-1}) \leqq \sum_{i=1}^{k} i(\gamma_i - \gamma_{i-1})$ for all $k \geqq 1$, $\beta_0 = \gamma_0 = 0$.

DEFINITION 6.2. Let $(A, \mathscr{B})$ be a given linear system with list of controllability indices $\{\sigma_1, \cdots, \sigma_m\}$, $\sigma_i \geqq \sigma_{i+1}$ for $i \in \underline{m-1}$. Then a nonincreasing list of positive

integers $\{\xi_1, \cdots, \xi_l\}$ is called *proper* with respect to $\{\sigma_1, \cdots, \sigma_m\}$, if and only if the following hold:

(i) $l \leqq m$,

(ii) $\sum_{i=1}^{l} \xi_i \leqq \sum_{i=1}^{m} \sigma_i$.

(iii) If $\xi_r < \sigma_s$ for some $r \in \underline{l}$ and some $s \in \underline{m-1}$, then

$$\sum_{i=r}^{l} \xi_i \leqq \sum_{i=s+1}^{m} \sigma_i,$$

(iv) $\xi_r \geqq \sigma_{r+m-l}$ for all $r \in \underline{l}$.

The following proposition relates Definitions 6.1 and 6.2.

PROPOSITION 6.3. *Let $(A, \mathscr{B})$ and $(A', \mathscr{B}')$ be two linear systems with input lists $\{\gamma_i\}$ and $\{\beta_i\}$, respectively, and with lists of controllability indices $\{\sigma_1, \cdots, \sigma_m\}$ and $\{\xi_1, \cdots, \xi_l\}$, respectively. Then the list $\{\beta_i\}$ is proper with respect to $\{\gamma_i\}$; if and only if the list $\{\xi_1, \cdots, \xi_l\}$ is proper with respect to $\{\sigma_1, \cdots, \sigma_m\}$.*

Proof. By direct computation (see also Theorem 5.7 (ii)). □

In the remainder of this section we will prove that for a given linear system $(A, \mathscr{B})$, a list is in $I\{(A, \mathscr{B})\}$ (resp. in $J\{(A, \mathscr{B})\}$), if and only if it is proper with respect to the corresponding list of $(A, \mathscr{B})$.

One important property of proper lists which is easily proved and used in the sequel, is stated in the following.

LEMMA 6.4. *Let $(A, \mathscr{B})$ be a linear system with input list $\{\gamma_i\}$. Let $\{\beta_i\}$ be a proper list with respect to $\{\gamma_i\}$, and assume that $\lim \beta_i = \lim \gamma_i$. Then $\beta_i = \gamma_i$ for all $i = 1, 2, \cdots$.*

Remark 6.5. An immediate consequence of Lemma 6.4 and Proposition 6.3 which is also easily verified using Definition 6.2 directly, is that if $\{\sigma_1, \cdots, \sigma_m\}$ is a list of controllability indices and $(\xi_1, \cdots, \xi_l\}$ is a proper list with respect to $\{\sigma_1, \cdots, \sigma_m\}$, then $l = m$ implies $\sigma_i = \xi_i$ for all i.

THEOREM 6.6. *Let $(A, \mathscr{B})$ be a given system with input list $\{\gamma_i\}$ and list of controllability indices $\{\sigma_1, \cdots, \sigma_m\}$. Let $(A + \hat{A}, \mathscr{V})$ be any system in $G\{(A, \mathscr{B})\}$ with input list $\{\beta_i\}$ and list of controllability indices $\{\xi_1, \cdots, \xi_l\}$. Then the lists of $(A + \hat{A}, \mathscr{V})$ are proper with respect to the corresponding lists of $(A, \mathscr{B})$.*

Proof. We will prove the theorem only for the input lists. The result for the lists of controllability indices then follows from Proposition 6.3. If $\{\mathscr{B}_i\}$ and $\{\mathscr{V}_i\}$ are, respectively, the input chains of $(A, \mathscr{B})$ and of $(A + \hat{A}, \mathscr{V})$, then (as was seen previously) $\mathscr{V}_i \subset \mathscr{B}_i$ for all $i \geqq 1$, and hence $\dim(\mathscr{V}_i) \leqq \dim(\mathscr{B}_i)$ and (i) of Definition 6.1 holds. That (ii) also holds follows from the fact that if $\mathscr{R}$ is any controllability subspace of $(A + \hat{A}, \mathscr{V})$ which is contained in $\langle A + \hat{A} | \mathscr{V} \rangle_k$, then it is also a controllability subspace of $(A, \mathscr{B})$ and is contained in $\langle A | \mathscr{B} \rangle_k$. The terms in the inequality (ii) of Definition 6.1 are simply upper bounds on the dimensions of controllability subspaces in $\langle A + \hat{A} | \mathscr{V} \rangle_k$ and $\langle A | \mathscr{B} \rangle_k$. □

The proof of Theorem 6.6 relies on the fact that the chains of every system in $G\{(A, \mathscr{B})\}$ are subchains of the corresponding chains of $(A, \mathscr{B})$. The properness conditions for the lists are thus nothing more than the dimensionality constraints

thereby imposed. We will prove below that given a system $(A, \mathscr{B})$, every proper list is indeed realized by some element in $G\{(A, \mathscr{B})\}$. Since we will use a construction in our proof, it will turn out to be more convenient to use Definition 6.2 for proper lists of controllability indices. By Proposition 6.3, the results then hold also for proper input lists.

LEMMA 6.7. *Let $(A, \mathscr{B})$ be a given system with list of controllability indices $\{\sigma_1, \cdots, \sigma_m\}$, $(\sigma_i \geqq \sigma_{i+1})$. Let $\{\xi_1, \cdots, \xi_{m-1}\}$ be a nonincreasing list of positive integers such that $\xi_i \geqq \sigma_i$, $i = 1, \cdots, m-1$, and $\xi_i > \sigma_i$ for some $i \in \underline{m-1}$. Then there exists $\mathscr{V} \subset \mathscr{B}$ with $\dim(\mathscr{V}) = m - 1$ and an $\hat{A} \in \mathscr{M}(\mathscr{B})$ such that the list $\{s_1, \cdots, s_{m-1}\}$, $(s_i \geqq s_{i+1})$, of controllability indices of $(A + \hat{A}, \mathscr{V})$ satisfies*

(i) $$\xi_i \geqq s_i \geqq \sigma_1 \quad \text{for all } i \in \underline{m-1},$$

(ii) $$\sum_{i=1}^{m-1} s_i = \min\left\{\sum_{i=1}^{m-1} \xi_i, \sum_{i=1}^{m} \sigma_i\right\}.$$

Proof. We shall prove the Lemma by construction. For each $i \in \underline{m-1}$, define $q_i = \xi_i - \sigma_i$. Let $k_1 = \min(q_1, \sigma_m)$, and for each $i = 2, 3, \cdots, m-1$, set $k_i = \min(q_i, \sigma_m - \sum_{j=1}^{i-1} k_j)$. Since we assumed that $q_i > 0$ for some $i \in \underline{m-1}$, it is clear that $k_i > 0$ for some $i \in \underline{m-1}$. Let t be the maximal integer for which $k_t > 0$, and set $\theta_j = \sigma_j - \sigma_m + \sum_{i=j}^{t} k_i$ for all $j \in \underline{t}$. Assume $(A, \mathscr{B})$ is in normal form and let $v_1, \cdots, v_m$ be a basis for $\mathscr{B}$ such that $L_{\mathscr{B}}(\mathrm{sp}\{v_i\}) = \sigma_i$ for all $i \in \underline{m}$. Let $\hat{A} \in \mathscr{M}(\mathscr{B})$ be defined by

$$\hat{A}(A^j v_i) = \begin{cases} v_m & \text{for } j = \theta_i, \quad i = 1, \cdots, t, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathscr{V}_t = \mathrm{sp}\{v_1, \cdots, v_t\}$ and $\mathscr{V}^t = \mathrm{sp}\{v_{t+1}, \cdots, v_{m-1}\}$ and define $\mathscr{V}$ by $\mathscr{V} = \mathscr{V}_t \oplus \mathscr{V}^t$. It is then readily noted that $\langle A + \hat{A} | \mathscr{V} \rangle = \langle A + \hat{A} | \mathscr{V}_t \rangle \oplus \langle A + \hat{A} | \mathscr{V}^t \rangle$ and the list of controllability indices of $(A + \hat{A}, \mathscr{V})$ is the union of the lists of controllability indices of $(A + \hat{A}, \mathscr{V}_t)$ and of $(A + \hat{A}, \mathscr{V}^t)$. Furthermore, the list of controllability indices of $(A + \hat{A}, \mathscr{V}^t)$ is the same as that of $(A, \mathscr{V}^t)$, i.e., $\{\sigma_{t+1}, \cdots, \sigma_{m-1}\}$. Hence we confine our attention to $(A + \hat{A}, \mathscr{V}_t)$. We shall distinguish between two cases:

(i) Assume $\sum_{i=1}^{m} \sigma_i \leqq \sum_{i=1}^{m-1} \xi_i$. Then t is the least integer such that $\sum_{i=1}^{t} (\xi_i - \sigma_i) \geqq \sigma_m$, and hence $\sum_{i=1}^{r} (\xi_i - \sigma_i) < \sigma_m$ for all $r \in \underline{t-1}$, whence $\xi_t - \sigma_t > 0$. We then have $\sigma_m - \sum_{i=1}^{r-1} (\xi_i - \sigma_i) > \xi_r - \sigma_r$ for all $r \in \underline{t-1}$ and consequently,

$$k_i = \xi_i - \sigma_i \quad \text{for all } i \in \underline{t-1},$$

$$k_t = \sigma_m - \sum_{i=1}^{t-1} k_i.$$

The reader can verify directly that in this case the list $\{s_1, \cdots, s_t\}$ of controllability indices of $(A + \hat{A} | \mathscr{V}_t)$ is given by

$$s_i = \xi_i \qquad \text{for } i \in \underline{t-1},$$

$$s_t = \sigma_t + k_t,$$

and

$$s_t = \sigma_t + \sigma_m - \sum_{i=1}^{t-1} (\xi_i - \sigma_i) = \sigma_m + \xi_t - \sum_{i=1}^{t} (\xi_i - \sigma_i) \leqq \xi_t.$$

Hence we immediately conclude that $s_i \geqq s_{i+1}$ for all $i \in \underline{t-1}$ and, moreover, $\sigma_i \leqq s_i \leqq \xi_i$ for all $i \in \underline{t}$. Also,

$$\sum_{i=1}^{t} s_i = \sum_{i=1}^{t-1} \xi_i + \sigma_m + \xi_t - \sum_{i=1}^{t} (\xi_i - \sigma_i) = \sum_{i=1}^{t} \sigma_i + \sigma_m$$

and hence,

$$\sum_{i=1}^{m-1} s_i = \sum_{i=1}^{m} \sigma_i.$$

(ii) Assume $\sum_{i=1}^{m} \sigma_i > \sum_{i=1}^{m-1} \xi_i$. Clearly, $\sigma_m > \sum_{i=1}^{k} (\xi_i - \sigma_i)$ for all $k \in \underline{m-1}$, so that $k_i = \xi_i - \sigma_i$ for all $i \in \underline{m-1}$ and t is then the maximal integer for which $\xi_i > \sigma_i$.

Again, the reader can verify that $s_i = \xi_i$ for all $i \in \underline{t}$ and, consequently, $s_i = \xi_i$ for all $i \in \underline{m-1}$. We then have

$$\sum_{i=1}^{m-1} s_i = \sum_{i=1}^{m-1} \xi_i,$$

and the proof is complete. □

We can now prove the following.

THEOREM 6.8. *Let $(A, \mathscr{B})$ be a given system with list of controllability indices $\{\sigma_1, \cdots, \sigma_m\}$, $(\sigma_i \geqq \sigma_{i+1})$. Let $\{\xi_1, \cdots, \xi_l\}$ be a proper list with respect to $\{\sigma_1, \cdots, \sigma_m\}$. Then there exist $\mathscr{V} \subset \mathscr{B}$, $\dim(\mathscr{V}) = l$, and a map $\hat{A} \in \mathscr{M}(\mathscr{B})$ such that the list of controllability indices of $(A + \hat{A}, \mathscr{V})$ is $\{\xi_1, \cdots, \xi_l\}$.*

Proof. First note that if $l = m$, there is nothing to prove in view of Lemma 6.4 (see also Remark 6.5). Hence assume $l < m$ and let p be the least integer such that $\xi_1 \geqq \sigma_p$. It is then easily verified that the list $\{\xi_1, \cdots, \xi_l\}$ is also proper with respect to the list $\{\sigma_p, \cdots, \sigma_m\}$. Hence, since we can then split the system into a direct sum of two sub-systems, one of which has controllability indices $\{\sigma_p, \cdots, \sigma_m\}$, we may assume at the outset, without loss of generality, that $p = 1$ and $\xi_1 \geqq \sigma_1$. Let $(A, \mathscr{B})$ be in normal form, and let $v_1, \cdots, v_m$ be a basis for $\mathscr{B}$ such that for each $i \in \underline{m}$, $L_{\mathscr{B}}(\text{sp}\{v_i\}) = \sigma_i$. Suppose that for some $k > 1$, $\xi_k < \sigma_k$. It then follows that $\sum_{i=k}^{l} \xi_i \leqq \sum_{i=k+1}^{m} \sigma_i$ and hence the list $\{\xi_k, \cdots, \xi_l\}$ is proper with respect to $\{\sigma_{k+1}, \cdots, \sigma_m\}$. Write $\langle A|\mathscr{B}\rangle = \langle A|\mathscr{V}_1\rangle \oplus \langle A|\mathscr{V}_2\rangle$, where $\mathscr{V}_1 = \text{sp}\{v_1, \cdots, v_k\}$ and $\mathscr{V}_2 = \text{sp}\{v_{k+1}, \cdots, v_m\}$. Apply Lemma 6.7 to the system $(A, \mathscr{V}_1)$ with respect to the list $\{\xi_1, \cdots, \xi_{k-1}\}$. We then obtain a system $(A + \hat{A}, \bar{\mathscr{V}}_1)$ (with $\dim(\bar{\mathscr{V}}_1) = k - 1$) with list of controllability indices $\{s_1, \cdots, s_{k-1}\}$ such that (i) and (ii) of Lemma 6.7 hold. The system $(A + \hat{A}, \bar{\mathscr{V}}_1 \oplus \mathscr{V})$ then has list of controllability indices $\{s_1, \cdots, s_{k-1}, \sigma_{k+1}, \cdots, \sigma_m\}$, and it is readily verified that the list $\{\xi_1, \cdots, \xi_l\}$ is also proper with respect to it. However, this new list of controllability indices has only $m - 1$ elements. If $l = m - 1$, we are done by Lemma 6.4. If $l < m - 1$, the above reduction procedure can be repeated until either the lengths of the lists coincide or we have a list $\{\bar{\sigma}_1, \cdots, \bar{\sigma}_t\}$ $(t > l)$ with respect to which the list $\{\xi_1, \cdots, \xi_l\}$ is proper and $\xi_i \geqq \bar{\sigma}_i$ for all $i \in \underline{l}$. We then apply Lemma 6.7 to the lists $\{\bar{\sigma}_1, \cdots, \bar{\sigma}_{l+1}\}$ and $\{\xi_1, \cdots, \xi_l\}$, and by repeating this procedure,

the lists eventually coincide in length, and by Lemma 6.4, the lists coincide in value. This completes the proof. □

For the sake of completeness we summarize Theorems 6.6 and 6.8 in terms of Definition 6.1 in the following.

THEOREM 6.9. *Let $(A, \mathscr{B})$ be a given system with input list $\{\gamma_i\}$. Then a list $\{\beta_i\}$ is in $I\{(A, \mathscr{B})\}$, if and only if it is proper with respect to $\{\gamma_i\}$.*

We conclude this section with the following.

THEOREM 6.10. *Let $(A, \mathscr{B})$ be a given system with input list $\{\gamma_i\}$. Let $\{\beta_i\}$ be any proper list with respect to $\{\gamma_i\}$. Then there exists another proper list $\{\delta_i\}$ such that the following hold:*

(i) $\delta_i \leqq \beta_i$ *for all* $i \geqq 1$,

(ii) $\lim \delta_i = \lim \beta_i$,

(iii) $$\sum_{i=1}^{\infty} i(\delta_i - \delta_{i-1}) = \sum_{i=1}^{\infty} i(\gamma_i - \gamma_{i-1}), \qquad \delta_0 = \gamma_0 = 0.$$

Proof. Denote $p = \sum_{i=1}^{\infty} i(\gamma_i - \gamma_{i-1}) - \sum_{i=1}^{\infty} i(\beta_i - \beta_{i-1})$. Let t be the maximal integer such that $\beta_t > \beta_{t-1}$ and write $\delta_i = \beta_i$ for $i = 1, \cdots, t-1$, $\delta_i = \beta_t - 1$ for $i = t, \cdots, t + p - 1$, and $\delta_i = \beta_t$ for $i \geqq t + p$. An easy calculation verifies that (i)–(iii) are satisfied. □

7. Feedback simulation. In this section we shall investigate the feedback simulation problem as was defined in the Introduction. In all our discussions of feedback simulation, it will be assumed, without further mention, that the systems under consideration are controllable. If $\Sigma = (A, \mathscr{B})$ and $\Sigma' = (A', \mathscr{B}')$ are two systems defined over spaces $\mathscr{X}$ and $\mathscr{X}'$, respectively, we shall frequently use the terminology "T is a linear map of Σ into Σ'" in reference to a map $T: \mathscr{X} \to \mathscr{X}'$.

In conformity with our discussion in the previous sections, we can rephrase the concept of feedback simulation as follows.

DEFINITION 7.1. Let $\Sigma = (A, \mathscr{B})$ be a given linear system. A system $\Sigma' = (A', \mathscr{B}')$ *is in the simulation orbit* of Σ (denoted $\Sigma' \in O\{\Sigma\}$), if and only if there exist a map $\hat{A} \in \mathscr{M}(\mathscr{B})$ and a linear epimorphism $T: \Sigma \to \Sigma'$ such that

$$T(A + \hat{A}) = A'T, \tag{7.1}$$

$$\mathscr{B}' \subset T\mathscr{B}. \tag{7.2}$$

The following is immediate.

LEMMA 7.2. *Let $\Sigma = (A, \mathscr{B})$ be given and let $\Sigma' = (A', \mathscr{B}')$ be any other system. Then $\Sigma' \in O\{\Sigma\}$ if and only if $\hat{\Sigma}' \in O\{\hat{\Sigma}\}$ for every $\hat{\Sigma}' \in E\{\Sigma'\}$ and every $\hat{\Sigma} \in E\{\Sigma\}$.*

We can now prove the following.

THEOREM 7.3. *Let $\Sigma = (A, \mathscr{B})$ and $\Sigma' = (A', \mathscr{B}')$ be given linear systems and let $\{\alpha_i\}$ be the (infinite) input list of $(A', \mathscr{B}')$. Then $\Sigma' \in O\{\Sigma\}$ if and only if there exists a map $\hat{A} \in \mathscr{M}(\mathscr{B})$ and a subspace $\mathscr{V} \subset \mathscr{B}$ with $\dim(\mathscr{V}) = \dim(\mathscr{B}')$, such that the (infinite) input list $\{\beta_i\}$ of $(A + \hat{A}, \mathscr{V})$ satisfies $\beta_i \leqq \alpha_i$ for all $i \geqq 1$.*

Remark 7.4. Theorem 7.3 has the following interpretation in terms of the controllability indices. The system Σ' is in the simulation orbit of Σ if and only if Σ has a controllability subspace $\mathscr{R} = \langle A + \hat{A} | \mathscr{V} \rangle$ such that if $(\sigma_1, \cdots, \sigma_{m'})$ is

the (ordered) list of controllability indices of $(A + \hat{A}, \mathscr{V})$, and $\{\sigma'_1, \cdots, \sigma'_{m'})$ is the (ordered) list of controllability indices of $(A', \mathscr{B}')$, then $\sigma'_i \leqq \sigma_i$ for all $i \in \underline{m}'$.

Proof of Theorem 7.3. Assume first that $\Sigma' \in O\{\Sigma\}$ and that the maps T and $\hat{A}$ exist such that (7.1) and (7.2) hold. Surely then, there exists a subspace $\mathscr{V} \subset \mathscr{B}$ such that $\dim(\mathscr{V}) = \dim(\mathscr{B}')$ and $\mathscr{B}' = T\mathscr{V}$. Let $\hat{\hat{A}} \in \mathscr{M}(\mathscr{V})$ be chosen such that the system $(A^*, \mathscr{V})$ is in normal form, where $A^* = A + \hat{A} + \hat{\hat{A}}$. (That $\hat{A}$ can be selected this way follows from Theorem 5.10.) We then have

$$\operatorname{Im}(TA^* - A'T) \subset T\mathscr{V} \tag{7.3}$$

and

$$\mathscr{B}' = T\mathscr{V}; \quad \dim(\mathscr{V}) = \dim(\mathscr{B}'). \tag{7.4}$$

It is readily verified that (7.3) implies that for any integer $j > 0$ and any vector $v \in \mathscr{V}$,

$$[T(A^*)^j - (A')^j T]v \in \langle A' | \mathscr{B}' \rangle_j. \tag{7.5}$$

Let $\{\mathscr{B}'_i\}$ and $\{\mathscr{V}_i\}^*$ be, respectively, the input chains of $(A', \mathscr{B}')$ and of $(A^*, \mathscr{V})$. Then, since $(A^*, \mathscr{V})$ is in normal form, $(A^*)^k v = 0$ for any $v \in \mathscr{V}_k$, and consequently by (7.5),

$$(A')^k Tv \in \langle A' | \mathscr{B}' \rangle_k$$

so that $Tv \in \mathscr{B}'_k$. We conclude that $T\mathscr{V}_k \subset \mathscr{B}'_k$ for all $k > 0$, and combining this fact with (7.4), we obtain $\beta_k = \dim(\mathscr{V}_k) \leqq \dim(\mathscr{B}'_k) = \alpha_k$ for all $k > 0$.

Conversely, assume there exist $\hat{A} \in \mathscr{M}(\mathscr{B})$ and $\mathscr{V} \subset \mathscr{B}$ such that $\dim(\mathscr{V}) = \dim(\mathscr{B}')$ and the input list $\{\beta_i\}$ of $(A^*, \mathscr{V})$ (where $A^* = A + \hat{A}$) satisfies $\beta_i \leqq \alpha_i$ for all $i \geqq 1$, $\{\alpha_i\}$ being the input list of $(A', \mathscr{B}')$. In view of Theorems 6.9 and 6.10, we may also assume that $(A^*, \mathscr{V})$ is controllable. We wish to exhibit the existence of a map T such that (7.1) and (7.2) hold. By Lemma 7.2, it is sufficient to demonstrate the existence of T under the assumption that both $(A', \mathscr{B}')$ and $(A^*, \mathscr{V})$ are in normal form. Let $v_1, \cdots, v_{m'}$ be a basis for $\mathscr{V}$ such that for each $i \geqq 1$, the set $v_1, \cdots, v_{\beta_i}$ is a basis for $\mathscr{V}_i$, and denote $\sigma_k = L_{\mathscr{V}}(\operatorname{sp}\{v_k\})$ for all $k \in \underline{m}'$. Similarly, let $b_1, \cdots, b_{m'}$ be a basis for $\mathscr{B}'$ such that for each $i \geqq 1$, the set $b_1, \cdots, b_{\alpha_i}$ is a basis for $\mathscr{B}'_i$, and denote $\sigma'_k = L_{\mathscr{B}'}(\operatorname{sp}\{b_k\})$ for all $k \in \underline{m}'$. Define the map T as follows: for each $i \in \underline{m}'$, let

$$T((A^*)^j v_i) = \begin{cases} (A')^j b_i & \text{for } j = 0, 1, \cdots, \sigma'_i - 1, \\ 0 & \text{for } \sigma'_i, \cdots, \sigma_i - 1 \quad \text{(whenever } \sigma_i > \sigma'_i\text{)}. \end{cases}$$

In view of the controllability assumption on $(A^*, \mathscr{V})$, the map T is completely specified, and an easy calculation shows that (7.1) holds. This completes the proof. □

We can now combine Theorem 6.9 with Theorem 7.3 to obtain a complete solution of the feedback simulation problem which is the main result of this section.

THEOREM 7.5. (Feedback simulation.) *Let $(A, \mathscr{B})$ be a linear system with input list $\{\gamma_i\}$. Then a system $(A', \mathscr{B}')$ is in the simulation orbit of $(A, \mathscr{B})$, if and only if there exists a nondecreasing sequence of positive integers $\{\beta_i\}$ such that the input list*

$\{\alpha_i\}$ *of* $(A', \mathscr{B}')$ *satisfies the following conditions*:

(i) $1 \leqq \beta_i \leqq \min\{\alpha_i, \gamma_i\}$,

(ii) $\lim \beta_i = \lim \alpha_i$,

(iii) $\sum_{i=1}^{k} i(\beta_i - \beta_{i-1}) \leqq \sum_{i=1}^{k} i(\gamma_i - \gamma_{i-1})$ *for all* $k \geqq 1$, $\beta_0 \triangleq 0$.

Remark 7.6. Theorem 7.5 involves an auxiliary list of integers $\{\beta_i\}$. One might hope that this list can be somehow eliminated and that a simpler condition involving only the input lists $\{\gamma_i\}$ and $\{\alpha_i\}$ holds. This however appears to be impossible. The computational verification of (i)–(iii) requires finding a sequence $\{\beta_i\}$ (whenever such a sequence exists), which is essentially a combinatorial problem. However, since most practical linear systems are of relatively low dimension, the computation is likely to be easily tractable even by exhaustive search procedures. Theorem 7.5 could of course also be stated in terms of the lists of controllability indices, but the conditions in that case are much more complicated.

We conclude with an easy and interesting corollary.

COROLLARY 7.7. *Let* $(A, \mathscr{B})$ *be a linear system with input list* $\{\gamma_i\}$, *and let* $(A', \mathscr{B}')$ *be a linear system such that* $\dim(\mathscr{B}') = \dim(\mathscr{B})$. *Then* $(A', \mathscr{B}')$ *is in the simulation orbit of* $(A, \mathscr{B})$, *if and only if the input list* $\{\alpha_i\}$ *of* $(A', \mathscr{B}')$ *satisfies the condition that* $\alpha_i \geqq \gamma_i$ *for all* $i \geqq 1$. *If, in addition,* $\dim(\langle A|\mathscr{B}\rangle) = \dim(\langle A'|\mathscr{B}'\rangle)$, *then the condition is* $\alpha_i = \gamma_i$ *for all* $i \geqq 1$.

Proof. Apply Lemma 6.4 to Theorem 7.5.

Remark 7.8. In terms of controllability indices, Corollary 7.7 reads as follows. Let $\{\sigma_1, \cdots, \sigma_m\}$ and $\{\sigma'_1, \cdots, \sigma'_m\}$ be the lists of controllability indices, respectively, of $(A, \mathscr{B})$ and $(A', \mathscr{B}')$ (where the lists are of the same length). Then $(A', \mathscr{B}')$ is in the simulation orbit of $(A, \mathscr{B})$, if and only if $\sigma'_1 \leqq \sigma_i$ for all $i \in \underline{m}$. If, in addition, both (controllable) systems are in spaces of the same dimension, the simulation condition is $\sigma'_i = \sigma_i$ for all $i \in \underline{m}$, which is precisely the feedback equivalence result of Brunovsky.

Acknowledgment. The author wishes to thank W. M. Wonham for various useful discussions.

REFERENCES

[1] P. BRUNOVSKY, *A classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–187.

[2] W. M. WONHAM AND A. S. MORSE, *Feedback invariants of linear multivariate systems*, Automatica, 8 (1972), pp. 93–100.

[3] R. E. KALMAN, *Kronecker invariants and feedback*, Proc. of Conf. on Ordinary Differential Equations, L. Weiss, ed., NRL Math. Res. Cent., 1971, pp. 459–471.

[4] A. S. MORSE, *Structural invariants of linear multivariate systems*, this Journal, 11 (1973), pp. 446–465.

[5] M. HEYMANN, *Structure and realization problems in the theory of dynamical systems*, Lecture notes CISM, Udine, Italy, 1972; Springer-Verlag, Wien, 1975.

[6] W. M. WONHAM AND A. S. MORSE, *Decoupling and pole assignment in linear multivariate systems: A geometric approach*, this Journal, 8 (1970), pp. 1–18.

[7] W. M. WONHAM, *Linear Multivarate Control, A Geometric Approach*, Springer-Verlag Lecture Notes in Economics and Methematical Systems, vol. 101, Springer-Verlag, Berlin, 1974.

[8] W. H. GREUB, *Linear Algebra*, Springer-Verlag, New York, 1967.

[9] M. E. WARREN AND A. E. ECKBERG, *On the dimensions of controllability subspaces: A characterization via polynomial matrices and Kronecker invariants*, this Journal, 13 (1975), pp. 434–445.

[10] M. HEYMANN, M. PACHTER AND R. J. STERN, *max-min control problems: A system theoretic approach*, IEEE Trans. Automatic Control, 1976.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

INCREMENTAL AND TOTAL OPTIMIZATION OF SEPARABLE FUNCTIONALS WITH CONSTRAINTS*

LAWRENCE D. STONE†

Abstract. Functionals E (real-valued) and C (vector-valued) are defined by $E(q) = \int_X e(x, q(x))\mu(dx)$ and $C(q) = \int_X c(x, q(x))\mu(dx)$, where μ is a Borel regular, nonatomic measure defined on a Borel subset X of a complete separable metric space. Let ω be the positive integers. Let $q_0, q_1, \cdots$, be extended real functions such that $q_0 = -\infty$ and $q_i \geqq q_{i-1}$ for $i \in \omega$. A function q^* is called *optimal* if $E(q^*) = \max\{E(q) : C(q) = C(q^*)\}$. The sequence $(q_1, q_2, \cdots)$ is *incrementally optimal* if $E(q_i) = \max\{E(p) : p \geqq q_{i-1}$ and $C(p) = C(q_i)\}$ for $i \in \omega$ and *totally optimal* if q_i is optimal for $i \in \omega$. Under appropriate measurability assumptions, it is shown that if $c(x, \cdot)$ is real-valued and increasing for $x \in X$, then an incrementally optimal sequence such that $|E(q_i)| < \infty$ and $C(q_i) \in$ interior range C for $i \in \omega$ is totally optimal. A counterexample is given to show that an extension of this result to multiple constraints fails even if $e(x, \cdot)$ and $c(x, \cdot)$ are linear for $x \in X$. In the case of a single constraint, the existence of optimal functions is proved under conditions which allow the range of C to be unbounded above.

1. Introduction. This paper investigates the relationship between incremental and total optimality for constrained optimization problems involving a real-valued separable effectiveness functional and a vector-valued separable cost functional. Existence of optimal functions is also considered.

A primary motivation for this investigation arises from search theory. In mathematical terms, the problem of finding the optimum distribution of search effort to detect a stationary object located in a subset X of Euclidean n-space becomes: find a function $q^* : X \to [0, \infty)$ such that $\int_X c(x, q^*(x))\,dx \leqq \Phi$ and

$$\int_X b(x, q^*(x))f(x)\,dx = \max\left\{\int_X b(x, q(x))f(x)\,dx : q \geqq 0 \quad \text{and} \quad \int_X c(x, q(x)) \leqq \Phi\right\}. \tag{1.1}$$

In (1.1), the function f gives the probability density of the target's location, $b(x, \cdot)$ is the local effectiveness function and $c(x, \cdot)$ the cost density function. In probability terms, $b(x, y)$ gives the conditional probability of detecting the target given it is located at x and the effort density is y at x. The above problem has an obvious analogue in case the search space X is discrete.

For the case where $b(x, y) = 1 - e^{-y}$ and $c(x, y) = y$ for $x \in X$ and $y \geqq 0$, Koopman [6, p. 617] made the following observation. Suppose one allocates Φ_1 amount of effort in an optimal fashion but fails to detect the target. An increment Φ_2 of effort then becomes available. If one allocates this additional effort in an incrementally optimal manner (i.e., optimal considering the previous allocation of Φ_1 amount of effort), then one obtains an optimal allocation of $\Phi_1 + \Phi_2$ effort. That is, two incrementally optimal allocations produce a totally optimal allocation. Koopman commented "This very convenient state of affairs seems to be a

* Received by the editors October 23, 1972, and in revised form January 3, 1974.

† Daniel H. Wagner, Associates, Paoli, Pennsylvania 19301. This research was supported by the Naval Analysis Programs, Office of Naval Research, under Contract N00014-69-C-0435.

characteristic property of the basic exponential law of search [i.e., local effectiveness function] assumed throughout."

For the case where $c(x, y) = y$ for $y \geqq 0$ and $x \in X$, the following results are known. Theorem 2.1 of [7] shows that for virtually any local effectiveness function, incrementally optimal allocations are totally optimal whenever the target's probability distribution is given by a density function as in (1.1). In the case where X is discrete, it is proven that concavity of $b(x, \cdot)$ for $x \in X$ guarantees that incrementally optimal allocations are totally optimal, and it is shown by counterexample that this property need not hold for discrete X if $b(x, \cdot)$ is not concave. Another paper on this subject is [3] (see discussion in [7]).

The results in [7] are very satisfactory for the case where $c(x, y) = y$ for $y \geqq 0$ and $x \in X$. However, search problems are not limited to this situation. Thus one is led to ask if these results hold for most general cost functions. In this paper, two possible generalizations are considered. The first, Theorem 3.1, generalizes the result of Theorem 2.1 of [7] to allow $c(x, \cdot)$ to be real-valued and increasing. In Example 3.2 it is shown by counterexample that the assumption that $c(x, \cdot)$ is increasing cannot be dropped. The second possible generalization is to allow for multiple constraints. That is, we allow $c(x, \cdot)$ to be vector-valued. For example, one might consider a search in which there is a constraint on both cost (in dollars) and time. For vector-valued cost functions, it is shown in Example 3.3 that no extension of the results of Theorem 3.1 is possible even if one assumes the linearity of $c(x, \cdot)$ and $e(x, \cdot)$ for $x \in X$. Example 3.3 also shows that total optimality of incrementally optimal allocations is not a consequence of the convexity of the range of the functionals or of the satisfaction of the pointwise multiplier rule by an optimal allocation.

While motivated by search theory, the results of this paper are stated in terms of a real-valued "effectiveness" functional E subject to an equality or inequality constraint on a vector-valued "cost" functional C. These functionals are separable, which means they are given by $E(q) = \int_X e(x, q(x))\mu(dx)$ and $C(q) = \int_X c(x, q(x))\mu(dx)$, where $q(x) \in Y(x)$ for $x \in X$ and X, Y, μ, e, and c are fixed. If we take μ to be Lebesgue measure on n-space and for $x \in X$ let $Y(x) = [0, \infty)$, and $e(x, y) = f(x)b(x, y)$ for $y \in Y(x)$, then we obtain the search situation considered above.

Section 4 of this paper concerns the existence of uniformly optimal search plans. In search theory, a uniformly optimal plan is an allocation in time and space which maximizes the probability of detection at each time t. This is the most desirable search plan since one can proceed with a plan which yields the long term goal of maximizing probability of detection by the end of the time allotted for the search without sacrificing any short term gain. In Theorem 4.3, it is shown that if $b(x, \cdot)$ is increasing and right-continuous, $c(x, \cdot)$ is extended real-valued and continuous and the probability distribution of the target is given by a probability density function with respect to a nonatomic measure, then a uniformly optimal plan exists under very general conditions which cover most situations likely to occur in search theory. In fact, the uniformly optimal plan is, in a sense, constructed in Theorem 4.3.

Theorem 4.3 is stated in terms of general separable functionals (one-dimensional) E and C. One feature of interest beyond search theory in Theorem 4.3,

is that it gives an existence theorem for optimal allocations without assuming that the range of C is bounded above. Theorem 4.3 is an extension of Theorem 3.3 of [7] to allow more general cost density functions c.

2. Preliminary definitions and assumptions. Throughout this paper, we assume that X is a Borel subset of a complete separable metric space and μ is a measure on X. For $x \in X$, we let $Y(x)$ be a subset of the extended real line $\bar{\mathscr{E}}_1$ with the usual topology. In order to use the results of [4] and [8], we use the definitions of measure, measurable set, and measurable function given in [4]. In order to use Corollary 5.2 of [8], we further assume that for each measurable $P \subset X$ for which $\mu(P) > 0$, there exists a measurable $Q \subset P$ such that $0 < \mu(Q) < \infty$. Following [4], we say that μ is Borel regular, if and only if all open sets of X are measurable and each set $A \subset X$ is contained in a Borel set B for which $\mu(A) = \mu(B)$.

For definiteness, the reader may wish to think of X as a Borel subset of Euclidean n-space and μ as Lebesgue measure on X. This identification will satisfy the measurability and topological hypotheses of all of the theorems in this paper. The phrase $x \in X$ is understood to mean almost every (in μ measure) $x \in X$, and a.e. stands for almost everywhere in μ measure.

Let $\Omega = \{(x, y) : x \in X \text{ and } y \in Y(x)\}$ be a Borel subset of $X \times \bar{\mathscr{E}}_1$, and let $c_1, \cdots, c_k$ and e be extended real-valued Borel functions defined on Ω. Using a framework very similar to [8] (which, however, does not permit e or c_i to assume $\pm\infty$), we let

$$\Psi = \{q : q \text{ is a function on } X \text{ and } q(x) \in Y(x) \text{ for } x \in X\},$$

$$\Xi = \{q : q \in \Psi \text{ and } c_1(\cdot, q(\cdot)), \cdots, c_k(\cdot\, q(\cdot)), e(\cdot, q(\cdot)) \text{ are measurable}\},$$

$$\Phi = \{q : q \in \Xi \text{ and } c_1(\cdot, q(\cdot)), \cdots, c_k(\cdot, q(\cdot)), e(\cdot, q(\cdot)) \text{ are integrable}\},$$

and

$$C_i(q) = \int_X c_i(x, q(x))\, \mu(dx) \quad \text{for } i = 1, \cdots, k, \quad q \in \Phi,$$

$$E(q) = \int_X e(x, q(x))\, \mu(dx) \quad \text{for } q \in \Phi.$$

Let ω be the set of positive integers. For $n \in \omega$, we let $\mathscr{E}_n$ be Euclidean n-space. If $a, b \in \mathscr{E}_n$, then $a \geqq b$ means $a_i \geqq b_i$ for $i = 1, \cdots, n$. Let $\mathscr{E}_n^+ = \{a : a \in \mathscr{E}_n \text{ and } a_i \geqq 0 \text{ for } i = 1, \cdots, n\}$. If a and $b \in \mathscr{E}_n$, 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 if $\infty - \infty$ occurs. We let $c = (c_1, \cdots, c_k)$, $C = (C_1, \cdots, C_k)$.

We define $q^* \in \Phi$ to be *optimal* if

$$E(q^*) = \max \{E(q) : C(q) = C(q^*)\}, \tag{2.1}$$

and we say that q^* is *strongly optimal* if

$$E(q^*) = \max \{E(q) : C(q) \leqq C(q^*)\}. \tag{2.2}$$

In this and similar usage, it is understood that $E(p) \in \{E(q) : C(q) = C(q^*)\}$ implies $E(q)$ exists.

An extended real-valued function f defined on a subset of the extended real line is said to be increasing if $y \geqq x$ implies $f(y) \geqq f(x)$.

3. Incremental optimization. For $i \in \omega$, let $q_i \in \Phi$ be such that $q_1 \leqq q_2 \cdots$. Let $q_0(x) = -\infty$ for $x \in X$, and define $C(q_0) = (-\infty, \cdots, -\infty)$, whether or not $q_0 \in \Phi$. If for $i \in \omega$,

$$E(q_i) = \max \{E(q) : q \geqq q_{i-1},\ q \in \Phi \text{ and } C(q) = C(q_i)\}, \tag{3.1}$$

then we say that $(q_1, q_2, \cdots)$ is an *incrementally optimal* sequence. If for $i \in \omega$,

$$E(q_i) = \max \{E(q) : q \geqq q_{i-1},\ q \in \Phi \text{ and } C(q_{i-1}) \leqq C(q) \leqq C(q_i)\}, \tag{3.2}$$

then we say that $(q_1, q_2, \cdots)$ is a *strong incrementally optimal* sequence. If q_i is optimal (strongly optimal) for each $i \in \omega$, then $(q_1, q_2, \cdots)$ is said to be a *totally optimal* (*strong totally optimal*) sequence.

Conceptually, an incrementally optimal sequence $(q_1, q_2, \cdots)$ is one such that for $i \in \omega$, q_{i+1} obtains the maximum effectiveness from the increment of cost $C(q_{i+1}) - C(q_i)$ given the previous allocation q_i. If for each i, q_i is an optimal allocation of $C(q_i)$, then the sequence is totally optimal.

Under the primary conditions of a single cost constraint and increasing cost function, we show that an incrementally optimal sequence is totally optimal. Define

$$\begin{aligned} l(x, y, \lambda) &= e(x, y) - \lambda \cdot c(x, y) && \text{for } (x, y) \in \Omega \text{ and } \lambda \in \mathscr{E}_k, \\ M(x, \lambda) &= \sup \{l(x, y, \lambda) : y \in Y(x)\} && \text{for } x \in X, \quad \lambda \in \mathscr{E}_k. \end{aligned} \tag{3.3}$$

when neither $\infty - \infty$ nor $0 \cdot \infty$ occurs.

Following [8], we say that $q \in \Psi$ *satisfies* (*strongly satisfies*) *the pointwise multiplier rule* if for some $\lambda \in \mathscr{E}_k$ (some $\lambda \in \mathscr{E}_k^+$),

$$l(x, q(x), \lambda) = M(x, \lambda) \quad \text{for } x \in X.$$

In order to make use of Corollary 5.2 of [8], we note that the extended real line is a complete separable metric space under the following metric:

$$d(x, y) = |\arctan(x) - \arctan(y)| \quad \text{for } x, y \in \bar{\mathscr{E}}_1,$$

where $\arctan(-\infty) = -\pi/2$ and $\arctan(\infty) = \pi/2$.

THEOREM 3.1. *Assume Ω, e, and c are Borel. Let μ be Borel regular and non-atomic. Let e be real-valued and for $x \in X$, let $c(x, \cdot)$ be real-valued and increasing. If $(q_1, q_2, \cdots)$ is a (strong) incrementally optimal sequence such that for $i \in \omega$, $|E(q_i)| < \infty$ and $C(q_i)$ is in the interior of the range of C, then $(q_1, q_2, \cdots)$ is a (strong) totally optimal sequence.*

Proof. By Corollary 5.2 of [8], there exists a $\lambda^1 \in \mathscr{E}_1$ such that for $x \in X$,

$$l(x, q_1(x), \lambda^1) = M(x, \lambda^1) \tag{3.4}$$

and $\lambda^2 \in \mathscr{E}_1$ such that for $x \in X$,

$$l(x, q_2(x), \lambda^2) \geqq l(x, y, \lambda^2) \quad \text{for } q_1(x) \leqq y \in Y(x). \tag{3.5}$$

In order to prove that q_2 is optimal (strongly optimal), it is sufficient, by Theorem 2.1 of [8], to show that there exists $\lambda \in \mathscr{E}_1(\lambda \geqq 0)$ such that

$$l(x, q_2(x), \lambda) = M(x, \lambda) \quad \text{for } x \in X. \tag{3.6}$$

By (3.4) and (3.5),

$$\begin{aligned} \lambda^2[c(x, q_2(x)) - c(x, q_1(x))] &\leqq e(x, q_2(x)) - e(x, q_1(x)) \\ &\leqq \lambda^1[c(x, q_2(x)) - c(x, q_1(x))] \quad \text{for } x \in X. \end{aligned} \tag{3.7}$$

If $c(x, q_1(x)) = c(x, q_2(x))$ for $x \in X$, then by (3.7), $e(x, q_1(x)) = e(x, q_2(x))$, and

$$e(x, q_2(x)) - \lambda^1 c(x, q_2(x)) = e(x, q_1(x)) - \lambda^1 c(x, q_1(x)) = M(x, \lambda^1) \quad \text{for } x \in X.$$

Thus (3.6) is satisfied for $\lambda = \lambda^1$. If $c(x, q_2(x)) > c(x, q_1(x))$ for x in a set of positive measure, then $\lambda^2 \leqq \lambda^1$ by (3.7). Suppose $y \in Y(x)$ and $y < q_1(x)$; then for $x \in X$,

$$\begin{aligned} 0 &\leqq e(x, q_1(x)) - e(x, y) - \lambda^1[c(x, q_1(x)) - c(x, y)] \\ &\leqq e(x, q_1(x)) - e(x, y) - \lambda^2[c(x, q_1(x)) - c(x, y)] \end{aligned}$$

and

$$l(x, y, \lambda^2) \leqq l(x, q_1(x), \lambda^2) \leqq l(x, q_2(x), \lambda^2) \quad \text{for } q_1(x) > y \in Y(x). \tag{3.8}$$

Combining (3.8) and (3.5), we obtain (3.6) with $\lambda = \lambda^2$. Thus q_2 satisfies the pointwise multiplier rule and, by Theorem 2.1 of [8], q_2 is optimal. By repeating the above argument for $q_3, q_4, \cdots$, the theorem is proved for optimality. The assertions concerning strong optimality follow by observing that in this case, Corollary 5.2 of [8] yields $\lambda^1, \lambda^2 \geqq 0$. Thus the number λ obtained to satisfy (3.6) is nonnegative.

Example 3.2. Theorem 3.1 does not remain true if one drops the assumption that $c(x, \cdot)$ is increasing for $x \in X$. To see this, we consider the situation where $X = [0, 1]$ and for $x \in X$, $Y(x) = [0, 3]$,

$$e(x, y) = \begin{cases} y, & 0 \leqq y \leqq 1, \\ -y + 2, & 1 \leqq y \leqq 3, \end{cases} \qquad c(x, y) = \begin{cases} y, & y \neq \frac{1}{2}, \\ 2, & y = \frac{1}{2}. \end{cases}$$

For $x \in X$, let $q_1(x) = 1$ and $q_2(x) = 2$. One may check that (q_1, q_2) is an incrementally optimal sequence by noting that for $x \in X$,

$$l(x, 1, 1) = M(x, 1),$$

$$e(x, 2) + c(x, 2) \geqq e(x, y) + c(x, y) \quad \text{for } y \geqq 1.$$

However, by taking $h(x) = \frac{1}{2}$, for $x \in X$, we find that

$$E(h) = \tfrac{1}{2} > 0 = E(q_2) \quad \text{and} \quad C(h) = 2 = C(q_2),$$

so that q_2 is not totally optimal.

A similar example can be constructed even if one requires that $c(x, \cdot)$ be continuous for $x \in X$.

Example 3.3. We now show that Theorem 3.1 cannot be extended to multiple constraints even when one requires that $e(x, \cdot)$, $c_i(x, \cdot)$ be linear for $i = 1, \cdots, k$ and $x \in X$. Since a linear function is both concave and convex, this shows that no combination of convexity/concavity assumptions will guarantee that incre-

mentally optimal sequences are totally optimal. That strict concavity or convexity may be sufficient is not ruled out by the counterexample; however, it seems unlikely that these assumptions would suffice.

Let $X = [0, 4]$, $Y(x) = [0, \infty)$ for $x \in X$. For $y \geqq 0$, let

$$e(x, y) = y, \quad c_1(x, y) = 4y, \quad c_2(x, y) = 2y \quad \text{for } 0 \leqq x \leqq 1,$$
$$e(x, y) = \tfrac{3}{4}y, \quad c_1(x, y) = y, \quad c_2(x, y) = 2y \quad \text{for } 1 \leqq x \leqq 2,$$
$$e(x, y) = \tfrac{1}{2}y, \quad c_1(x, y) = y, \quad c_2(x, y) = \tfrac{9}{8}y \quad \text{for } 2 < x \leqq 3.$$
$$e(x, y) = 0, \quad c_1(x, y) = y, \quad c_2(x, y) = 4y \quad \text{for } 3 < x \leqq 4.$$

By Blackwell's extension (see [2]) of Lyapunov's theorem, the range of C is convex. Let (z_1, z_2) represent a point in 2-space. Then one can check that the range of C is the region in the nonnegative quadrant of 2-space which lies between the line $z_2 = \frac{1}{2}z_1$ and the line $z_2 = 4y_1$. Let

$$q_1(x) = \begin{cases} 0, & 0 \leqq x \leqq 1, \\ 1, & 1 < x \leqq 2, \\ 0, & 2 < x \leqq 4, \end{cases} \qquad q_2(x) = \begin{cases} 1, & 0 \leqq x \leqq 1, \\ 1, & 1 < x \leqq 2, \\ 0, & 2 < x \leqq 4. \end{cases}$$

Note that $C[q_1] = (1, 2)$ and $C[q_2] = (5, 4)$ are in the interior of the range of C. By choosing $\lambda^1 = (\frac{1}{4}, \frac{1}{4})$ and $\lambda^2 = (\frac{1}{40}, \frac{9}{20})$, one may check that for $x \in X$,

$$e(x, q_1(x)) - \lambda^1 \cdot c(x, q_1(x)) = M(x, \lambda^1)$$

and

$$e(x, q_2(x)) - \lambda^2 \cdot c(x, q_2(x)) \geqq e(x, y) - \lambda^2 \cdot c(x, y) \quad \text{for } y \geqq q_1(x).$$

Thus (q_1, q_2) is a strong incrementally optimal sequence. However, q_2 is not optimal, much less strongly optimal. To see this, define

$$h(x) = \begin{cases} 0, & 0 \leqq x \leqq \frac{7}{48}, \\ 1, & \frac{7}{48} < x \leqq 1, \\ 0, & 1 < x \leqq 1 + \frac{5}{12}, \\ 1, & 1 + \frac{5}{12} < x \leqq 3, \\ 0, & 3 < x \leqq 4. \end{cases}$$

Then one may check that $C(h) = C(q_2) = (5, 4)$, but $E(h) = 1 + \frac{19}{24} > 1 + \frac{3}{4} = E(q_2)$.

Observe that this example satisfies the conditions of Corollary 5.2 of [8] so that optimal allocations satisfy a pointwise multiplier rule. Although this is the main tool used to prove Theorem 3.1, it is not sufficient for the analog of Theorem 3.1 to hold for multiple constraints.

4. Existence of optimal allocations. In this section, we prove the existence of uniformly optimal allocations for a single constraint involving a continuous, increasing cost function under assumptions which are natural, for example, for search theory. Lemmas 4.1 and 4.2 below will be used to prove the main existence

result, Theorem 4.3. Throughout this section, we take $c(x, \cdot)$ to be extended real-valued (i.e., $k = 1$) and denote c_1 by c.

The following lemma is an extension of Halkin's Proposition 8.3 in [5] which is proved for totally finite measure spaces. The extension to σ-finite measure spaces stated in the lemma is routine.

LEMMA 4.1. *Let μ be a nonatomic σ-finite measure on X. Then there is a family $\{S_\alpha : \alpha \in [0, 1]\}$ of measurable sets such that*

(4.1) (i) $S_0 = \phi, \quad S_1 = X \quad$ *and* $\quad \alpha < \beta$ *implies* $S_\alpha \subset S_\beta$,

(4.1) (ii) $\mu(S_\alpha) < \infty \quad$ *for* $0 \leqq \alpha < 1$,

(4.1)(iii) $\lim_{\alpha \to \beta} \mu(S_\alpha) = \mu(S_\beta) \quad$ *for* $\beta \in [0, 1]$,

(4.1)(iv) $\lim_{\alpha \uparrow \beta} S_\alpha = S_\beta \quad$ *for* $\beta \in [0, 1]$.

For $x \in X$, let

$$T(x) = \inf \{y : y \in Y(x)\}, \qquad U(x) = \sup \{y : y \in Y(x)\}. \tag{4.2}$$

Suppose $Y(x)$ is compact in $\bar{\mathscr{E}}_1$, $e(x, \cdot)$ is real-valued, increasing and right-continuous and $c(x, \cdot)$ is continuous and increasing for $x \in X$. Then for $x \in X$ we define

$$\varphi(x, \lambda) = \sup \{y : y \in Y(x) \text{ and } l(x, y, \lambda) = M(x, \lambda)\} \quad \text{for } \lambda > 0,$$

$$\xi(x, \lambda) = \lim_{\lambda' \downarrow \lambda} \varphi(x, \lambda') \quad \text{for } \lambda \geqq 0.$$

Note that $Y(x)$ need not be an interval. Since $Y(x)$ is compact and $l(x, \cdot, \lambda)$ is upper semicontinuous for $\lambda > 0$, $M(x, \lambda)$ is achieved on $Y(x)$ and $\varphi(x, \lambda)$ is well-defined. We let

$$I(\lambda) = \int_X c(x, \varphi(x, \lambda))\mu(dx) \quad \text{for } \lambda > 0,$$

when the integral exists.

LEMMA 4.2. *Assume Ω, e, and c are Borel. Let μ be σ-finite, nonatomic and Borel regular. Suppose $-\infty < E(T) \leqq E(U) < \infty$, $|C(T)| < \infty$,* and *for $x \in X$, $Y(x)$ is compact in $\bar{\mathscr{E}}_1$, $e(x, \cdot)$ is increasing and right-continuous, and $c(x, \cdot)$ is continuous, increasing, and extended real-valued. Then the following hold*:

(a) *$\varphi(\cdot, \lambda) \in \Phi$ and $I(\lambda)$ is finite for $\lambda > 0$;*

(b) *$\varphi(x, \cdot)$ is decreasing for $x \in X$ and I is decreasing,*

(c) *$\varphi(x, \cdot)$ is left-continuous for $x \in X$ and I is left-continuous,*

(d) *For $\lambda > 0$ we may find $f : X \times [C(\xi(\cdot, \lambda)), I(\lambda)] \to \bar{\mathscr{E}}_1$ such that* (i) *$f(x, \cdot)$ is increasing for $x \in X$,* (ii) *for $t \in [C(\xi(\cdot, \lambda)), I(\lambda)]$, $C(f(\cdot, t)) = t$, and* (iii),

$$l(x, f(x, t), \lambda) = M(x, \lambda) \quad \text{for } x \in X.$$

Proof. Since $Y(x)$ is compact for $x \in X$, $U \in \Psi$. To see that $U \in \Xi$, let $\pi(x, y) = x$ for $(x, y) \in \Omega$ and $R = \Omega \cap \{(x, y) : y \geqq a\}$. Then R is Borel, and $\{x : U(x) \geqq a\} = \pi(R)$. Since X and $\bar{\mathscr{E}}_1$ are complete separable metric spaces, so is $X \times \bar{\mathscr{E}}_1$. Thus by 2.2.13 of [4], $\pi(R)$ is measurable and $U \in \Xi$. A similar argument shows $T \in \Xi$.

Since $-\infty < E(T) \leqq E(U) < \infty$, we must have $e(x, \cdot)$ real-valued for $x \in X$. The compactness of $Y(x)$ and the upper semi-continuity of $l(x, \cdot, \lambda)$ guarantee that $\varphi(x, \lambda) \in Y(x)$ and

$$l(x, \varphi(x, \lambda), \lambda) = M(x, \lambda) \quad \text{for } x \in X \quad \text{and} \quad \lambda > 0.$$

To show that $\varphi(\cdot, \lambda) \in \Xi$, take $\lambda > 0$, $a \in \bar{\mathscr{E}}_1$, and let

$$R = \Omega \cap \{(x, y) : l(x, y, \lambda) > a\}.$$

Since Ω, e, and c are Borel, R is Borel. By the same reasoning as above, $\{x : M(x, \lambda) > a\} = \pi(R)$ is measurable, and by 2.3.6 of [4], $M(\cdot, \lambda)$ is equal a.e. to a Borel function $\tilde{M}(\cdot, \lambda)$. Similarly,

$$\{x : \varphi(x, \lambda) > a\} = \pi\{(x, y) : l(x, y, \lambda) = \tilde{M}(x, y) \text{ and } y > a\}$$

is measurable and $\varphi(\cdot, \lambda) \in \Xi$.

By virtue of $|C(T)| < \infty$, we have $c(x, T(x))$ is finite for $x \in X$. Since $e(x, \cdot)$ is increasing, we have

$$\begin{aligned} e(x, U(x)) - e(x, T(x)) &\geqq e(x, \varphi(x, \lambda)) - e(x, T(x)) \\ &\geqq \lambda[c(x, \varphi(x, \lambda)) - c(x, T(x))]. \end{aligned}$$

Hence

$$-\infty < C(T) \leqq I(\lambda) \leqq (1/\lambda)[E(U) - E(T)] + C(T) < \infty,$$

which shows that I is finite, and $\varphi(\cdot, \lambda) \in \Phi$. This proves (a).

Suppose $0 < \lambda^1 < \lambda^2$. Let $y_1(x) = \varphi(x, \lambda^1)$ and $y_2(x) = \varphi(x, \lambda^2)$ for $x \in X$. Then for $x \in X$,

$$l(x, y_1(x), \lambda^1) \geqq l(x, y_2(x), \lambda^1) \quad \text{and} \quad l(x, y_2(x), \lambda^2) \geqq l(x, y_1(x), \lambda^2)$$

which implies

$$\begin{aligned} \lambda^1[c(x, y_1(x)) - c(x, y_2(x))] &\leqq e(x, y_1(x)) - e(x, y_2(x)) \\ &\leqq \lambda^2[c(x, y_1(x)) - c(x, y_2(x))]. \end{aligned} \tag{4.3}$$

By virtue of the fact that $0 < \lambda^1 < \lambda^2$, we must have $c(x, y_1(x)) \geqq c(x, y_2(x))$ for $x \in X$; otherwise, (4.3) yields a contradiction. If $c(x, y_1(x)) = c(x, y_2(x))$, then $e(x, y_2(x)) = e(x, y_1(x))$, and using the definition of $\varphi(x, \cdot)$, one may show that $y_2(x) = y_1(x)$. If $c(x, y_1(x)) > c(x, y_2(x))$, then the increasing nature of $c(x, \cdot)$ yields $y_1(x) > y_2(x)$. Thus $\varphi(x, \cdot)$ is a decreasing function for $x \in X$. Since $c(x, \cdot)$ is continuous and increasing for $x \in X$, one may show that I is decreasing. This proves (b).

To prove (c), we first show that for $x \in X$, $M(x, \cdot)$ is continuous. Choose $0 < \lambda^1 < \lambda^2 < \infty$. Observe that $\varphi(x, \lambda^1) \geqq \varphi(x, \lambda^2)$ and define

$$K(x) = \sup\{|c(x, y)| : \varphi(x, \lambda^2) \leqq y \leqq \varphi(x, \lambda^1)\} \quad \text{for } x \in X.$$

Suppose that $M(x, \lambda^1) \geqq M(x, \lambda^2)$. Since $l(x, \varphi(x, \lambda_1), \lambda_2) \leqq l(x, \varphi(x, \lambda_2), \lambda_2)$,

$$\begin{aligned}|M(x, \lambda^1) - M(x, \lambda^2)| &\leqq |l(x, \varphi(x, \lambda^1), \lambda^1) - l(x, \varphi(x, \lambda^1), \lambda^2)| \\ &\leqq \sup \{|l(x, y, \lambda^1) - l(x, y, \lambda^2)| : \varphi(x, \lambda^2) \leqq y \leqq \varphi(x, \lambda^1)\} \\ &\leqq |\lambda^1 - \lambda^2| \sup \{|c(x, y)| : \varphi(x, \lambda^2) \leqq y \leqq \varphi(x, \lambda^1)\} \\ &\leqq |\lambda^1 - \lambda^2| K(x).\end{aligned}$$

Making a similar argument when $M(x, \lambda^1) < M(x, \lambda^2)$, we show that $M(x, \cdot)$ is continuous for $x \in X$.

Fix $x \in X$ such that $M(x, \cdot)$ is continuous. Let $\lambda_i \uparrow \lambda_0$ and define $y_i = \varphi(x, \lambda_i)$ for $i = 0, 1, 2, \cdots$. Then $\{y_i\}_{i=1}^{\infty}$ is a decreasing sequence with a limit $z \in Y(x)$. Moreover, $z \geqq y_0$, and

$$\lim_{i \to \infty} l(x, y_i, \lambda_i) = \lim_{i \to \infty} M(x, \lambda_i) = M(x, \lambda_0) = l(x, y_0, \lambda_0).$$

However, by the upper semi-continuity of $e(x, \cdot)$ and continuity of $c(x, \cdot)$,

$$l(x, y_0, \lambda_0) = \lim_{i \to \infty} l(x, y_i, \lambda_i) \leqq l(x, z, \lambda_0).$$

Hence $l(x, z, \lambda_0) = l(x, y_0, \lambda_0)$, and by definition of $y_0, z = y_0$. It follows that $\varphi(x, \cdot)$ is left-continuous for $x \in X$. Since $c(x, \cdot)$ is continuous and increasing for $x \in X$, we may apply the monotone convergence theorem to complete the proof of (c).

We claim $l(x, \xi(x, \lambda), \lambda) = M(x, \lambda)$ for $x \in X$. This follows from the continuity of $M(x, \cdot)$ and $c(x, \cdot)$ along with the upper semi-continuity of $e(x, \cdot)$ for $x \in X$ as follows:

$$\begin{aligned}l(x, \xi(x, \lambda), \lambda) &\geqq \lim_{\lambda' \downarrow \lambda} l(x, \varphi(x, \lambda'), \lambda') \\ &= \lim_{\lambda' \downarrow \lambda} M(x, \lambda') = M(x, \lambda).\end{aligned}$$

Thus $l(x, \xi(x, \lambda), \lambda) = M(x, \lambda)$ for $x \in X$.

To prove (d) we use the family $\{S_\alpha : \alpha \in [0, 1]\}$ of Lemma 4.1. Fix $\lambda > 0$ and let

$$h_\alpha(x) = \begin{cases} \varphi(x, \lambda) & \text{for } x \in S_\alpha, \\ \xi(x, \lambda) & \text{for } x \in X - S_\alpha, \quad \alpha \in [0, 1]. \end{cases}$$

Note that $l(x, h_\alpha(x), \lambda) = M(x, \lambda)$ for $x \in X$. Since $\varphi(x, \lambda) \geqq \xi(x, \lambda)$ for $x \in X$, one may use (i) and (iv) of Lemma 4.1 and the monotone convergence theorem to show that $C(h_\alpha)$ is a left continuous function of α. A similar argument using (i)–(iii) of Lemma 4.1 shows that $C(h_\alpha)$ is a right-continuous (hence continuous) function of α. The existence of f in (d) now follows readily. This proves the lemma.

Theorem 4.3 below proves the existence of uniformly optimal allocation schedules. Let J be an interval of extended reals. An *allocation schedule* over J is a Borel function η defined on $X \times J$ such that for $x \in X$, $\eta(x, \cdot)$ is increasing and $\eta(x, v) \in Y(x)$ for $v \in J$. We say that an allocation schedule η is *uniformly optimal* over J if

$$E(\eta(\cdot, v)) = \max \{E(q) : C(q) \leqq v\} \quad \text{for } v \in J.$$

This definition is an extension of the definition of uniform optimality given in [1] in conjunction with search theory. The notion is a natural one for search in that one desires a search plan (allocation schedule) such that one is always achieving the maximum probability of detection from the effort expended.

In Theorem 4.3 below, one would like to prove the existence of a uniformly optimal allocation schedule over $[C(T), C(U)]$. However, this is not always possible since, for example, the range of C need not be connected.

THEOREM 4.3. *Assume Ω, e, and c are Borel. Let μ be σ-finite, nonatomic, and Borel regular. Suppose $-\infty < E(T) \leqq E(U) < \infty$, $|C(T)| < \infty$ and for $x \in X$, $Y(x)$ is compact in $\bar{\mathscr{E}}_1$, $e(x, \cdot)$ is increasing and right-continuous, and $c(x, \cdot)$ is continuous increasing, and extended real-valued. Then there exists an allocation schedule η which is uniformly optimal over $[C(T), C(V))$, where $V = \xi(\cdot, 0)$. Moreover, $E(V) = \max\{E(q) : C(q) < \infty\}$, and if $C(U) < \infty$, then*

$$E(V) = \max\{E(q) : q \in \Phi\} \tag{4.4}$$

and extending η by defining $\eta(\cdot, C(V)) = V$, one obtains an allocation schedule which is uniformly optimal over $[C(T), C(V)]$.

Proof. First we define

$$g(x) = \lim_{\lambda \uparrow \infty} \varphi(x, \lambda).$$

We claim that $c(x, g(x)) = c(x, T(x))$ for $x \in X$. To see this, we note that $\varphi(x, \lambda) \geqq g(x)$ for $0 < \lambda < \infty$, and suppose $c(x, g(x)) > c(x, T(x))$ for x in a set of positive measure. Then for such x,

$$\infty > \lambda_0 \equiv \frac{e(x, U(x)) - e(x, T(x))}{c(x, g(x)) - c(x, T(x))} \geqq \frac{e(x, \varphi(x, \lambda)) - e(x, T(x))}{c(x, \varphi(x, \lambda)) - c(x, T(x))} \quad \text{for } \lambda > 0,$$

which contradicts $l(x, \varphi(x, \lambda), \lambda) = M(x, \lambda)$ for $\lambda > \lambda_0$. Thus the claim is verified.

In addition, g is strongly optimal. To see this, we observe that $C(g) = C(T)$ and suppose that g is not strongly optimal. That is, there exists $q^* \in \Phi$ such that $E(q^*) > E(g)$ and $C(q^*) \leqq C(g)$. Since $c(x, \cdot)$ is increasing,

$$c(x, q^*(x)) = c(x, T(x)) = c(x, g(x)) \quad \text{for } x \in X. \tag{4.5}$$

Since $E(q^*) > E(g)$, there exists a set P with positive measure such that

$$e(x, q^*(x)) > e(x, g(x)) \quad \text{for } x \in P. \tag{4.6}$$

For $x \in P$, the increasing nature of $e(x, \cdot)$ yields that $q^*(x) > g(x)$. Hence we may choose λ_x such that $q^*(x) > \varphi(x, \lambda_x) > g(x)$. Then

$$e(x, \varphi(x, \lambda_x)) - \lambda_x c(x, \varphi(x, \lambda_x)) \geqq e(x, q^*(x)) - \lambda_x c(x, q^*(x)). \tag{4.7}$$

Again, the increasing nature of $c(x, \cdot)$ along with (4.5) implies that $c(x, \varphi(x, \lambda_x)) = c(x, q^*(x))$. Equation (4.7) now yields $e(x, \varphi(x, \lambda_x)) \geqq e(x, q^*(x))$. The increasing nature of $e(x, \cdot)$ implies that $e(x, \varphi(x, \lambda_x)) = e(x, q^*(x))$. By the same argument as above, we may show that $e(x, \varphi(x, \lambda)) = e(x, q^*(x))$ for $\lambda \geqq \lambda_x$. The definition of $g(x)$ and the right continuity of $e(x, \cdot)$ may now be combined to prove that $e(x, q^*(x)) = e(x, g(x))$ for $x \in P$, which contradicts (4.6). Thus $E(q^*) \leqq E(g)$ and

g is strongly optimal. Define

$$\eta(x, C(T)) = g(x) \quad \text{for } x \in X.$$

Now I is monotone so it has only a countable number of discontinuities. Let N be a countable index set such that $\{\lambda_n : n \in N\}$ is the set of discontinuity points of I. Let $J_n = [C(\xi(\cdot, \lambda_n)), I(\lambda_n)]$ for $n \in N$. Then the intervals J_n are disjoint and are the jump intervals at the discontinuity points of I. Note that $\lim_{\lambda \downarrow 0} I(\lambda) = C(V)$. For $v \in (C(T), C(V)) - \bigcup_{n \in N} J_n$, let

$$\lambda^*(v) = \sup \{\lambda : I(\lambda) = v\}.$$

By the left continuity of I, $I(\lambda^*(v)) = v$. For $v \in J_n$, let $\lambda^*(v) = \lambda_n$. By (d) of Lemma 4.2, we may find a function f_n defined on $X \times J_n$ such that $f_n(x, \cdot)$ is increasing for $x \in X$ and for $v \in J_n$, $f_n(\cdot, v) \in \Phi$, $C(f_n(\cdot, v)) = v$, and

$$l(x, f_n(x, v), \lambda_n) = M(x, \lambda_n) \quad \text{for } x \in X.$$

We now define for $x \in X$,

$$\eta(x, v) = \begin{cases} \varphi(x, \lambda^*(v)) & \text{if } v \in (C(T), C(V)) - \bigcup_{n \in N} J_n, \\ f_n(x, v) & \text{if } v \in J_n \quad \text{for some } n \in N. \end{cases}$$

Then for $v \in (C(T), C(V))$, $C(\eta(\cdot, v)) = v$ and $\eta(\cdot, v)$ satisfies the pointwise multiplier rule with $\lambda = \lambda^*(v) \geqq 0$. Thus $\eta(\cdot, v)$ is strongly optimal by Theorem 2.1 of [8].

To prove that $\eta(x, \cdot)$ is increasing for $x \in X$, we let v and s be such that $C(T) < v < s < C(V)$. Then there exists a set P such that $\mu(P) > 0$ and

$$c(x, \eta(x, s)) > c(x, \eta(x, v)) \quad \text{for } x \in P.$$

Since (3.7) holds with λ^1, λ^2, q_1, and q_2 replaced by $\lambda(v)$, $\lambda(s)$, $\eta(\cdot, v)$ and $\eta(\cdot, s)$ respectively, $\lambda^*(s) \leqq \lambda^*(v)$. Suppose $\lambda^*(s) < \lambda^*(v)$. Then $\eta(x, s) \geqq \xi(x, \lambda^*(s)) \geqq \varphi(x, \lambda^*(v)) \geqq \eta(x, v)$ for $x \in X$. If $\lambda^*(s) = \lambda^*(v)$, then v and s are both in the same J_n for some $n \in N$ and $\eta(x, s) \geqq \eta(x, v)$ for $x \in X$ by construction. Since $\eta(x, C(T)) = g(x) \leqq \varphi(x, \lambda)$ for $x \in X$ and $\lambda > 0$, we have $\eta(x, \cdot)$ is increasing for $x \in X$. Thus η is uniformly optimal over $[C(T), C(V))$.

Suppose $q \in \Phi$ such that $|C(q)| < \infty$. Then $|c(x, q(x))| < \infty$ for $x \in X$, and we have

$$\begin{aligned} e(x, V(x)) &\geqq \lim_{\lambda \downarrow 0} e(x, \varphi(x, \lambda)) - \lambda c(x, \varphi(x, \lambda)) \\ &\geqq \lim_{\lambda \downarrow 0} e(x, q(x)) - \lambda c(x, q(x)) = e(x, q(x)) \quad \text{for } x \in X. \end{aligned}$$

Thus $E(V) = \max \{E(q) : |C(q)| < \infty\}$. If $C(U) < \infty$, then (4.4) follows, and by setting $\eta(x, C(V)) = V(x)$ for $x \in X$, the theorem follows.

REFERENCES

[1] V. I. ARKIN, *Uniformly optimal strategies in search problems*, Theor. Probability Appl., 9 (1964), pp. 674–680.

[2] D. BLACKWELL, *The range of certain vector integrals*, Proc. Amer. Math. Soc., 2 (1951), pp. 390–395.

[3] J. M. DOBBIE, *Search theory: A sequential approach*, Naval. Res. Logist. Quart., 4 (1963), pp. 323–334.

[4] H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.
[5] H. HALKIN, *On the necessary condition for optimal control of nonlinear systems*, J. Analyse Math., 7 (1964), pp. 1–82.
[6] B. O. KOOPMAN, *The theory of search. III: The optimum distribution of searching effort*, Operations Res., 5 (1957), pp. 613–626.
[7] L. D. STONE, *Total optimality of incrementally optimal allocations*, Naval Res. Logist. Quart., 20 (1973), pp. 419–430.
[8] 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.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

A STOCHASTIC MAXIMUM PRINCIPLE*

VIRGINIA M. WARFIELD†

Abstract. The major theorem of this paper is very closely parallel to the classical Pontryagin maximum principle. The classical case, very roughly stated, says that if $u(t)$ is a control function which has an associated trajectory $x(t)$, then there is a function $H(v, x, t)$ such that $u(t)$ is optimal only if for each t and for all v in the control set,

$$H(u(t), x(t), t) \leqq H(v, x(t), t).$$

Our stochastic case of the open loop problem, stated even more roughly, says that there is a function $H(v, x, t, \omega)$ such that a control function $u(t)$ with associated trajectory $x(t, \omega)$ is optimal only if for all t and for all v in the control set,

$$E\{H(u(t), x(t, \omega), t, \omega)\} \leqq E\{H(v, x(t, \omega), t, \omega)\}.$$

Using this result, we then proceed to define a process whereby a control can be tested for optimality in the closed loop case, where information is acquired at a finite number of times.

Throughout the paper, the trajectories are determined by a stochastic integral equation. The stochastic integrals used are McShane's first and second order belated integrals.

1. The stochastic maximum principle. A more detailed statement of the stochastic maximum principle requires a considerable amount of machinery. We shall initially state only what is needed for the open loop case.

Notations and conventions.

Q is a compact subset of $\mathbb{R}^m$.

$T > 0$; we work on $[0, T]$.

Trajectories $x(t, \omega)$ are in $\mathbb{R}^n$.

F^T is a function from $\mathbb{R}^n$ to $\mathbb{R}$.

$f^0, \cdots, f^n$ are functions from $[0, T] \times \mathbb{R}^n \times Q$ to $\mathbb{R}$.

$g^i_\rho, \cdots, g^n_\rho; G^1_{\rho\sigma}, \cdots, G^n_{\rho\sigma}$ $(\sigma, \rho = 1, \cdots, r)$ are functions from $[0, T] \times \mathbb{R}^n$ to $\mathbb{R}$.

u is a continuous function from $[0, T]$ to Q.

$z^\rho(t, \omega)$ are stochastic processes on $[0, T]$.

For $x \in \mathbb{R}^n$, $|x|$ is the usual Euclidean norm of x.

For a random variable X, $\|X\| = \sqrt{E(|X|^2)}$.

For a set A, 1_A is the indicator function for A.

Definitions and hypotheses. A control function $u(t)$ and a trajectory $x(t, \omega)$ correspond to each other with initial distribution $y(\omega)$ if

$$\text{(I)} \qquad \begin{aligned} x^i(t, \omega) = y^i(\omega) &+ \int_0^t f^i(s, x(s, \omega\ , u(s))\, ds + \int_0^t \sum_\rho g^i_\rho(s, x(s, \omega))\, dz^\rho \\ &+ \int_0^t \sum_{\rho\sigma} G^i_{\rho\sigma}(s, x(s, \omega))\, dz^\rho\, dz^\sigma \quad \text{a.e.} \end{aligned}$$

The integrals here are McShane's belated integrals ([4], [5] and [6]).

* Received by the editors March 29, 1974, and in revised form October 17, 1975.

† Seattle, Washington 98112.

If $x(t, \omega)$ is the trajectory corresponding to a control function $u(t)$ with initial distribution $y(\omega)$, we define the cost for $u(t)$ as follows:

$$\text{(II)} \qquad C(u, y, \omega) = \int_0^T f^0(s, x(s, \omega), u(s))\, ds + F^T(x(T, \omega)).$$

Then our problem is to choose a control function u such that $E\{C(u, y, \omega)\}$ is minimized over the set of u and x satisfying (I).

(III) $(\Omega, \mathscr{A}, P)$ is a probability triple, where P is a complete measure; $\{\mathscr{F}_s : s \in [0, T]\}$ is a family of σ-subalgebras of $\mathscr{A}$ such that if $0 \leqq s \leqq t \leqq T$, then $\mathscr{F}_s \subseteq \mathscr{F}_t$.

(IV) $z^\rho(t, \omega)$, $\rho = 1, \cdots, r$, is a stochastic process on $[0, T]$ such that $z^\rho(t, \cdot)$ is $\mathscr{F}_t$-measurable, $z^\rho(\cdot, \omega)$ is continuous except on a set of measure zero, and there are numbers K_1, K_2, K_4 such that if $0 \leqq s \leqq t \leqq T$, then a.s.

$$|E\{[z^\rho(t, \omega) = z^\rho(s, \omega)]|\mathscr{F}_s\}| \leqq K_1(t - s),$$

$$E\{[z^\rho(t, \omega) - z^\rho(s, \omega)]^c|\mathscr{F}_s\} \leqq K_c(t - s), \qquad c = 2, 4.$$

(V) For any $s \in [0, T]$ and for all ρ,

$$\lim_{\substack{t \to s \\ t \in [0,T]}} \operatorname{ess\,sup} \frac{E\{[z^\rho(t, \omega) - z^\rho(s, \omega)]^c|\mathscr{F}_s\}}{(t - s)^2} = 0$$

uniformly in s ($c = 6, 8$).

(VI) $y^i(\omega)$ is $\mathscr{F}_0$-measurable and has finite second moment.

(VII) f^i, g^i_ρ and $G^i_{\rho\sigma}$ are of class C^2.

(VIII) f^i, g^i_ρ and $G^i_{\rho\sigma}$ have bounded first and second derivatives.

(IX) F^T is bounded; $(\partial/\partial x^i)F^T$ and $(\partial^2/\partial x^i\, \partial x^j)F^T$ exist and are bounded.

These hypotheses are more than sufficient to prove a preliminary result which is of interest on its own, showing Lipschitzian dependence of solutions on initial points.

We also now have sufficient hypotheses and definitions to state the open loop case of our maximum principle.

Stochastic maximum principle (open loop case). Suppose hypotheses (I)–(IX) hold. Then there exist a stochastic process $\mathbf{Z}(t, \omega)$ which is a solution to the system of differential equations

$$\begin{aligned} dZ_m(t, \omega) = {} & -\frac{\partial}{\partial x^m} f^0(t, x(t, \omega), u(t))\, dt \\ & - \sum_\beta Z_\beta(t, \omega) \left\{ \frac{\partial}{\partial x^m} f^\beta(t, x(t, \omega), u(t))\, dt + \sum_\rho \frac{\partial}{\partial x^m} g^\beta_\rho(t, x(t, \omega))\, dz^\rho \right. \\ & \left. + \sum_{\rho\sigma} \left[\frac{\partial}{\partial x^m} G^\beta_{\rho\sigma}(t, x(t, \omega)) \right. \right. \\ & \left. \left. - \sum_{k=1}^n \frac{\partial}{\partial x^k} g^\beta_\sigma(t, x(t, \omega)) \frac{\partial}{\partial x^m} g^k_\rho(t, x(t, \omega)) \right] dz^\rho\, dz^\sigma \right\} \end{aligned} \qquad \text{(M)}$$

with terminal condition

$$Z_m(T, \omega) = \frac{\partial}{\partial x^m} F^T(x(T, \omega)).$$

We define a function H as follows:

$$H(t, x, u, Z) = f^0(t, x, u) + \sum_m Z_m(t) f^m(t, x, u).$$

Then in order for $u(t)$ to be an optimal control function, it is necessary that if $x(t, \omega)$ is the trajectory corresponding to $u(t)$, then for each $t \in [0, T]$ $u(t)$ is a point of Q that minimizes

$$E\{H(t, x(t, \omega)), \cdot, Z(t, \omega)\} \quad \text{over } Q.$$

It should be remarked that if Z were originally defined in this manner, we would be faced with the necessity of proving the existence of a solution for a system of equations with the complicating feature of having terminal rather than initial conditions. As it happens, however, we actually define Z by a messier but less dangerous tactic; then observe that it does satisfy (M).

The closed loop case of the theorem can be stated in almost identical form, but the similarity in form hides a considerable change in surrounding circumstance and a fair amount of new machinery. The fundamental change is that we now assume that at a finite number of times $t_1, \cdots, t_{p-1}$ in the interval $[0, T]$, we find out the exact value of the state variables. Clearly, we then wish to be able to adjust our controls so as to make use of the information as we acquire it. To this end, we replace our control function $u(t)$ by a *control program* U, which is a function $U(t, w_1, \cdots, w_{p-1})$ with values in Q, where $t \in [0, T]$, $w_i \in \mathbb{R}^n$, with the following property:

If $t < t_j$, then for any $w_j, \cdots, w_{p-1}$; $w'_j, \cdots, w'_{p-1}$,

$$U(t, w_1, \cdots, w_{p-1}) = U(t, w_1, \cdots, w_{j-1}, w'_j, \cdots, w'_{p-1}).$$

If this U is the control program, and the states of the system at times $t_1, \cdots, t_{p-1}$ are $x(t_1), \cdots, x(t_{p-1})$ respectively, the control to be used at time t is $U(t, x(t_1), \cdots, x(t_{p-1}))$. If $t_j \leqq t \leqq t_{j+1}$, the states $x(t_1), \cdots, x(t_j)$ are known, and U is independent of the others, so the control is determined. It should be observed that this choice depends only on the available data $x(t_1), \cdots, x(t_j)$, and not on the values of $z^\rho(s)$ $(0 \leqq s \leqq t)$, which are usually unavailable. Since on (t_j, t_{j+1}) U depends only on t, it can (and will) be referred to as $u(t)$. In order to be able to work freely within each interval, we also need to require that the increments of $z^\rho(t, \omega)$ on $[t_j, t_{j+1}]$ be independent of those on $[t_0, t_j]$.

Our system for making the open loop case of the maximum principle applicable on the interval $[t_{j-1}, t_j]$ consists of a device for absorbing the information to be received into a penalty function $F^t j$ for that particular interval. This device requires the following changes:

We replace hypothesis (IX) by

(IX′) F^T is Lipschitzian.

We also add the following.

(X) For any subinterval $[a, b]$ of $[0, T]$, any control function and any initial distribution at a, if $x(t, \omega)$ is a solution of (I) on $[a, b]$, then the distribution of $x(b, \omega)$ is absolutely continuous with respect to Lebesgue measure; i.e., if $N \in \mathbb{R}^n$ and N has Lebesgue measure zero, then

$$P(x(b, \omega) \in N) = 0.$$

Under these hypotheses we can define multipliers Z_m on each interval $[t_{j-1}, t_j]$ which again turn out to be the solutions to the system (M), this time with terminal condition

$$Z_m(t_j, \omega) = \frac{\partial^+}{\partial x^m} F^t j(x(t_j, \omega)),$$

where $\partial^+/\partial x^m$ denotes the upper right derivative. Then the statement of the maximum principle for the closed loop case, including the definition of the function H, is identical to that for the open loop case. Note, however, that the values on (t_{j-1}, t_j) of u, and hence x and hence Z are undefined unless $x(t_1), \cdots, x(t_{j-1})$ are known.

An interesting feature of our maximum principle is that in practice, for a given problem, it is frequently possible to formulate (I) in many different ways, yielding apparently different, but actually identical solutions. For instance, in a problem involving Brownian motion, the fact that $(dz)^2 = dt$ leads to a variety of choices for f and $G_{\rho\sigma}$. If one wishes to eliminate this ambiguity, a possible solution is to put the problem in McShane's "canonical form" [8, Chap. 3, § 3]. In this form, the special case of our theorem involving Brownian motion coincides with the finite-dimensional case of the maximum principle for Banach space recently published by Kuo [2].

Results related to this maximum principle have been published by Fleming [1] and Kushner [3]. In both cases, the statement of the maximum principle is extremely similar to ours—the difference is in the contexts. Fleming is working with solutions of an equation of the form

$$X(t) = X_0 + \int_0^t f(r, X(r), Y(r, X))\, dr + \int_0^t \sigma(r, X(r))\, dw(r),$$

where w is Brownian motion; his Lagrange multipliers are arrived at through partial differential equations, and it is not clear how to relate them to ours. Kushner works with solutions of an equation of the form

$$dx(\omega, t) = f(x(\omega, t), u(\omega, t))\, dt + dz(\omega, t).$$

His maximum principle comes out in terms of expectation conditioned on the minimum σ-field with respect to which $u(\omega, t)$ is measurable. The special case of his theorem where u does not depend on ω coincides with the special case of our theorem where $g_\rho^i \equiv 1$ and $G_{\rho\sigma}^i \equiv 0$.

2. Lemmas. Rather than launching ourselves directly into a series of proofs of lemmas of varying degrees of intrinsic interest, we will now give statements alone of the necessary lemmas, relegating their proofs to the final section.

THEOREM 1 (Lipschitzian dependence on initial value). *There is a constant K such that for any two solutions $x(t, \omega)$ and $\tilde{x}(t, \omega)$ of the system of equations*

$$x^i(t, \omega) = y^i(\omega) + \int_0^t f^i(s, x(s, \omega))\, ds + \int_0^t \sum_\rho g^i_\rho(s, x(s, \omega))\, dz^\rho + \int_0^t \sum_{\rho\sigma} G^i_{\rho\sigma}(s, x(s, \omega))\, dz^\rho\, dz^\sigma$$

with initial distributions, respectively, $y(\omega)$ and $\tilde{y}(\omega)$,

$$\|x(t, \omega) - \tilde{x}(t, \omega)\| \leqq K\|y(\omega) - \tilde{y}(\omega)\|.$$

LEMMA 1. *Assume hypotheses* (III)–(V) *hold. Let $x(t, \omega)$ and $\tilde{x}(t, \omega)$ be n-vector-valued processes adapted to $\mathscr{F}$. Suppose $g^i_{\rho\alpha}$ and $G^i_{\rho\sigma\alpha}$ are functions: $[0, T] \times \mathbb{R}^n \to \mathbb{R}$ $(i, \alpha = 1, \cdots, n;\ \rho, \sigma = 1, \cdots, r)$, which are continuous and bounded in all variables for all i, α, ρ, σ. Then there is a constant K such that if $h(t, \omega)$ is a solution to*

$$h^i(t, \omega) = y^i(\omega) + \int_0^t \sum_{\alpha\rho} \int_0^1 g^i_{\rho\alpha}(s, x(s, \omega) + \theta[\tilde{x}(s, \omega) - x(s, \omega)])\, d\theta h^\alpha(s, \omega)\, dz^\rho + \int_0^t \sum_{\alpha\beta\sigma} \int_0^1 G^i_{\rho\sigma\alpha}(s, x(s, \omega) + \theta[\tilde{x}(s, \omega) - x(s, \omega)])\, d\theta h^\alpha(s, \omega)\, dz^\rho\, dz^\sigma,$$

then $\|h(t, \omega)\| < K\|y(\omega)\|$ for all $t \in [0, T]$.

LEMMA 2.[1] *If X is continuous in L_2-norm and almost all of its sample functions are continuous, and if $\Omega[N] = \{\omega \in \Omega : \sup_{0 \leqq t \leqq T} |X(t)| \geqq N\}$, then as $N \to \infty$, $\|X(t)1_{\Omega[N]}\|$ converges uniformly to 0, where $1_{\Omega[N]}$ is the characteristic for $\Omega[N]$.*

LEMMA 3. *Assume hypotheses* (III)–(VI) *and* (VIII) *all hold. Assume $y(\omega)$ has finite second moment and is $\mathscr{F}_a$-measurable. Let $x_0(t, \omega)$, $x_\varepsilon(t, \omega)$ be solutions for* (I) *with the same control function $u_0(t)$ and with initial values, respectively, $y_0(\omega)$ and $y_0(\omega) + \varepsilon y(\omega)$. Let $q(t|\varepsilon) = (1/\varepsilon)[x_\varepsilon(t, \omega) - x_0(t, \omega)]$. Let $X(t, \omega)$ be a solution to*

$$X^i(t, \omega) = y^i(\omega) + \int_0^t \sum_\alpha \frac{\partial}{\partial x^\alpha} f^i(s, x_0(s, \omega), u_0(s)) X^\alpha(s, \omega)\, ds + \int_0^t \sum_{\rho\alpha} \frac{\partial}{\partial x^\alpha} g^i_\rho(s, x_0(s, \omega)) X^\alpha(s, \omega)\, dz^\rho + \int_0^t \sum_{\rho\sigma\alpha} \frac{\partial}{\partial x_\alpha} G^i_{\rho\sigma}(s, x_0(s, \omega)) X^\alpha(s, \omega)\, dz^\rho\, dz^\sigma.$$

Then $\lim_{\varepsilon \to 0} \|q(t|\varepsilon) - X(t)\| = 0$ uniformly in t.

[1] This result, which simplifies drastically the proofs of several of the lemmas from my dissertation, was communicated to me by E. J. McShane.

LEMMA 4. *Assume hypotheses* (III)–(VIII) *hold. Suppose that on the interval* $[t_0 - \varepsilon, t_0]$, x_ε *and* x_0 *are solutions of*

$$x_\varepsilon^i(t,\omega) = y^i(\omega) + \int_{t_0-\varepsilon}^t f^i(s, x_\varepsilon(s,\omega), \bar{u})\,ds + \int_{t_0-\varepsilon}^t \sum_\rho g_\rho^i(s, x_\varepsilon(s,\omega))\,dz^\rho + \int_{t_0-\varepsilon}^t \sum_{\rho\sigma} G_{\rho\sigma}^i(s, x_\varepsilon(s,\omega))\,dz^\rho\,dz^\sigma,$$

$$x_0^i(t,\omega) = y^i(\omega) + \int_{t_0-\varepsilon}^t f^i(s, x_0(s,\omega), u_0(s))\,ds + \int_{t_0-\varepsilon}^t \sum_\rho g_\rho^i(s, x_0(s,\omega))\,dz^\rho + \int_{t_0-\varepsilon}^t \sum_{\rho\sigma} G_{\rho\sigma}^i(s, x_0(s,\omega))\,dz^\rho\,dz^\sigma.$$

Then

$$\| x_\varepsilon^i(t_0,\omega) - x_0^i(t_0,\omega) - \varepsilon[f^i(t_0, x_0(t_0,\omega), \bar{u}) - f^i(t_0, x_0(t_0,\omega), u_0(t_0))] \| = o(\varepsilon).$$

3. Open loop case. We now proceed directly to the proof of the open loop case of the maximum principle.

Our objective is to find a necessary condition for a given control function $u(t)$ to minimize the expected cost. Clearly one thing that must be true is that any control function that differs from u_0 by only a "small amount" must give an expected cost greater than that given by u_0. Thus in particular, if we define u_ε for $\varepsilon > 0$ as follows: for a given ε, $\bar{u} \in Q$, and t_0 in the interval (ε, T),

$$u_\varepsilon(t) = \begin{cases} u_0(t), & 0 \leqq t < t_0 - \varepsilon, \\ \bar{u}, & t_0 - \varepsilon \leqq t < t_0, \\ u_0(t), & t_0 \leqq t \leqq T, \end{cases}$$

and if $y(\omega)$ is a given starting distribution, then it must be true that for small ε,

$$E\{C(u_\varepsilon, y, \omega)\} \geqq E\{C(u_0, y, \omega)\}.$$

Hence, assuming the expression makes sense (which we must prove), we must have

$$\frac{d}{d\varepsilon}(E\{C(u_\varepsilon, \omega)\})|_{\varepsilon=0} \geqq 0.$$

To prove that the expression does make sense, we must first show that u_ε gives a solvable trajectory equation, i.e., that for a given initial distribution $y(\omega)$, there is a solution for

$$x_\varepsilon^i(t,\omega) = y^i(\omega) + \int_0^t f^i(s, x_\varepsilon(s,\omega), u_\varepsilon(s))\,ds + \int_0^t \sum_\rho g_\rho^i(s, x_\varepsilon(s,\omega))\,dz^\rho + \int_0^t \sum_{\rho\sigma} G_{\rho\sigma}^i(s, x_\varepsilon(s,\omega))\,dz^\rho\,dz^\sigma.$$

On $[0, t_0 - \varepsilon)$, $x_\varepsilon = x_0$. On $[t_0 - \varepsilon, t_0)$, the initial condition $x_0^i(t_0 - \varepsilon)$ is determined from the integral on $[0, t_0 - \varepsilon)$. Hence by Theorem 2 of [9] $x_0^i(t_0 - \varepsilon)$ has finite second moment; since it is a limit of Cauchy polygons, $x_0^i(t_0 - \varepsilon)$ is $\mathscr{F}_{t_0-\varepsilon}$-measurable. Using hypotheses (III) and (IV), we deduce that the equation on $[t_0 - \varepsilon, t_0]$ has a solution (cf. [5]). On $[t_0, T]$, the same argument applies.

Now if x is a trajectory corresponding to u and an initial distribution y, then

$$C(u, y, \omega) = \int_0^T f^0(s, x(s, \omega), u(s))\, ds + F^T(x(T, \omega)),$$

so

$$\begin{aligned}
&\frac{d}{d\varepsilon}(E\{C(u_\varepsilon, y, \omega)\})|_{\varepsilon=0} \\
&\quad = \lim_{\varepsilon\to 0} E\left\{\frac{1}{\varepsilon}\int_0^T f^0(s, x_\varepsilon(s, \omega), u_\varepsilon(s)) - f^0(s, x_0(s, \omega)u_0(s))\, ds\right\} \\
&\qquad + \lim_{\varepsilon\to 0} E\left(\frac{1}{\varepsilon}[F^T(x_\varepsilon(T, \omega)) - F^T(x_0(T, \omega))]\right),
\end{aligned}$$

provided this limit exists. We prove that the limit does exist by chipping away at it industriously:

$$\begin{aligned}
&\frac{1}{\varepsilon}\int_0^T f^0(s, x_\varepsilon(s), u_\varepsilon(s)) - f^0(s, x_0(s), u_0(s))\, ds \\
&\quad = \frac{1}{\varepsilon}\int_{t_0-\varepsilon}^{t_0} f^0(s, x_\varepsilon(s), u_\varepsilon(s)) - f^0(s, x_0(s), u_0(s))\, ds \\
&\qquad + \frac{1}{\varepsilon}\int_{t_0}^T f^0(s, x_\varepsilon(s), u_\varepsilon(s)) - f^0(s, x_0(s), u_0(s))\, ds.
\end{aligned}$$

By Lemma 4,

$$\begin{aligned}
&\left\| \int_{t_0-\varepsilon}^{t_0} [f^0(s, x_\varepsilon(s), u_\varepsilon(s)) - f^0(s, x_0(s), u_0(s))]\, ds \right. \\
&\qquad \left. - \varepsilon[f^0(t_0, x_0(t_0), \bar{u}) - f^0(t_0, x_0(t_0), u_0(t_0))] \right\| = o(\varepsilon).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\lim_{\varepsilon\to 0} E\left\{\frac{1}{\varepsilon}\int_{t_0-\varepsilon}^{t_0} [f^0(s, x_\varepsilon(s), u_\varepsilon(s)) - f^0(s, x_0(s), u_0(s))]\, ds\right\} \\
&\qquad = E\{f^0(t_0, x_0(t_0), \bar{u}) - f^0(t_0, x_0(t_0), u_0(t_0))\}.
\end{aligned}$$

To take care of the situation on $[t_0, T]$, we first observe that by Taylor's theorem,

$$f^0(s, x_\varepsilon(s), u_0(s)) = f^0(s, x_0(s), u_0(s)) + \sum_{\alpha=1}^{n} \frac{\partial}{\partial x^\alpha} f^0(s, x_0(s), u_0(s))(x_\varepsilon^\alpha(s) - x_0^\alpha(s)) + \frac{1}{2} \sum_{\alpha,\rho=1}^{n} \frac{\partial^2}{\partial x^\alpha \partial x^\rho} f^0(s, x^*(s), u_0(s))(x_\varepsilon^\alpha(s) - x_0^\alpha(s))(x_\varepsilon^\rho(s) - x_0^\rho(s)),$$

where x^* is between x_0 and x_ε. Hence if $q^\alpha(s|\varepsilon)$ and $X^\alpha(s)$ are as defined in Lemma 3,

$$\begin{aligned} &E\left\{\frac{1}{\varepsilon}\int_{t_0}^{T} f^0(s, x_\varepsilon(s), u_0(s)) - f^0(s, x_0(s), u_0(s))\, ds\right\} \\ &\quad = E\left\{\int_{t_0}^{T} \sum_\alpha \frac{\partial}{\partial x^\alpha} f^0(s, x_0(s), u_0(s)) q^\alpha(s|\varepsilon)\, ds \right. \\ &\qquad \left. + \frac{\varepsilon}{2}\int_{t_0}^{T} \sum_{\alpha\rho} \frac{\partial^2}{\partial x^\alpha \partial x^\beta} f^0(s, x^*(s), u_0(s)) q^\alpha(s|\varepsilon) q^\beta(s|\varepsilon)\, ds\right\} \\ &\quad = E\left\{\int_{t_0}^{T} \sum_\alpha \frac{\partial}{\partial x^\alpha} f^0(s, x_0(s), u_0(s)) X^\alpha(s)\, ds\right\} \\ &\qquad + E\left\{\int_{t_0}^{T} \sum_\alpha \frac{\partial}{\partial x^\alpha} f^0(s, x_0(s), u_0(s))(q^\alpha(s|\varepsilon) - X^\alpha(s))\, ds\right\} \\ &\qquad + E\left\{\frac{\varepsilon}{2}\int_{t_0}^{T} \sum_{\alpha\beta} \frac{\partial^2}{\partial x^\alpha \partial x^\beta} f^0(s, x^*(s), u_0(s)) q^\alpha(s|\varepsilon) q^\beta(s|\varepsilon)\, ds\right\}. \end{aligned}$$

We will show that the second and third terms of this expression go to zero as ε goes to zero.

Second term: Since the integrands for $q^\alpha(t|\varepsilon)$ and $X^\alpha(t)$ are linear in q and X, respectively, they satisfy hypotheses H3 and H4 of Theorem 3 in [5] trivially. Hence from that theorem, we may deduce that $q^\alpha(t|\varepsilon)$ and $X^\alpha(t)$ are processes adapted to $\mathcal{F}$. Since f was assumed to be continuous, this means we may apply Fubini's theorem to deduce that

$$\begin{aligned} &E\left\{\int_{t_0}^{T} \sum_\alpha \frac{\partial}{\partial x^\alpha} f^0(s, x_0(s), u_0(s))(q^\alpha(s|\varepsilon) - X^\alpha(s))\, ds\right\} \\ &\quad = \int_{t_0}^{T} E\left\{\sum_\alpha \frac{\partial}{\partial x^\alpha} f^0(s, x_0(s), u_0(s))(q^\alpha(s|\varepsilon) - X^\alpha(s))\right\} ds \\ &\quad \leqq \int_{t_0}^{T} \sum_\alpha \left\| \frac{\partial}{\partial x^\alpha} f^0(s, x_0(s), u_0(s)) \right\| \| q^\alpha(s|\varepsilon) - X^\alpha(s) \|\, ds. \end{aligned}$$

But by hypothesis (VIII), $\partial/\partial x^\alpha f^0$ is bounded in all variables; hence $\|(\partial/\partial x^\alpha) f^0(s, x_0(s), u_0(s))\|$ is bounded. By Lemma 3,

$$\lim_{\varepsilon \to 0} \| q^\alpha(s|\varepsilon) - X^\alpha(s) \| = 0 \quad \text{uniformly in } s.$$

From these facts we conclude that the second term goes to 0.

Third term:

$$E\left\{\frac{\varepsilon}{2}\int_{t_0}^{T}\sum_{\alpha,\beta}\frac{\partial^2}{\partial x^\alpha\,\partial x^\beta}f^0(s,\,x^*(s),\,u_0(s))q^\alpha(s|\varepsilon)q^\beta(s|\varepsilon)\,ds\right\}$$
$$=\frac{\varepsilon}{2}\sum_{\alpha,\beta}\int_{t_0}^{T}\left[E\left\{\frac{\partial^2}{\partial x^\alpha\,\partial x^\beta}f^0(s,\,x^*(s),\,u_0(s))\right\}q^\alpha(s|\varepsilon)q^\beta(s|\varepsilon)\right]ds.$$

If B is the bound on $|(\partial^2/\partial x^\alpha\,\partial x^\beta)f^0|$,

$$\sum_{\alpha,\beta}E\left(\left|\left[\frac{\partial^2}{\partial x^\alpha\,\partial x^\beta}f^0(s,x^*(s),u_0(s))\right]^2 q^\alpha(s|\varepsilon)q^\beta(s|\varepsilon)\right|\right)$$
$$\leqq\sum_{\alpha,\beta}B\cdot E(|q^\alpha(s|\varepsilon)|\cdot|q^\beta(s|\varepsilon)|)$$
$$\leqq\sum_{\alpha,\beta}B\|q^\alpha(s|\varepsilon)\|\,\|q^\beta(s|\varepsilon)\|.$$

By Theorem 1, $\|q^\alpha(s|\varepsilon)\|$ is bounded on $[t_0,\,T]$; say $\|q^\alpha(s|\varepsilon)\|\leqq C$ on $[t_0,\,T]$ for all α. Then we have

$$E\left\{\frac{\varepsilon}{2}\int_{t_0}^{T}\sum_{\alpha,\beta}\left|\left[\frac{\partial^2}{\partial x^\alpha\,\partial x^\beta}f^0(s,\,x^*(s),\,u_0(s))\right]^2 q^\alpha(s|\varepsilon)q^\beta(s|\varepsilon)\right|ds\right\}\leqq n^2BC^2T\left(\frac{\varepsilon}{2}\right).$$

So the third term goes to zero with ε. It follows that

$$\lim_{\varepsilon\to0}E\left\{\frac{1}{\varepsilon}\int_{t_0}^{T}[f^0(s,\,x_\varepsilon(s),\,u_0(s))-f^0(s,\,x_0(s),\,u_0(s))]\,ds\right\}$$
$$=E\left\{\int_{t_0}^{T}\sum_{\alpha}\frac{\partial}{\partial x^\alpha}f^0(s,\,x_0(s),\,u_0(s))X^\alpha(s)\,ds\right\}.$$

This takes care of the first term of the expression for $(d/d\varepsilon)[E(C(u_\varepsilon))]$. Identical treatment of the second term yields that

$$\lim_{\varepsilon\to0}E\left\{\frac{1}{\varepsilon}[F^T(x_\varepsilon(T))-F^T(x_0(T))]\right\}=\sum_{\alpha}\frac{\partial}{\partial x^\alpha}F^T(x_0(T))X^\alpha(T).$$

Making a massive collection of terms from the past several pages, we find that we have proved that

$$\lim_{\varepsilon\to0}\frac{1}{\varepsilon}[E(C(u_\varepsilon))-E(C(u_0))]$$
$$=E\{f^0(t_0,\,x_0(t_0),\,\bar{u})-f^0(t_0,\,x_0(t_0),\,u_0(t_0))\}$$
$$+E\left\{\int_{t_0}^{T}\sum_{\alpha}\frac{\partial}{\partial x^\alpha}f^0(s,\,x_0(s),\,u_0(s))X^\alpha(s)\,ds\right\}+E\left\{\sum_{\alpha}\frac{\partial}{\partial x^\alpha}F^T(x_0(T))X^\alpha(T)\right\}.$$

So we have proved that $(d/d\varepsilon)[E(C(u_\varepsilon))]|_{\varepsilon=0}$ exists. As previously remarked, it must be nonnegative if u_0 is to minimize the expected cost. This gives us our first form

of minimum condition: if u_0 is to provide a minimum for the function $E\{C(u)\}$, it must be true that for all $t_0 \in [0, T]$ and all $\bar{u} \in Q$,

$$E\Big\{f^0(t_0, x_0(t_0), \bar{u}) - f^0(t_0, x_0(t_0), u_0(t_0)) + \int_{t_0}^{T} \sum_{\alpha} \frac{\partial}{\partial x^\alpha} f^0(s, x_0(s), u_0(s)) X^\alpha(s)\, ds + \sum_{\alpha} \frac{\partial}{\partial x^\alpha} F^T(x_0(T)) X^\alpha(T)\Big\} \geqq 0.$$

Unfortunately, this is not a very useful form of condition. So we maneuver it around a bit to get a more helpful object. To aid these maneuvers, we use the following definition and theorem from [9].

DEFINITION. Let $x(t, \omega)$ and $y(t, \omega)$ be n-dimensional stochastic processes on $[0, T]$. Then x and y are said to be *adjoints* if

$$\sum_{i=1}^{n} x^i(t, \omega) y^i(t, \omega) = \sum_{i=1}^{n} x^i(0, \omega) y^i(0, \omega) \quad \text{a.s.}$$

ADJOINT THEOREM. *Assume that hypotheses* (III)–(V) *hold, that* $x(0, \omega)$ *and* $y(0, \omega)$ *are bounded for a.a.* ω *and that* $A^i_{h\rho}(t, \omega)$ *and* $B^i_{h\rho\sigma}(t, \omega)$ $(i, h = 1, \cdots, n;$ $\rho, \sigma = 1, \cdots, r)$ *are processes adapted to* $\mathcal{F}$ *which are bounded on* $[0, T]$ *and have a.s. continuous sample paths. Suppose* $x(t, \omega)$ *is a solution of the following system of equations*:

$$x^i(t, \omega) = x^i(0, \omega) + \int_0^t \sum_{h,\rho} A^i_{h\rho}(s, \omega) x^h(s, \omega)\, dz^\rho + \int_0^t \sum_{h,\rho,\sigma} B^i_{h\rho\sigma}(s, \omega) x^h(s, \omega)\, dz^\rho\, dz^\sigma.$$

Then there exists a stochastic process adjoint to x.

Consider the following system of equations on $[0, T]$:

$$\begin{aligned} dW^i(t) &= \sum_{\alpha} \frac{\partial}{\partial x^\alpha} f^i(t, x(t), u(t)) W^\alpha(t)\, dt \\ &\quad + \sum_{\alpha,\rho} \frac{\partial}{\partial x^\alpha} g^i_\rho(t, x(t)) W^\alpha(t)\, dz^\rho \\ &\quad + \sum_{\alpha,\rho,\sigma} \frac{\partial}{\partial x^\alpha} G^i_{\rho\sigma}(t, x(t)) W^\alpha(t)\, dz^\rho\, dz^\sigma. \end{aligned} \tag{1}$$

Let $X^i_1(t), \cdots, X^i_n(t)$ be n linearly independent solutions to (1) with the property that $X^i_j(0) = \delta^i_j$. Let $Y^i_1(t), \cdots, Y^i_n(t)$ be n linearly independent solutions to the system adjoint to (1) on $[0, T]$, with $Y^i_j(0) = \delta^i_j$, whose existence is guaranteed by the Adjoint theorem. By choice of $X(t)$, $X^i(t)$ satisfies (1) on $[t_0, T]$. Hence there exist random variables $C_1(\omega), \cdots, C_n(\omega)$ such that

$$X^h(t, \omega) = \sum_k C_k(\omega) X^h_k(t, \omega) \quad \text{a.s.}$$

Setting $t = t_0$, we get

$$\begin{aligned} \sum_h X^h(t_0, \omega) Y^h_j(t_0, \omega) &= \sum_{h,k} C_k(\omega) X^h_k(t_0, \omega) Y^h_j(t_0, \omega) \\ &= \sum_k C_k(\omega) \delta^k_j \quad \text{by Theorem 3 of [9]} \\ &= C_j(\omega). \end{aligned}$$

Therefore, we have

$$X^h(t, \omega) = \sum_{k,m} X^m(t_0, \omega) Y_k^m(t_0, \omega) X_k^h(t, \omega).$$

Substituting this into our minimum condition, we get a second form of necessary condition:

$$E\Big\{ f^0(t_0, x_0(t_0, \omega), \bar{u}) - f^0(t_0, x_0(t_0, \omega), u_0(t_0)) + \int_{t_0}^{T} \sum_{\alpha,m} \frac{\partial}{\partial x^\alpha} f^0(s, x_0(s, \omega), u_0(s)) X^m(t_0, \omega) Y_k^m(t_0, \omega) X_k^\alpha(s, \omega)\, ds + \sum_{\alpha,m} \frac{\partial}{\partial x^\alpha} F^T(x_0(T, \omega)) X^m(t_0, \omega) Y_k^m(t_0, \omega) X_k^\alpha(T, \omega) \Big\} \geqq 0.$$

In Lemma 4, we proved that

$$\lim_{\varepsilon \to 0} \left\| \frac{1}{\varepsilon} [x_\varepsilon^i(t_0, \omega) - x_0^i(t_0, \omega)] - [f^i(t_0, x_0(t_0, \omega), \bar{u}) - f^i(t_0, x_0(t, \omega), u_0(t_0))] \right\| = 0$$

But

$$\lim_{\varepsilon \to 0} \left\| \frac{1}{\varepsilon} [x_\varepsilon^i(t_0, \omega) - x_0^i(t_0, \omega)] - X^i(t_0, \omega) \right\| = \lim_{\varepsilon \to 0} \| q^i(t_0 | \varepsilon) - X^i(t_0, \omega) \| = 0.$$

Hence

$$\| X^i(t_0, \omega) - [f^i(t_0, x_0(t_0, \omega), \bar{u}) - f^i(t_0, x_0(t_0, \omega), u_0(t_0))] \| = 0.$$

We use this to replace the $X^m(t_0, \omega)$'s in our second form of minimum condition and achieve a mess. To write this mess more appealingly, we make the following definitions: For $t_0 \in [0, T]$,

$$Z_m(t_0, \omega) = \int_{t_0}^{T} \sum_{\alpha,k} \frac{\partial}{\partial x^\alpha} f^0(s, x_0(s, \omega), u_0(s)) Y_k^m(t_0, \omega) X_k^\alpha(s, \omega)\, ds + \sum_{\alpha,k} \frac{\partial}{\partial x^\alpha} F^T(x_0(T, \omega)) Y_k^m(t_0, \omega) X_k^\alpha(T, \omega),$$

$$H(t, x, \bar{u}, Z) = f^0(t, x, \bar{u}) + \sum_m Z_m f^m(t, x, \bar{u}).$$

Then our condition reads:

In order for a control function $u_0(t)$ to be optimal it is necessary that for all $t \in [0, T]$ and all $\bar{u} \in Q$,

$$E\{H(t, x_0(t, \omega), \bar{u}, Z(t, \omega)) - H(t, x_0(t, \omega), u_0(t), Z(t, \omega))\} \geqq 0;$$

i.e., for each $t \in [0, T]$, $u_0(t)$ is the point in Q whose associated control function minimizes

$$E\{H(t, x_0(t, \omega), \bar{u}, Z(t, \omega))\} \quad \text{over the set of } \bar{u} \in Q.$$

For the open loop case, there remains to be proved only that our multipliers Z_m satisfy both the system of equations (M) defined in the statement of the maximum principle in § 1, and the terminal condition that accompanies (M). The second of these facts is immediate from the choice of X_j^i and Y_j^i. The first uses two results from McShane's recently published book [6]. One result (Thms. III-5-1 and III-5-4) is that all integrals of the form

$$\int_a^b f(s, \omega)\, dz^\rho(s, \omega)\, dz^\sigma(s, \omega) \cdots dz^\sigma(s, \omega)$$

with three or more factors $dz^\rho, \cdots, dz^\sigma$ vanish, and so do all integrals with two factors $dz^\rho\, dz^\sigma$ in which $dz^\rho = dt$ or $dz^\sigma = dt$ or both. The other result (Thm. IV-3-7) is the "integration by parts" formula

$$d(uv) = u\, dv + v\, du + du\, dv,$$

in which du and dv are to be replaced by their expressions in terms of $dt, dz^1, \cdots, dz^T$, and in the product $du\, dv$ all terms containing three or more factors dz^ρ and all terms containing a factor dt and another factor dt or dz^ρ are to be discarded. The proof is a computation performed by applying, in order, the second of the above results, the formula for an adjoint system, the definitions of X and Y, the first of the above results and the definition of Z.

4. Closed loop case. In order to tackle the closed loop case of the theorem, we need to elaborate the set-up and modify the hypotheses as described in the Introduction. Since we will want to be able to apply the open loop theorem in this context, our first job is to prove that replacing hypothesis (IX) by hypotheses (IX′) and (X) leaves the open loop theorem intact. Inspection of the proof reveals that the only direct use made of hypothesis (IX) occurred in showing that

$$\lim_{\varepsilon \to 0} E\left\{\frac{1}{\varepsilon}[F^T(x_\varepsilon(t, \omega)) - F^T(x_0(t, \omega))]\right\} = E\left\{\sum_\alpha \frac{\partial}{\partial x^\alpha} F^T(x_0(T, \omega)) X^\alpha(T, \omega)\right\}.$$

To prove this equality in our new context we use the following theorem due to Rademacher [8].

THEOREM. *Every Lipschitzian function on $\mathbb{R}^n$ has a total differential except on a set of Lebesgue measure* 0.

Let N be the subset of $\mathbb{R}^n$ such that N has Lebesgue measure 0 and if $x \notin N$, then F^T has a total differential at x. By hypothesis (X), if $M = \{\omega \in \Omega : x_0(T, \omega) \in N\}$, then $P(M) = 0$.

Let $(\partial^+/\partial x^i)F^T(x(T, \omega))$ be the upper right partial derivative of F^T with respect to the ith coordinate:

$$\frac{\partial^+}{\partial x^i} F^T(x(T, \omega)) = \varlimsup_{\substack{\varepsilon \to 0 \\ \varepsilon > 0}} \frac{1}{\varepsilon}\{F^T(x^1(T, \omega), \cdots, x^i(T, \omega) + \varepsilon, \cdots, x^n(T, \omega)) - F^T(x^1(T, \omega), \cdots, x^i(T, \omega), \cdots, x^n(T, \omega))\}$$

Then consider the expression

$$\left| F^T(x_\varepsilon(T, \omega)) - F^T(x_0(T, \omega)) - \sum_i \frac{\partial^+}{\partial x^i} F^T(x_0(T, \omega))(x_\varepsilon^i(T, \omega) - x_0^i(T, \omega)) \right|,$$

to which we assign the label $\theta(\varepsilon)$. What we need is that

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} E\{\theta(\varepsilon)\} = 0.$$

If we introduce the additional notation $D(\varepsilon) = x_\varepsilon(T, \omega) - x_0(T, \omega)$, then we have

$$\begin{aligned}
\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} E\{\theta(\varepsilon)\} &= \lim_{\varepsilon \to 0} E\left\{ \frac{|D(\varepsilon)|}{\varepsilon} \cdot \frac{\theta(\varepsilon)}{|D(\varepsilon)|} \right\} \\
&\leqq \lim_{\varepsilon \to 0} \left\{ \left\| \frac{D(\varepsilon)}{\varepsilon} \right\| \cdot \left\| \frac{\theta(\varepsilon)}{D(\varepsilon)} \right\| \right\} \\
&\leqq K \lim_{\varepsilon \to 0} \left\{ \left\| \frac{\theta(\varepsilon)}{D(\varepsilon)} \right\| \right\} \text{ by Theorem 1.}
\end{aligned}$$

By Theorem 1, $\lim_{\varepsilon \to 0} \|D(\varepsilon)\| = 0$. It follows that for any sequence $\varepsilon_1, \varepsilon_2, \cdots$, where $\varepsilon_j \to 0$, there is a subsequence $\delta_1, \delta_2, \cdots$ such that $D(\delta_j) \to 0$ a.e. Take such a sequence and subsequence. Let Z be a subset of Ω such that $P(Z) = 0$ and for $\omega \in \Omega - Z$, $|D(\delta_j)| \to 0$. Then for $\omega \in \Omega - (Z \cup M)$

$$\lim_{\theta_j \to 0} \frac{|\theta(\delta_j)|}{|D(\delta_j)|} = 0$$

since F^T has a total differential on $\Omega - M$. So

$$\begin{aligned}
\left\| \lim_{\delta_j \to 0} \left| \frac{\theta(\delta_j)}{D(\delta_j)} \right| \right\| &= \left\{ \int_{\Omega - (M \cup Z)} \lim_{\delta_j \to 0} \left| \frac{\theta(\delta_j)}{D(\delta_j)} \right|^2 P(d\omega) + \int_{M \cup Z} \lim_{\delta_j \to 0} \left| \frac{\theta(\delta_j)}{D(\delta_j)} \right|^2 P(d\omega) \right\}^{1/2} \\
&\leqq \left\{ \int_{\Omega - (M \cup Z)} 0 P(d\omega) + \int_{M \cup Z} 4L_F^2 T P(d\omega) \right\}^{1/2} = 0.
\end{aligned}$$

Hence by the bounded convergence theorem,

$$\lim_{\delta_j \to 0} \left\| \frac{\theta(\delta_j)}{D(\delta_j)} \right\| = 0.$$

But if L_{F^T} is the Lipschitz constant for F^T, then $\|\theta(\varepsilon)/D(\varepsilon)\|$ is bounded by $2L_{F^T}$. Hence it takes values in a closed interval. Thus if there are an infinite number of points ε_i such that $\|\theta(\varepsilon_i)/D(\varepsilon_i)\|$ is outside of some neighborhood of 0, there must be an accumulation point p_0 elsewhere than 0. But then a sequence ε_j with $\|\theta(\varepsilon_j)/D(\varepsilon_j)\| \to p_0$ would have no subsequence such that $\|\theta(\delta_j)/D(\delta_j)\| \to 0$, a contradiction.

It follows that

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} E\{\theta(\varepsilon)\} = 0.$$

So if $q(T|\varepsilon)$ and $X(T, \omega)$ are as defined in the statement of Lemma 3,

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} E\{F^T(\underline{x}_\varepsilon(T, \omega))\} = F^T(\underline{x}_0(T, \omega))\}$$

$$= \lim_{\varepsilon \to 0} E\left\{\sum_i \frac{\partial^+}{\partial x^i} F^T(\underline{x}_0(T, \omega)) q^i(T|\varepsilon)\right\},$$

$$= E\left\{\sum_i \frac{\partial^+}{\partial x^i} F^T(\underline{x}_0(T, \omega)) X^i(T, \omega)\right\}.$$

This last equality follows from Lemma 3 by a standard Schwarz lemma argument. This enables us to proceed as in the proof of the open loop case, provided that for $(\partial/\partial x^\alpha)F^T$, we substitute $(\partial^+/\partial x^\alpha)F^T$.

We are now in a position to apply the open loop case of the maximum principle on each interval (t_{j-1}, t_j) provided that we can supply a penalty function F^{t_j} for each t_j. It will suffice for us to show how to define $F^{t_{p-1}}$ and have it be Lipschitzian. Assume we know $\underline{x}(t_{p-1}, \omega)$ accurately—call it $\underline{x}_{p-1}$. Fix $\underline{u}$. Then the cost $C(\underline{u}, \underline{x}_{p-1}, \omega)$ starting at $\underline{x}_{p-1}$ and using control function $\underline{u}$ is a random variable determined by the equation

$$C(\underline{u}, x_{p-1}, \omega) = \int_{t_{p-1}}^{T} f_0(s, \underline{x}(s, \omega), \underline{u}(s))\, ds + F^T(\underline{x}(T, \omega)).$$

The increments of the z on $[t_{p-1}, T]$ are independent of the increments on $[0, t_{p-1}]$, so this is an open loop problem of the type considred.What we want to take for our penalty function $F^{t_{p-1}}$ is the infimum over all control functions $\underline{u}$ of

$$E\{C(\bar{u}, \underline{x}_{p-1}, \omega)\}.$$

To have this turn out Lipschitzian, we need to know that for each $\underline{u}$,

$$E\{C(\underline{u}, \underline{x}_{p-1}, \omega)\}$$

is Lipschitzian in $\underline{x}$, and that the Lipschitz constant does not depend on $\underline{u}$.

f^0 is differentiable, hence Lipschitzian; call its Lipschitz constant L_{f^0}. Then if $\underline{x}(t, \omega)$ and $\underline{y}(t, \omega)$ are trajectories over $[t_{p-1}, T]$ having starting points $\underline{x}_{p-1}$ and $\underline{y}_{p-1}$, we have

$$\left\| \int_{t_{p-1}}^{T} f^0(s, \underline{x}(s, \omega), \underline{u}(s)) - f^0(s, \underline{y}(s, \omega), \underline{u}(s))\, ds \right\|$$

$$\leqq \int_{t_{p-1}}^{T} \| f^0(s, \underline{x}(s, \omega), \underline{u}(s)) - f^0(s, \underline{y}(s, \omega), \underline{u}(s)) \|\, ds$$

$$\leqq \int_{t_{p-1}}^{T} L_{f^0} \| \underline{x}(s, \omega) - \underline{y}(s, \omega) \|\, ds \quad \text{(independently of } \underline{u})$$

$$\leqq (T - t_{p-1}) \cdot L_{f^0} \cdot K \cdot |\underline{x}_{p-1} - \underline{y}_{p-1}| \quad \text{by Theorem 1.}$$

Let $L = (T - t_{p-1}) \cdot L_{f^0} \cdot K$.

Then

$$\begin{aligned}
&\left| E\int_{t_{p-1}}^{T} f^0(s, \underline{x}(s,\omega), \underline{u})\, ds - E\int_{t_{p-1}}^{T} f^0(s, \underline{y}(s,\omega), \underline{u})\, ds \right| \\
&\quad \leqq E\left| \int_{t_{p-1}}^{T} f^0(s, \underline{x}(s,\omega), \underline{u})\, ds - \int_{t_{p-1}}^{T} f^0(s, \underline{y}(s,\omega), \underline{u})\, ds \right| \\
&\quad \leqq \|1\| \left\| \int_{t_{p-1}}^{T} f^0(s, \underline{x}(s,\omega), \underline{u})\, ds - \int_{t_{p-1}}^{T} f^0(s, \underline{y}(s,\omega), \underline{u})\, ds \right\| \quad \text{(Schwarz)} \\
&\quad \leqq L|\underline{x}_{p-1} - \underline{y}_{p-1}|,
\end{aligned}$$

so the expected running cost does vary with the initial point in a Lipschitzian way, with the Lipschitz constant independent of $\underline{u}$.

For the chunk of $F^{t_{p-1}}$ that comes from F^T, we have

$$\begin{aligned}
E\{|F^T(\underline{x}(T,\omega)) - F^T(\underline{y}(T,\omega))|\} &\leqq E\{|L_{F^T}(\underline{x}(T,\omega) - \underline{y}(T,\omega))|\} \\
&\leqq L_{F^T} E\{|\underline{x}(T,\omega) - \underline{y}(T,\omega)|\} \\
&\leqq L_{F^T}\|1\|\, \|\underline{x}(T,\omega) - \underline{y}(T,\omega)\| \quad \text{(Schwarz)} \\
&\leqq L_{F^T} \cdot K \cdot \|\underline{x}_{p-1} - \underline{y}_{p-1}\| \quad \text{(Theorem 1)} \\
&= L_{F^T} \cdot K \cdot |\underline{x}_{p-1} - \underline{y}_{p-1}|.
\end{aligned}$$

For both cases, we observe that the constant K in Theorem 1 did not depend on the choice of $\underline{u}$. Hence we do have the form of Lipschitzianness desired.

It follows that if we define $F^{t_{p-1}}$ by

$$F^{t_{p-1}}(\underline{x}(t_{p-1})) = \inf_{\underline{u}} E\{C(\underline{u}, \underline{x}(t_{p-1}), \omega)\},$$

then $F^{t_{p-1}}$ is Lipschitzian with Lipschitz constant $L + L_{FT} \cdot K$.

Clearly to define F^{t_j}, we proceed backwards from $F^{t_{p-1}}$ repeating the above process for each interval successively. This gives us a procedure by which a solution to the closed loop problem can be found. It should be pointed out, however, that since the use of each piece of information requires the solution of an open loop problem not merely for one initial value, but for all initial values in $\mathbb{R}^n$, the computational cost of actually solving even a simple problem is likely to be astronomical.

5. Proofs of lemmas. Before proving the lemmas stated in § 2, we need to state a pair of inequalities.

PRELIMINARY LEMMA. *Suppose hypotheses* (I)–(IV) *hold, and suppose* $f(t,\omega)$ *has the property that for all* $t \in [0, T]$ *and all* $f(t,\omega) \in L_2(\Omega, \mathscr{A}, P)$, $\|f(t)\|_2$ *is bounded on* $[0, T]$ *and* f *is a.e.* (*with respect to Lebesgue measure*) *continuous in* L_2*-norm on* $[0, T]$. *Then*

$$\left\| \int_0^t f(s)\, dz^\rho\, dz^\sigma \right\| \leqq K_9 \left[\int_0^t \|f(s)\|^2\, ds \right]^{1/2},$$

$$\left\| \int_0^t f(s)\, dz^\rho \right\| \leqq K_{10} \left[\int_0^t \|f(s)\|^2\, ds \right]^{1/2},$$

where $K_9 = K_4 + 2K_2T$ *and* $K_{10} = 2K_1T^{1/2} + K_2^{1/2}$.

The proof is an immediate application of the Cauchy–Schwarz inequality and of Corollary 4.1 of [4].

Proof of Lemma 1. Since $\underline{x}$ and $\tilde{x}$ are fixed processes,

$$\int_0^1 g^i_{\rho\alpha}(s, \underline{x}(s,\omega) + \theta[\tilde{\underline{x}}(s,\omega) - \underline{x}(s,\omega)])\, d\theta$$

and

$$\int_0^1 G^i_{\rho\sigma\alpha}(s, \underline{x}(s,\omega) + \theta[\tilde{\underline{x}}(s,\omega) - \underline{x}(s,\omega)])\, d\theta$$

are functions of s and ω alone.

Label them, respectively, $A^i_{\rho\alpha}(s,\omega)$ and $B^i_{\rho\sigma\alpha}(s,\omega)$. Since $\underline{x}$ and $\tilde{\underline{x}}$ are adapted to $\mathscr{F}$ and $g^i_{\rho\sigma}$ and $G^i_{\rho\sigma\alpha}$ are continuous, $A^i_{\rho\alpha}$ and $B^i_{\rho\sigma\alpha}$ are adapted to $\mathscr{F}$. They are also bounded, since $g^i_{\rho\alpha}$ and $G^i_{\rho\sigma\alpha}$ are bounded in all variables. Say $|A^i_{\rho\alpha}(s,\omega)| \leqq A$, $|B^i_{\rho\sigma\alpha}(s,\omega)| \leqq B$. Then

$$\begin{aligned}
\|h^i(t)\| &= \left\| y^i + \int_0^t \sum_{\rho\alpha} A^i_{\rho\alpha}(s) h^\alpha(s)\, dz^\rho + \int_0^t \sum_{\rho\sigma\alpha} B^i_{\rho\sigma\alpha}(s) h^\alpha(s)\, dz^\rho\, dz^\sigma \right\| \\
&\leqq \|y^i\| + \left\| \int_0^t \sum_{\rho\alpha} A^i_{\rho\alpha}(s) h^\alpha(s)\, dz^\rho \right\| + \left\| \int_0^t \sum_{\rho\sigma\alpha} B^i_{\rho\sigma\alpha}(s) h^\alpha(s)\, dz^\rho\, dz^\sigma \right\| \\
&\leqq \|y^i\| + \underline{K}\left[\int_0^t \|h(s)\|^2\, ds\right]^{1/2}, \quad \text{where } \underline{K} = rn(K_{10}A + rK_9B).
\end{aligned}$$

Let $\|y^j\| = \max \|y^i\|$. Then we have

$$\|h(t)\| \leqq n\|y^j\| + n\underline{K}\left[\int_0^t \|h(s)\|^2\, ds\right]^{1/2},$$

and hence

$$\|h(t)\|^2 \leqq 2n^2\|y^j\|^2 + 2n^2\underline{K}^2 \int_0^t \|h(s)\|^2\, ds.$$

By a standard argument using Gronwall's lemma (see, e.g., [9]), this gives us that

$$\|h(t)\|^2 \leqq K^2 \max \|y^i\|^2, \quad \text{where } K^2 = 2n^2\, e^{2n^2\underline{K}^2T}.$$

It follows that

$$\begin{aligned}
E\{[h(t,\omega)]^2\} &\leqq K^2 \max E\{[y^i]^2\} \\
&\leqq K^2 E\{\sum_i [y^i]^2\} \\
&\leqq K^2 E\{\underline{y}^2\}.
\end{aligned}$$

Proof of Theorem 1. We first observe that all of the integrals with respect to s can be subsumed into integrals with respect to z by defining $z_\rho^{r+1}(t, \omega)$ to be t for all ω and $g_{r+1}^i(t, \omega)$ to be $f^i(t, x)$. To simplify notation, we make three definitions:

$$\varepsilon = \|y(\omega) - \tilde{y}(\omega)\|,$$

$$q(t|\varepsilon) = \frac{[x(t, \omega) - \tilde{x}(t, \omega)]}{\varepsilon},$$

$$y_0(\omega) = \frac{y(\omega) - \tilde{y}(\omega)}{\|y(\omega) - \tilde{y}(\omega)\|}.$$

Then

$$q(t|\varepsilon) = \frac{1}{\varepsilon}[y(\omega) - \tilde{y}(\omega)] + \frac{1}{\varepsilon}\int_0^t \sum_\rho [g_\rho^i(s, x(s, \omega)) - g_\rho^i(s, \tilde{x}(s, \omega))]\, dz^\rho$$

$$+ \frac{1}{\varepsilon}\int_0^t \sum_{\rho\sigma} [G_{\rho\sigma}^i(s, x(s, \omega)) - G_{\rho\sigma}^i(s, \tilde{x}(s, \omega))]\, dz^\rho\, dz^\sigma.$$

Since

$$g^i(s, x(s, \omega)) - g_\rho^i(s, \tilde{x}(s.\, \omega)) = g_\rho^i(s, x(s, \omega) + \theta[\tilde{x}(s, \omega) - x(s, \omega)])\Big|_0^1$$

$$= \int_0^1 \sum_{\alpha=1}^n \frac{\partial}{\partial x^\alpha} g_\rho^i(s, x(s, \omega) + \theta[\tilde{x}(s, \omega) - x(s, \omega)])$$

$$\cdot (x^\alpha(s, \omega) - \tilde{x}^\alpha(s, \omega))\, d\theta$$

and $G_{\rho\sigma\alpha}^i$ behaves the same, we have

$$q^i(t|\varepsilon) = y_0(\omega) + \int_0^t \sum_{\rho\alpha} \int_0^1 \frac{\partial}{\partial x^\alpha} g_\rho^i(s, x(s, \omega) + \theta[\tilde{x}(s, \omega) - x(s, \omega)])\, d\theta\, q^\alpha(s|\varepsilon)\, dz^\rho$$

$$+ \int_0^t \sum_{\rho\sigma\alpha} \int_0^1 \frac{\partial}{\partial x^\alpha} G_{\rho\sigma\alpha}^i(s, x(s, \omega) + \theta[\tilde{x}(s, \omega) - x(s, \omega)])\, d\theta\, q^\alpha(s|\varepsilon)\, dz^\rho\, dz^\sigma.$$

Since hypothesis (VIII) tells us that g_ρ^i and $G_{\rho\sigma}^i$ have bounded and continuous first partial derivatives, Lemma 1 applies and we have a constant K such that

$$\|q(t|\varepsilon)\| < K\|y_0\| = K.$$

But

$$\|q(t|\varepsilon)\| = \frac{1}{\varepsilon}\{\|x(t, \omega) - \tilde{x}(t, \omega)\|\},$$

so we have

$$\|x(t, \omega) - \tilde{x}(t, \omega)\| < K\|y(\omega) - \tilde{y}(\omega)\|.$$

Proof of Lemma 2. Except on a set of P-measure 0, for each t in $[0, T]$, we have

$$\lim_{s \to t} [|X(s, \omega) - X(t, \omega)|1_{N''}(\omega)]^2 = 0,$$

the quantity in brackets being at most $2N$. By the dominated convergence theorem $X(t)1_{N''}$ is continuous in L_2-norm. So is $X(t)$, and, therefore, so is $X(t)1_{N'} = X(t) - X(t)1_{N''}$. In particular, $\|X(t)1_{N'}\|$ is continuous. For each t, it decreases as N increases. Since $[X(t, \omega)1_{N'}(\omega)]^2 \to 0$ as $N \to \infty$ for all ω except those in the set of P-measure 0 on which $X(\cdot, \omega)$ is discontinuous, by the dominated convergence theorem $X(t)1_{N'} \to 0$ for all t. For each positive ε, the set $\{t \in [0, T] : \|X(t)1_{N'}\| \geqq \varepsilon\}$ is compact. It shrinks as N increases, and no t is in this set for all N. It follows that there is an N_0 for which the set is empty, and for $N \geqq N_0$, we have $\|X(t)1_N\| < \varepsilon$ for all t in $[0, T]$. That is, $\|X(t)1_{\Omega(N)}\| = \|X(t)1_{N'}\|$ converges uniformly to 0.

Proof of Lemma 3. Once again we start by subsuming the *fds* terms. Then we adopt the following notation:

$$v^i(t) = \|q^i(t|\varepsilon) - X^i(t)\|^2,$$

$$v(t) = \|q(t|\varepsilon) - X(t)\|^2,$$

$$h^i_{\rho\alpha}(t) = \int_0^1 \frac{\partial}{\partial x^\alpha} g^i_\rho(t, x_0(t) + \theta[x_\varepsilon(t) - x_0(t)])\, d\theta,$$

$$H^i_{\rho\sigma\alpha}(t) = \int_0^1 \frac{\partial}{\partial x^\alpha} G^i_{\rho\sigma}(t, x_0(t) + \theta[x_\varepsilon(t) - x_0(t)])\, d\theta,$$

$$A^i_\rho(t) = \sum_\alpha h^i_{\rho\alpha}(t)[q^\alpha(t|\varepsilon) - X^\alpha(t)],$$

$$B^i_\rho(t) = \sum_\alpha \left[h^i_{\rho\alpha}(t) - \frac{\partial}{\partial x^\alpha} g^i_\rho(t, x_0(t)) \right] X^\alpha(t),$$

$$C^i_{\rho\sigma}(t) = \sum_\alpha H^i_{\rho\sigma\alpha}(t)[q^\alpha(t|\varepsilon) - X^\alpha(t)],$$

$$D^i_{\rho\sigma}(t) = \sum_\alpha \left[H^i_{\rho\sigma\alpha}(t) - \frac{\partial}{\partial x^\alpha} G^i_{\rho\sigma}(t, x_0(t)) \right] X^\alpha(t).$$

Then

$$\begin{aligned} \|q(t|\varepsilon) - X(t)\|^2 &= \left\| \int_0^t \sum_\rho A_\rho(s)\, dz^\rho + \int_0^t \sum_\rho B_\rho(s)\, dz^\rho \right. \\ &\quad \left. + \int_0^t \sum_\rho C_{\rho\sigma}(s)\, dz^\rho\, dz^\sigma + \int_0^t \sum_{\rho\sigma} D_{\rho\sigma}(s)\, dz^\rho\, dz^\sigma \right\|^2 \\ &\leqq \left[\int_0^t \left\| \sum_\rho A_\rho(s)\, dz^\rho \right\| + \left\| \int_0^t \sum_\rho B_\rho(s)\, dz^\rho \right\| \right. \\ &\quad \left. + \left\| \int_0^t \sum_{\rho\sigma} C_{\rho\sigma}(s)\, dz^\sigma \right\| + \left\| \int_0^t \sum_{\rho\sigma} D_{\rho\sigma}(s)\, dz^\rho\, dz^\sigma \right\| \right]^2 \end{aligned} \tag{2}$$

$$\leqq 4 \max \left\{ \left\| \int_0^t \sum_\rho A_\rho(s)\, dz^\rho \right\|, \left\| \int_0^t \sum_\rho B_\rho(s)\, dz^\rho \right\|, \right.$$
$$\left. \left\| \int_0^t \sum_{\rho\sigma} C_{\rho\sigma}(s)\, dz^\rho\, dz^\sigma \right\|, \left\| \int_0^t \sum_{\rho\sigma} D_{\rho\sigma}(s)\, dz^\rho\, dz^\sigma \right\| \right\}^2$$
$$\leqq 4 \left[\left\| \int_0^t \sum_\rho A_\rho(s)\, dz^\rho \right\|^2 + \left\| \int_0^t \sum_\rho B_\rho(s)\, dz^\rho \right\|^2 \right.$$
$$\left. + \left\| \int_0^t \sum_{\rho\sigma} C_{\rho\sigma}(s)\, dz^\rho\, dz^\sigma \right\|^2 + \left\| \int_0^t \sum_{\rho\sigma} D_{\rho\sigma}(s)\, dz^\rho\, dz^\sigma \right\|^2 \right].$$

We tackle this expression by pieces:

$$\left\| \int_0^t \sum_\rho A^i_\rho(s)\, dz_\rho \right\|^2 = \left\| \int_0^t \sum_{\rho\alpha} h^i_{\rho\alpha}(s)[q^\alpha(s|\varepsilon) - X^\alpha(s)]\, dz^\rho \right\|^2$$
$$< K_{10}^2 \int_0^t \sum_{\alpha\rho} \| h^i_{\rho\alpha}(s)[q^\alpha(s|\varepsilon) - X^\alpha(s)] \|^2\, ds$$
$$= K_{10}^2 \int_0^t \sum_{\rho\alpha} E\{(h^i_{\rho\alpha}(s))^2 (q^\alpha(s|\varepsilon) - X^\alpha(s))^2\}\, ds.$$

Let $M = \sup\{h^i_{\rho\alpha}(t) : t \in [0, T],\ \rho = 1, \cdots, r,\ \alpha = 1, \cdots, n\}$, which is finite because of hypothesis (VIII). By the preceding inequality,

$$\left\| \int_0^t \sum_\rho A^i_\rho(t)\, dz^\rho \right\|^2 \leqq M^2 K_{10}^2 \int_0^t \sum_\alpha E\{q^\alpha(s|\varepsilon) - X^\alpha(s)\}^2\, ds$$
$$= M^2 K_{10}^2 \int_0^t \sum_\alpha \| q^\alpha(s|\varepsilon) - X^\alpha(s) \|^2\, ds$$
$$= M^2 K_{10}^2 \int_0^t v(s)\, ds.$$

Similarly, if we let $\underline{M} = \sup\{H^i_{\rho\sigma\alpha}(t) : t \in [0, T];\ \rho, \sigma = 1, \cdots, r;\ \alpha = 1, \cdots, n\}$, then

$$\left\| \int_0^t \sum_{\rho\sigma} C^i_{\rho\sigma}(t)\, dz^\rho\, dz^\sigma \right\|^2 \leqq K_9^2 \int_0^t \sum_{\alpha\sigma\rho} \left\| H^i_{\rho\sigma\alpha}(s)[q^\alpha(s|\varepsilon) - X^\alpha(s)] \right\|^2 ds$$
$$\leqq K_9^2 \underline{M}^2 \int_0^t v(s)\, ds.$$

Let $C = nK_9^2 \underline{M}^2 + nK_{10}^2 M^2$. Then the preceding inequalities combine to yield

$$\sum_i \left[\left\| \int_0^t \sum_\rho A^i_\rho(s)\, dz^\rho \right\|^2 + \left\| \int_0^t \sum_{\rho\sigma} C^i_{\rho\sigma}(s)\, dz^\rho\, dz^\sigma \right\|^2 \right] \leqq C \int_0^t v(s)\, ds. \tag{3}$$

Next we let C_3 be the bound on $\|q(t|\varepsilon)\|$ whose existence was established in Theorem

1, and B the bound on $|(\partial/\partial x^\alpha)g^i_\rho|$, $|(\partial/\partial x^\alpha)G^i_{\rho\sigma}|$ and their first derivatives ($i, \alpha = 1, \cdots, n, \rho, \sigma = 1, \cdots, r$), whose existence is guaranteed by hypothesis (VIII). Then for each ω

$$\left| h^i_{\rho\alpha}(t) - \frac{\partial}{\partial x^\alpha} g^i_\rho(t, x_0(t)) \right| \leqq 2B.$$

Furthermore, the functions $|(\partial/\partial x^\alpha)g^i_\rho|$ and $|(\partial/\partial x^\alpha)G^i_{\rho\sigma}|$ are Lipschitzian in x for fixed t, with Lipschitz constant $Bn^{1/2}$. Hence for each ω,

$$\begin{aligned}
&\left| h^i_{\rho\alpha}(t) - \frac{\partial}{\partial x^\alpha} g^i_\rho(t, x_0(t)) \right| \\
&\quad \leqq \int_0^1 \left| \frac{\partial}{\partial x^\alpha} g^i_\rho(t, x_0(t) + \theta[x_\varepsilon(t) - x_0(t)]) - \frac{\partial}{\partial x^\alpha} g^i_\rho(t, x_0(t)) \right| d\theta \\
&\quad \leqq \int_0^1 Bn^{1/2} |\theta(x_\varepsilon(t) - x_0(t))| \, d\theta.
\end{aligned}$$

As in the proof of Lemma 2, for each positive N, we define $1_{N'}$ to be

$$\left\{ \omega \in \Omega : \sup_{0 \leqq t \leqq T} |X(t, \omega)| \geqq N \right\}$$

and $1_{N''}$ to be $\Omega - 1_{N'}$. For each α and ρ, we apply the preceding inequality at all ω in $1_{N''}$ and the inequality before that at all ω in $1_{N'}$, and obtain

$$\begin{aligned}
&\left| \left[h^i_{\rho\alpha}(t) - \frac{\partial}{\partial x^\alpha} g^i_\rho(t, x_0(t)) \right] X^\alpha(t) \right|^2 \\
&\qquad \leqq \tfrac{1}{4} B^2 n |x_\varepsilon(t) - x_0(t)|^2 N^2 1_{N''} + 4B^2 |X^\alpha(t)|^2 1_{N'}.
\end{aligned}$$

By Lemma 2, the last term tends uniformly to zero. Hence if we fix $\delta > 0$, we can choose an N for which the last term is less that $\delta/(4K_{10}^2 nr)$ for all t in $[0, T]$. With this N fixed, we then choose an ε' such that if $\varepsilon < \varepsilon'$, then the first term is less than $\delta/(4K_{10}^2 nr)$ for all t; this is possible by Theorem 1. Then for such ε we have

$$\|B^i_\rho(t)\|^2 < \delta/(2nrK_{10}^2) \quad \text{for all } t.$$

By a similar proof, for all sufficiently small ε,

$$\|D^i_{\rho\sigma}(t)\|^2 < \delta/(2nr^2K_9^2).$$

By the two preceding inequalities,

$$\begin{aligned}
&\sum_i \left[\left\| \int_0^t \sum_\rho B^i_\rho(s) \, dz^\rho \right\|^2 + \left\| \int_0^t \sum_{\rho\sigma} D_{\rho\sigma}(s) \, dz^\rho \, dz^\sigma \right\|^2 \right] \\
&\qquad \leqq nrK_{10}^2 \int_0^t \left\| \sup_{i,\rho} B^i_\rho(s) \right\|^2 ds + nr^2 K_9^2 \int_0^t \left\| \sup_{i,\rho,\sigma} D^i_{\rho\sigma}(s) \right\|^2 ds < \delta t.
\end{aligned}$$

This and inequalities (2) and (3) and the definition of v imply

$$v(t) \leqq C \int_0^t v(s) \, ds + \delta t.$$

To this we would like to apply Gronwall's lemma. So we consider the equation

$$w(t) = C\int_0^t w(s)\,ds + \delta t.$$

This equation has the solution

$$w(t) = (\delta/C)[e^{Ct} - 1].$$

It follows by Gronwall's lemma that

$$v(t) \leqq (1/C)[e^{CT} - 1]\delta \quad \text{for all } t \in [0, T].$$

So $\lim_{\varepsilon\to 0} \|q(t|\varepsilon) - X(t)\|^2 = 0$ uniformly in t.

Proof of Lemma 4. Let $v_\varepsilon = x_\varepsilon - x_0$. Then

$$v_\varepsilon^i(t) = \int_{t_0-\varepsilon}^t [f^i(s, x_0(s) + v_\varepsilon(s), \bar{u}) - f^i(s, x_0(s), u_0(s))]\,ds$$

$$+ \int_{t_0-\varepsilon}^t \sum_\rho [g^i(s, x_0(s) + v_\varepsilon(s)) - g^i_\rho(s, x_0(s))]\,dz^\rho$$

$$+ \int_{t_0-\varepsilon}^t \sum_{\rho\sigma} [G^i_{\rho\sigma}(s, x_0(s) + v_\varepsilon(s)) - G^i_{\rho\sigma}(s, x_0(s))]\,dz^\rho\,dz^\sigma$$

(A) $$= \int_{t_0-\varepsilon}^t [f^i(s, x_0(s) + v_\varepsilon(s), \bar{u}) - f^i(s, x_0(s), \bar{u})]\,ds$$

(B) $$+ \int_{t_0-\varepsilon}^t [f^i(s, x_0(s), \bar{u}) - f^i(t_0, x_0(t_0), \bar{u})]\,ds$$

(C) $$+ \int_{t_0-\varepsilon}^t [f^i(t_0, x_0(t_0), \bar{u}) - f^i(t_0, x_0(t_0), u_0(t_0))]\,ds$$

(D) $$+ \int_{t_0-\varepsilon}^t [f^i(t_0, x_0(t_0), u_0(t_0)) - f^i(s, x_0(s), u_0(s))]\,ds$$

(E) $$+ \int_{t_0-\varepsilon}^t \left[\sum_\rho g^i_\rho(s, x_0(s) + v_\varepsilon(s)) - g^i_\rho(s, x_0(s))\right] dz^\rho$$

(F) $$+ \int_{t_0-\varepsilon}^t \sum_{\rho\sigma} [G^i_{\rho\sigma}(s, x_0(s) + v_\varepsilon(s)) - G^i_{\rho\sigma}(s, x_0(s))]\,dz^\rho\,dz^\sigma.$$

Since (C) $= \varepsilon[f^i(t_0, x_0(t_0), \bar{u}) - f^i(t_0, x_0(t_0), u_0(t_0))]$, the result will follow immediately if we can show $\|(A)\| + \|(B)\| + \|(D)\| + \|(E)\| + \|(F)\| = o(\varepsilon)$. We shall show first that $\|(A)\| + \|(E)\| + \|(F)\| = o(\varepsilon)$. Since f^i was assumed to be of class C_2 with bounded derivatives, there is a constant L_f such that

$$|f^i(s, x_0(s) + v_\varepsilon(s), \bar{u}) - f^i(s, x_0(s), \bar{u})| \leqq L_f|v_\varepsilon(s)|.$$

This inequality holds for all of the (invisible) ω's, so

$$\|f^i(s, x_0(s) + v_\varepsilon(s), \bar{u}) - f^i(s, x_0(s), \bar{u})\| \leqq L_f\|v_\varepsilon(s)\|.$$

By Theorem 1, there exists a bound, say W, for $\|v_\varepsilon(s)\|$ on $[t_0 - \varepsilon, t_0]$. Let $Q_1 = L_f W$. Then

$$\begin{aligned}\|(\mathrm{A})\| &= \left\| \int_{t_0-\varepsilon}^{t} [f^i(s, x_0(s) + v_\varepsilon(s), \bar{u}) - f^i(s, x_0(s), \bar{u})]\, ds \right\| \\ &\leqq \int_{t_0-\varepsilon}^{t} \| f^i(s, x_0(s) + v_\varepsilon(s), \bar{u}) - f^i(s, x_0(s), \bar{u})\|\, ds \\ &\leqq \int_{t_0-\varepsilon}^{t} L_f \|v_\varepsilon(s)\|\, ds \\ &\leqq L_f W \int_{t_0-\varepsilon}^{t} ds \leqq Q_1 \varepsilon.\end{aligned}$$

By identical arguments, there exist constants Q_2 and Q_3 such that $\|(\mathrm{B})\| \leqq Q_2\varepsilon$ and $\|(\mathrm{D})\| \leqq Q_3\varepsilon$.

$$\begin{aligned}\|(\mathrm{E})\| &= \left\| \int_{t_0-\varepsilon}^{t} \sum_{\rho} [g_\rho^i(s, x_0(s) + v_\varepsilon(s)) - g_\rho^i(s, x_0(s))]\, dz^\rho \right\| \\ &\leqq \sum_{\rho} K_{10} \left[\int_{t_0-\varepsilon}^{t} \|g^i(s, x_0(s) + v_\varepsilon(s)) - g^i(s, x_0(s))\|^2\, ds \right]^{1/2}\end{aligned}$$

g_ρ^i is Lipschitzian for all ρ, hence, as above, there exists L_g such that

$$\|g^i(s, x_0(s) + v_\varepsilon(s)) - g^i(s, x_0(s))\| \leqq L_g \|v_\varepsilon(s)\| \leqq L_g W.$$

So

$$\|(\mathrm{E})\| \leqq rK_{10} \left[\int_{t_0-\varepsilon}^{t} L_g^2 W^2\, ds \right]^{1/2} \leqq rK_{10} L_g W \varepsilon^{1/2}.$$

Let $Q_4 = rK_{10}L_gW$; then $\|(\mathrm{E})\| \leqq Q_4\varepsilon^{1/2}$. By an identical argument, if $Q_5 = r^2K_9L_GW$, then $\|(\mathrm{F})\| \leqq Q_5\varepsilon^{1/2}$. Since for small ε, $\varepsilon < \varepsilon^{1/2}$, if we let $Q = \sum_{i=1}^{5} Q_i$, we have shown that

$$\|v_\varepsilon(t)\| \leqq Q\varepsilon^{1/2}.$$

But using this fact and the previous argument concerning $\|(\mathrm{A})\|$, we now have

$$\begin{aligned}\|(\mathrm{A})\| &\leqq L_f \int_{t_0-\varepsilon}^{t} \|v_\varepsilon(s)\|\, ds \\ &\leqq L_f \int_{t_0-\varepsilon}^{t} Q\varepsilon^{1/2}\, ds \leqq L_f Q \varepsilon^{3/2}.\end{aligned}$$

Similarly,

$$\begin{aligned}\|(\mathrm{E})\| &\leqq rK_{10} \left[\int_{t_0-\varepsilon}^{t} L_d^2 \|v_\varepsilon(s)\|^2\, ds \right]^{1/2} \\ &\leqq rK_{10} [L_g^2 \varepsilon (Q\varepsilon^{1/2})^2]^{1/2} \\ &\leqq rK_{10} L_g Q \varepsilon,\end{aligned}$$

and

$$\|(\mathrm{F})\| \leqq r^2 K_9 L_G Q \varepsilon.$$

Since for small ε, $\varepsilon^{3/2} < \varepsilon$, if we let $M = L_f Q + Q + rK_{10} L_g Q + r^2 K_9 L_G Q$, we have $\|v_\varepsilon(s)\| \leqq M\varepsilon$. Running the argument one more time on $\|(\mathrm{E})\|$ and $\|(\mathrm{F})\|$, we get $\|(\mathrm{E})\| \leqq rK_{10} L_g M \varepsilon^{3/2}$.

$$\|(\mathrm{F})\| \leqq r^2 K_9 L_G M \varepsilon^{3/2}.$$

So we have shown that $\|(\mathrm{A})\|$, $\|(\mathrm{E})\|$ and $\|(\mathrm{F})\|$ are all $o(\varepsilon)$.

As equipment for working on $\|(\mathrm{B})\|$ and $\|(\mathrm{D})\|$, we need the following observation:

Let $\Omega[N] = \{\omega | \sup_{t\in[0,T]} |x_0(t,\omega)| \geqq N\}$. Then if L_f is the Lipschitz constant for f regarded as a function of x (guaranteed to exist by hypothesis (VIII)), we have

$$|f^i(t,x,\bar{u})| \leqq |f^i(t,0,\bar{u})| + L_f|x(t)| \quad \text{for } i = 1,\cdots,n, \quad t\in[0,T].$$

Let $K = \max f^i(t,0,\underline{u})$. Then

$$|f^i(t,x,\bar{u})| \leqq K + L_f|x(t)|,$$

and hence

$$\|f^i(t,x_0(t),\bar{u})1_{N'}\| \leqq \|K1_{N'}\| + \|L_f x_0(t)1_{N'}\|,$$

for $i = 1,\cdots,n$ and $t\in[0,T]$. By Lemma 2, the right-hand term tends to zero uniformly in t as N tends to ∞. It follows that the left-hand term does so.

To simplify the rest of the work with $\|(\mathrm{B})\|$, we define a function $\theta(s)$ by

$$\theta(s) = f^i(s,x_0(s),\bar{u}) - f^i(t_0,x_0(t_0),\bar{u}).$$

Then

$$\begin{aligned}\|(\mathrm{B})\| &\leqq \int_{t_0-\varepsilon}^{t_0} \|\theta(s)\|\,ds \\ &\leqq \int_{t_0-\varepsilon}^{t_0} \|f^i(s,x_0(s),\bar{u})1_{N'}\|\,ds + \int_{t_0-\varepsilon}^{t_0} \|f^i(t_0,x_0(t_0),\bar{u})1_{N'}\|\,ds \\ &\qquad + \int_{t_0-\varepsilon}^{t_0} \|\theta(s)1_{N''}\|\,ds.\end{aligned}$$

Let $\delta > 0$ be given. Then the observation above and the fact that $f^i(t_0,x_0(t_0),\bar{u})$ is constant permit us to choose N such that each of the first two integrands is less than $\delta/4$. $\theta(s)1_{N''}$ is bounded, since all f^i are continuous and we are working over the compact region $[0,T]\times\{x|\,\|x\| \leqq N\}\times U$. Furthermore, $\theta(s)1_{N''}$ is continuous in s for almost all ω. Hence

$$\lim_{s\to t_0} [\theta(s)1_{N''}]^2 = 0 \quad \text{a.s.}$$

Thus using the bounded convergence theorem,

$$\|\theta(s)1_{N''}\| \to 0 \quad \text{as } s \to t_0.$$

Thus there is an $\varepsilon' > 0$ such that on $[t_0 - \varepsilon', t_0]$,

$$\|\theta(s)1_{N''}\| < \delta/2.$$

It follows that if $0 < \varepsilon < \varepsilon'$,

$$\|(B)\| < \int_{t_0-\varepsilon}^{t_0} \delta/4\, ds + \int_{t_0-\varepsilon}^{t_0} \delta/4\, ds + \int_{t_0-\varepsilon}^{t_0} \delta/2\, ds = \varepsilon\delta.$$

Hence $\|(B)\|$ is $o(\varepsilon)$. The argument for $\|(D)\|$ is identical.

Acknowledgment. The statements in this paper form part of the author's doctoral dissertation for Brown University, written under the helpful direction of Professor Wendell Fleming. Some of the proofs have since been simplified. The author also wishes to thank Professor E. J. McShane for his patient assistance at many stages of the work on both dissertation and paper.

REFERENCES

[1] W. H. FLEMING, *Optimal control of partially observable diffusions*, this Journal, 6 (1968), pp. 194–214.
[2] H.-H. KUO, *On the stochastic maximum principle in Banach space*, J. Functional Analysis, 14 (1973), pp. 146–161.
[3] H. J. KUSHNER, *On the stochastic maximum principle: Fixed time of control*, J. Math. Anal. Appl., 11 (1965), pp. 78–92.
[4] E. J. MCSHANE, *Toward a stochastic calculus I and II*, Proc. Nat. Acad. Sci. U.S.A., 63 (1969), pp. 275–280, 1084–1087.
[5] ———, *Stochastic integrals and stochastic functional equations*, SIAM J. Appl. Math., 17 (1969), pp. 287–306.
[6] ———, *Stochastic Calculus and Stochastic Models*, Academic Press, New York, 1974.
[7] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962 (English transl.).
[8] H. RADEMACHER, *Uber partielle und totale differenzierbarkeit*, Math. Ann., 79 (1919), pp. 340–359.
[9] V. M. WARFIELD, *Existence and adjoint theorems for linear stochastic differential equations*, Pacific J. Math., 51 (1974), pp. 305–320.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

STOCHASTIC APPROXIMATION ALGORITHMS OF THE MULTIPLIER TYPE FOR THE SEQUENTIAL MONTE CARLO OPTIMIZATION OF STOCHASTIC SYSTEMS*

HAROLD J. KUSHNER† AND MILTON L. KELMANSON‡

Abstract. Many stochastic control (or parametrized) systems have (expected value) objective functions of largely unknown form, but where noise corrupted observations can be taken at any selected value of a finite-dimensional parameter x. The parameter x must satisfy equality and inequality constraints. The usual numerical techniques of nonlinear programming on control theory are not usually helpful here. The paper discusses a number of algorithms (with convergence proofs) for selecting a sequence of parameter values $\{X_n\}$, where X_n depends on X_{n-1} and observations taken at X_{n-1}, and the limit points are both feasible and satisfy the Kuhn–Tucker necessary condition (w.p. 1 (with probability 1)). The algorithms are stochastic "small step" versions of the deterministic combined penalty function-multiplier methods.

1. Introduction. For some integers s, t, let $f(\cdot)$, $\phi_i(\cdot)$, $i = 1, \cdots, s$, $q_i(\cdot)$, $i = 1, \cdots, t$ denote continuous, twice differentiable real-valued functions on Euclidean r-space R^r, with uniformly bounded mixed second derivatives. $\phi(\cdot), q(\cdot)$ denote the vectors with the components $\phi_i(\cdot), q_i(\cdot)$, respectively. Define the sets $C = \{x : q_i(x) \leqq 0,\ i = 1, \cdots, t\}$, and $B = \{x : \phi_i(x) = 0,\ i = 1, \cdots, s\}$. For each $x \in R^r$, let $H(\cdot|x)$ and $\hat{H}(\cdot|x)$ denote distribution functions of real-valued and R^r-valued, respectively, random variables with uniformly (in x) bounded variance (covariance, resp)., and $\int y\, dH(y|x) = f(x)$, $\int v\, d\hat{H}(v|x) = f_x(x)$, where $f_x(\cdot)$ is the gradient of $f(\cdot)$. The paper is concerned with several algorithms for finding (sequentially) a local minimum of $f(x)$ in C, B or $C \cap B$. The functions $\phi_i(\cdot), q_i(\cdot)$ are known and their values or values of their derivatives can be calculated at any x. We do not assume that $f(\cdot)$ is known but, given a parameter X, we can draw one or more random variables from the distribution (with parameter value X) $H(\cdot|X)$ or $\hat{H}(\cdot|X)$, depending on the case. If $\tilde{X}_i$, $i \leqq n$, $\tilde{Y}_i$, $i \leqq n$, are the first n parameter values at which draws are made, and the values, respectively, and $\tilde{X}_{n+1}$ is the $(n+1)$st parameter value at which a draw (denoted by $\tilde{Y}_{n+1}$) is to be made, then we suppose that $E[\tilde{Y}_{n+1}|\tilde{X}_i,\ i \leqq n+1,\ \tilde{Y}_i,\ i \leqq n] = f(\tilde{X}_{n+1})$ (or $f_x(X_{n+1})$, according to the case).

The algorithms are roughly of the stochastic approximation type. An initial estimate, X_0, of a local minimum, is made, one or more observations are taken at X_0, a new estimate X_1 is calculated, etc. As is generally true in nonlinear programming, it is quite difficult (except under certain convexity conditions) to devise practical computational algorithms which are guaranteed to (eventually) find a true local or global minimum. In this paper, the algorithms generate (as is usual in deterministic nonlinear programming) a sequence $\{X_n\}$ whose limit points are feasible (meaning that they satisfy the constraints; they are in C, B or $C \cap B$, according to the case) and which satisfy one of the usual local necessary

† Division of Applied Mathematics and Engineering, Brown University, Providence, Rhode Island 02912. The work of this author was supported by the Air Force Office of Scientific Research under AF-AFOSR 71-2078C and in part by the National Science Foundation under Eng-73-03846-AO1, Office of Naval Research NONR N1467-AD-191001805.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was supported in part by CAPES, Ministry of Education, Brasil.

conditions for minimality; in particular, either the Kuhn–Tucker condition, or the necessary condition of the calculus, according to the case.

References Kushner [1] and Kushner and Gavin [2] dealt with a family of (inequality constrained only) algorithms, based on the deterministic methods of feasible directions, in which each iterate satisfied the constraints. The search was divided into cycles, X_n, denoting the initial point of the nth search cycle, and all limit points (w.p. 1) of $\{X_n\}$ satisfied a necessary condition for minimality. The general conditions on each search cycle implicitly required that the "search effort" per cycle increase as the cycle number increased. Part of the reason for the requirement for the increasing effort/cycle is the difficulty of analyzing the algorithm when the iterates are on or near the boundary of the feasible region. Numerical experiments (such as those reported in [2] suggest that more efficient use is made of the observation if the effort per cycle does not increase. In Kushner and Sanvicente [3], a penalty function-like method was developed (for inequality constrained case). There the iterates were not constrained to be feasible (the algorithm guaranteed that the limits would be), but the method shares with the deterministic penalty function method the numerical disadvantages that the penalty functions ultimately increase extremely rapidly for x outside of the feasible set.

Here, several stochastic approximation-like versions of the so-called deterministic methods of multipliers [4]–[7] will be developed. The methods in [4]–[7] do not require feasibility, and avoid some of the numerical problems associated with penalty function techniques. Intuitively, it seems very reasonable to expect that the numerical advantages which the techniques have in the deterministic case (say, with the methods in [4]–[7]) would also hold for the stochastic algorithms discussed below.

For the sake of simplicity of notation in the proofs, we do the pure equality constraint case in § 2, and the pure inequality constraint case in §§ 3 and 4. It should be fairly clear that the combined problem can be handled by a combination of the ideas in the proofs of §§ 2, 3 and 4.

There are numerous applications of these systematic Monte Carlo optimization techniques. Typically $f(x)$ represents the average response of a physical system with parameter x. Only noise corrupted data $f(x) + \xi$ is available at each chosen value of x (ξ = observation noise). If the system is complex, $f(\cdot)$ will not be known, and we may have to resort to an "experimental" method for optimization. Experience with such methods in the stochastic case suggests that "small step" methods are probably preferable. The paper is concerned with convergence theorems for several such methods.

2. Equality constraints. The algorithms in this section are stochastic approximations of those discussed by Miele et al. [4] (in the sense that the Kiefer–Wolfowitz method is a stochastic version of Newton's method), where no actual convergence proofs are given. In order to minimize the number of terms in our expansions, we assume that the observations are taken with $\hat{H}(\cdot\,|x)$; i.e., given X_k, observe Y_n where Y_n is distributed as $\hat{H}(\cdot\,|X_n)$. Let $\mathscr{B}_n$ denote the smallest σ-algebra determined by $X_0, \cdots, X_n$. Then $E_{\mathscr{B}_n} Y_n = f_x(X_n)$, $\operatorname{covar}_{\mathscr{B}_n} \xi_n \leqq \sigma^2 I$ for some real σ^2, where $\xi_n \equiv Y_n - f_x(X_n)$. There are only minor changes in the

assumptions if $f_x(X_n)$ must be estimated via finite differences, and the changes will be discussed later. Let k denote a (henceforth fixed) positive real number, and define the functions $P(\cdot)$, $\Phi(\cdot)$, $L(\cdot,\cdot)$ and $W(\cdot,\cdot)$ by $P(x) \equiv |\phi(x)|^2$, $\Phi(x) \equiv$ Jacobian of $\phi(\cdot)$ at x, $\Phi_n \equiv \Phi(X_n)$, $L(x,\lambda) \equiv f(x) + \lambda'\phi(x)$ for a vector λ with real components (where $'$ denotes transpose, and the norm is Euclidean), and $W(x,\lambda) \equiv L(x,\lambda) + (k/2)P(x)$.

The following assumptions will be used. We require $s \leqq r$.

(A1) Let $\{a_n\}$ be a sequence of positive real numbers with $\sum_n a_n = \infty$.

(A2) $\sum_n a_n^2 < \infty$.

(A3) $\Phi'(x)\phi(x) = 0$ implies that $\Phi(x)$ is of full rank (hence also that $\phi(x) = 0$).

For each $h \in R^r$, define $\pi(x)h$ to be the projection of h on the orthogonal complement to the subspace of R^r determined by the rows of $\Phi(x)$. If $\Phi(x)$ is of full rank, then[1]

$$\pi(x)h = [I - \Phi'(x)(\Phi(x)\Phi'(x))^{-1}\Phi(x)]h. \tag{2.1}$$

In any case, (2.1) holds if the inverse is interpreted as the pseudoinverse, which we will do.

For each $\varepsilon > 0$, define the set $G_\varepsilon = \{x : |\pi(x)f_x(x)|^2 \leqq \varepsilon\}$, and write $G_0 = G$. $G \cap B$ *is the set of feasible stationary points.* It is closed and is the union of a collection of disjoint closed and connected sets $S_i, \cdots$, on each of which $f(x)$ is constant, say, taking value f_i on S_i. Assumption (A4) is not a serious practical restriction.

(A4) There are only finitely many sets $S_1, \cdots$.

In the equality constrained case, we must show that the sequences $\{X_n\}$ generated by the algorithms converge to $G \cap B$.

ALGORITHM 1. Given the iterate X_n (with components denoted by X_n^i, $i = 1, \cdots, r$), X_{n+1} is given by the parameterized (by λ_n) form

$$\begin{aligned} X_{n+1} &= X_n - a_n\left[Y_n + \Phi_n'\lambda_n + \frac{k}{2}P_x(X_n)\right] \\ &= X_n - a_n\left[f_x(X_n) + \Phi_n'\lambda_n + \frac{k}{2}P_x(X_n) + \xi_n\right] \end{aligned} \tag{2.2}$$

where $P_x(x) = 2\Phi'(x)\phi(x)$.

If x is a constrained minimum, then we know from calculus that there is a vector $\lambda = (\lambda_1, \cdots, \lambda_s)$ and scalar λ_0 (not both zero) so that $\lambda_0 f_x(x) + \sum_i \lambda_i \phi_{i,x}(x) = \lambda_0 f_x(x) + \Phi'(x)\lambda = 0$. By (A3), the $\{\phi_{i,x}(x)\}$ are linearly independent, and we can take $\lambda_0 \neq 0$. This suggests that we choose λ_n so that the norm of the estimated gradient of the Lagrangian is minimized. Namely, we let λ_n minimize $|L_x(X_n,\lambda_n) + \xi_n|^2$. Equivalently, we let (following the idea in [4]) λ_n satisfy the orthogonality relationship

$$\Phi_n(f_x(X_n) + \xi_n + \Phi_n'\lambda_n) = 0. \tag{2.3}$$

If Φ_n is of full rank, then $\Phi_n\Phi_n'$ is invertible and λ_n is unique and is given by

$$\begin{aligned} \lambda_n &= -[\Phi_n\Phi_n']^{-1}\Phi_n f_x(X_n) - [\Phi_n\Phi_n']^{-1}\Phi_n\xi_n \equiv \bar{\lambda}_n + \hat{\lambda}_n \\ &\equiv E_{\mathscr{B}_n}\lambda_n + (\lambda_n - E_{\mathscr{B}_n}\lambda_n). \end{aligned} \tag{2.4}$$

[1] It may be computationally preferable to use the pseudoinverse of $\Phi'(x)$ for $(\Phi(x)\Phi'(x))^{-1}\Phi(x)$.

If $x = X_n$, we write $\pi_n h$ in (2.1). If Φ_n or $\Phi(x)$ are not of full rank, then (2.3) still has a solution, although not necessarily unique and we may suppose that it is (2.4) with the pseudoinverse of $[\Phi_n\Phi_n']$ replacing the inverse. We will use the forms (2.1), (2.4) with the inverse indicating either the inverse or pseudoinverse. Define the function $\bar{\lambda}(\cdot)$ by $\bar{\lambda}(x) = -[\Phi(x)\Phi'(x)]^{-1}\Phi(x)f_x(x)$, where the pseudoinverse is used. By (A3), $(\Phi(x)\Phi'(x))$ is invertible at all feasible x.

From (2.1), (2.2), (2.4),

$$X_{n+1} = X_n - a_n\left(\pi_n f_x(X_n) + \pi_n\xi_n + \frac{k}{2}P_x(X_n)\right). \tag{2.5}$$

In all proofs K denotes a positive real number. Its value may change from usage to usage.

THEOREM 2.1. *Assume the conditions in the introduction and also* (A1) *to* (A4). *Then there is a null set N so that if $\omega \notin N$ and $\sup_n |X_n(\omega)| < \infty$ and x is any limit point of $\{X_n(\omega)\}$, then $\phi(x) = 0$ and there is a vector (perhaps depending on the limit x) $\psi = (\psi^1, \cdots, \psi^s)$ for which*

$$f_x(x) + \Phi'(x)\psi = 0.$$

(An equivalent statement is that if $\omega \notin N$ and $\sup_n |X_n(\omega)| < \infty$, then all limit points are in $G \cap B$.)

Remark. We note that convergence takes place for all values of $k > 0$, unlike in the deterministic case [4]–[7] where k must be greater than some minimum value.

Remark. It follows from the arguments of Part 1 of the proof of Theorem 2.1 that if $P(x) \to \infty$ as $|x| \to \infty$ and if either $a_n|\pi_n f_x(X_n)|^2 \to 0$ w.p.1 as $n \to \infty$, or $|\pi(x)f_x(x)| \leqq K|\Phi'(x)\phi(x)|$ for large x and some real K, then $P(X_n) \to 0$ w.p.1, and $\sup_n |X_n(\omega)| < \infty$ w.p.1 is implied.

Proof. Until further notice, we suppose that there is some real M for which $|X_n| \leqq M$ w.p.1, all n, and that the generic variable x satisfies $|x| \leqq M$.

Part (i). By a straightforward Taylor series expansion and the use of (2.3) (which certainly holds if $P_x'(X_n)$ replaces Φ_n there) we get

$$\begin{aligned} P(X_{n+1}) - P(X_n) \leqq\ & -a_n P_x'(X_n)\left[f_x(X_n) + \Phi_n'\lambda_n + \xi_n + \frac{k}{2}P_x(X_n)\right] \\ & + a_n^2 K\left|f_x(X_n) + \Phi_n'\lambda_n + \xi_n + \frac{k}{2}P_x(X_n)\right|^2 \\ \leqq\ & -a_n\frac{k}{2}|P_x(X_n)|^2 + a_n^2 K[|f_x(X_n) + \Phi_n'\bar{\lambda}_n|^2 \\ & + |\xi_n + \Phi_n'\hat{\lambda}_n|^2 + |P_x(X_n)|^2] \end{aligned}$$

which yields

$$\begin{aligned} P(X_{n+1}) - P(X_n) \leqq\ & -a_n k|\Phi_n'\phi(X_n)|^2 + a_n^2 K[|\pi_n f_x(X_n)|^2 \\ & + |\pi_n\xi_n|^2 + |\Phi_n'\phi(X_n)|^2]. \end{aligned} \tag{2.6}$$

Define the set $\tilde{B}_\varepsilon = \{x : |\Phi'(x)\phi(x)|^2 \leqq \varepsilon\}$. By (A3), $\tilde{B}_\varepsilon \to B$ as $\varepsilon \to 0$. For each

$\varepsilon > 0$, there is a $\delta_\varepsilon > 0$ so that $\delta_\varepsilon \to 0$ as $\varepsilon \to 0$ and $\tilde{B}_\varepsilon \subset B_{\delta_\varepsilon} \equiv \{x: |P(x)| \leqq \delta_\varepsilon\}$, (by A3). By (A2)

$$\lim_{N\to\infty} \sum_{n=N}^{\infty} a_n^2 |\pi_n \xi_n|^2 = 0$$

w.p.1. For large n, the other second order (in a_n) terms in (2.6) are at most half of the absolute value of the first order term—for X_n in $R^r - \tilde{B}_\varepsilon$. The last two sentences (the convergence and the dominance) and the divergence of $\sum_n a_n$ imply that $X_n \in \tilde{B}_\varepsilon \subset B_{\delta_\varepsilon}$ infinitely often w.p.1, for any $\varepsilon > 0$. Fix ε. The same sentences and (2.6) imply that $\{X_n\}$ can go from B_{δ_ε} to $R^r - B_{\delta_{3\varepsilon}}$ only finitely often (w.p.1) without entering $B_{\delta_{2\varepsilon}} - B_{\delta_\varepsilon}$ first, and they also imply that the sequence $\{X_n\}$ can go from a point in $B_{\delta_{2\varepsilon}}$ to $R^r - B_{\delta_{3\varepsilon}}$ at most finitely often w.p.1. Since $\varepsilon > 0$ is arbitrary, $\Phi_n' \phi(X_n) \to 0$ and hence by (A3), $\phi(X_n) \to 0$ (w.p.1).

Part (ii). Now, we turn to evaluate the limits of $f(X_n)$. Similarly to (2.6), we have

$$\begin{aligned} f(X_{n+1}) - f(X_n) &\leqq -a_n f_x'(X_n)\left[f_x(X_n) + \Phi_n' \lambda_n + \frac{k}{2} P_x(X_n) + \xi_n \right] \\ &\quad + a_n^2 K[|\pi_n f_x(X_n)|^2 + |\Phi_n' \phi(X_n)|^2 + |\pi_n \xi_n|^2] \\ &\leqq -a_n f_x'(X_n)\pi_n f_x(X_n) - a_n f_x'(X_n)\pi_n \xi_n \\ &\quad - a_n \frac{k}{2} f_x'(X_n) P_x(X_n) \\ &\quad + a_n^2 K[|\pi_n f_x(X_n)|^2 + |\Phi_n' \phi(X_n)|^2 + |\pi_n \xi_n|^2]. \end{aligned} \tag{2.7}$$

We have $f_x'(X_n)\pi_n f_x(X_n) = |\pi_n f_x(X_n)|^2$ (by the definition of $\pi(x)$ and projection). Note that by (A2) and the bound M, $\sum a_n f_x'(X_n)\pi_n \xi_n$ is a square summable convergent martingale. Also $P_x(X_n) \to 0$ w.p.1 by Part (i). Using these two facts together with the divergence of $\sum_n a_n$ and (2.7), and an argument like that in Part (i) (to show $X_n \in \tilde{B}_\varepsilon$ or B_ε infinitely often w.p.1), we can show that[1] $X_n \in G_\varepsilon$ infinitely often w.p.1 for each $\varepsilon > 0$.

Since the[2] S_i are disjoint and closed, and since $f(x) = f_i$ on S_i, for each small $\delta > 0$, there is an $\varepsilon > 0$ so that we can write $G_\varepsilon \cap B_\varepsilon = \bigcup_i S_i^\varepsilon$, where $\{S_i^\varepsilon\}$ are closed, connected and disjoint and $S_i^\varepsilon \supset S_i$, and the maximum variation of $f(x)$ on each S_i^ε is less than δ. We can (and will) also suppose that if $f_i \neq f_j$, then $|f_i - f_j| \geqq 3\delta$. Let $f_i > f_j$. So for $\{X_n\}$ to go from S_j^ε to S_i^ε, the sequence $\{f(X_n)\}$ must increase by at least δ while outside of $S_j^\varepsilon \cup S_i^\varepsilon$. Now, the inequality (2.7) and the convergence of $\sum a_n f_n'(X_n)\pi_n \xi_n$ and $\sum a_n^2 |\pi_n \xi_n|^2$ and the asymptotic dominance of the other second order terms (in (2.7)) by the first order term for X_n outside of G_ε, and $P(X_n) \to 0$ w.p.1, together imply that $\{X_n\}$ can make only finitely many excursions from S_j to S_i(w.p.1). Indeed, by the same reasoning $\{X_n\}$ can make only finitely many excursions from S_j^ε into any set A, where $\inf_{x\in A} f(x) \geqq f_j + 2\delta$. Since δ is arbitrarily small, $\{f(X_n)\}$ converges w.p.1.

[2] See paragraph above (A4) for the definition.

Let $A_d(y)$ denote a ball in R^r with center y and diameter d, and suppose that for a real $\varepsilon_1 > 0$,

$$\inf_{x \varepsilon A_{2d}(y)} |\pi(x) f_x(x)| \geqq \varepsilon_1 .$$

Define $m_1 = \min\{n : X_n \in A_d(y)\}$, $m_1' = \min\{n : X_n \notin A_{2d}(y), n > m_1\}$ and,[3] in general, $m_i = \min\{n : X_n \in A_d(y), m > m_{i-1}'\}$, $m_i' = \min\{n : X_n \notin A_{2d}(y), n > m_i\}$. Summing (2.7) and using the convergence and dominance cited in the last paragraph yields, w.p.1,

$$\lim_{i \to \infty} \sum_{n=m_i}^{m_i'-1} [f(X_{n+1}) - f(X_n)] \leqq - \lim_{i \to \infty} \varepsilon_1 K \sum_{m_i}^{m_i'-1} a_i .$$

But by summing $X_{n+1} - X_n$ from (2.5) over $[m_i, m_i' - 1]$ and using the summability of $\sum a_n \pi_n \xi_n$ and convergence of $P_x(X_n)$ to 0, and the fact that the distance traveled over those iterates is at least d, we get

$$\lim_{i \to \infty} \sum_{n=m_i}^{m_i'-1} a_n \geqq Kd,$$

which contradicts the convergence of $\{f(X_n)\}$, unless $m_i < \infty$ only finitely often w.p.1. Since y and d are arbitrary, we conclude that $\{X_n\}$ must eventually stay in $G_\varepsilon \cap B_\varepsilon$ w.p.1 for any $\varepsilon > 0$, and hence that $X_n \to G \cap B$ w.p.1, if $\sup_n |X_n| \leqq M$ w.p.1.

The proof without the bound M, but with $\sup_n |X_n| < \infty$ w.p.1 proceeds similarly. We repeat the above proof, but stop the iteration $\{X_n\}$ at the first instant that $|X_n| > M$. Then we conclude that $X_n \to G \cap B$ with a probability $\geqq P\{\sup_n |X_n| \leqq M\}$. Since M is arbitrary, the theorem holds as stated. Q.E.D.

Remark. If we replaced (A3) by "$\Phi'(x)\phi(x) = 0$ implies $\phi(x) = 0$", then the theorem would read: there is a number ψ^0, vector ψ, $(\psi^0, \psi) \neq 0$ such that $\psi^0 f_x(x) + \Phi'(x)\psi = 0$ at almost all limit points.

Finite differences. Let e_i denote the unit vector in the ith coordinate direction, and $\{c_n\}$ a sequence of positive real numbers which converges to zero. Let X_n be given, and let $Y_n(X_n \pm c_n e_i)$ denote a random draw from $H(\cdot\,|X_n \pm c_n e_i)$. Define (the finite difference version of (2.2))

$$X_{n+1}^i = X_n^i - a_n \left[\frac{Y_n(X_n + c_n e_i) - Y_n(X_n - c_n e_i)}{2c_n} + \Phi_n' \lambda_n + \frac{k}{2} P_x(X_n) \right], \quad i = 1, \cdots, r, \tag{2.8}$$

Let λ_n be determined by (2.3) but with the finite difference estimate replacing $f_x(X_n) + \xi_n$ there.

THEOREM 2.2. *Assume the conditions of Theorem* (2.1), *but with* a_n/c_n *replacing* a_n *in* (A2). *Let* $E_{\mathscr{B}_n} Y_n(X_n \pm c_n e_i) = f(X_n \pm c_n e_i)$ w.p.1. *Then the conclusion of Theorem* 2.1 *holds.*

The proof is almost exactly the same as that of Theorem 2.1, except that $f(X_n \pm c_n e_i)$ must be expanded, and the "noise" in the iteration (2.8) is proportional to a_n/c_n, rather than to a_n. Note that we do not require $\sum_n a_n c_n < \infty$ as is

[3] Undefined m_i or m_i' are set equal to ∞.

common in stochastic approximation (the $a_n c_n$ terms in the expansion are ultimately dominated by the a_n term outside of any B_ε).

Observe that, in both Theorem 2.1 and 2.2, if $X_{n_i}(\omega) \to X(\omega)$, then the limit $\bar{\lambda}(X(\omega))$ can be used as the multiplier ψ at $X(\omega)$.

ALGORITHM 2. Again, for the sake of simplicity, we use $\hat{H}(\cdot\,|x)$. In this algorithm (2.2) is still used, but λ_n is selected to assure that the "first order" change in $\phi(X_{n+1}) - \phi(X_n)$ is proportional to $-\phi(X_n)$. In particular, for some positive number k_1 we select λ_n to enforce the relationship (following the idea in [4])

$$\Phi_n(X_{n+1} - X_n) = -a_n k_1 \phi(X_n)$$

or, equivalently, the relationships (2.9) or (2.10).

$$\Phi_n\left[f_x(X_n) + \xi_n + \Phi_n'\lambda_n + \frac{k}{2}P_x(X_n)\right] = k_1\phi(X_n), \tag{2.9}$$

$$\Phi_n[f_x(X_n) + \xi_n + \Phi_n'\lambda_n] = [k_1 I - k\Phi_n\Phi_n']\phi(X_n). \tag{2.10}$$

There is not necessarily a solution to (2.10) unless we replace (A3) by (A3′).

(A3′) $\Phi(x)$ is of full rank for each $x \in R^n$.

Assuming (A3′) and solving (2.10) for λ_n yields

$$\begin{aligned}\lambda_n = &- [\Phi_n\Phi_n']^{-1}\Phi_n f_x(X_n) - [\Phi_n\Phi_n']^{-1}\Phi_n\xi_n \\ &+ [\Phi_n\Phi_n']^{-1}(k_1 I - k\Phi_n\Phi_n')\phi(X_n) \equiv \bar{\lambda}_n + \hat{\lambda}_n + \mathring{\lambda}_k.\end{aligned} \tag{2.11}$$

THEOREM 2.3. *Under the assumptions of Theorem* 2.1 *except with* (A3′) *replacing* (A3), *the conclusion of Theorem* 2.1 *holds for Algorithm* 2.

Proof. The proof is very close to that of Theorem 2.1 and will only be sketched. We will first suppose here, as there, that $|X_n| \leqq M$ w.p.1, for all n, and then let $M \to \infty$ as in that proof. The first inequality of Part (i) of the proof of Theorem 2.1 still holds. The replacement of λ_n in that inequality by its value in (2.11), noting (2.10), yields

$$\begin{aligned}P(X_{n+1}) - P(X_n) &\leqq - a_n k_1|\phi(X_n)|^2 \\ &\quad + a_n^2 K[|f_x(X_n) + \Phi_n'\bar{\lambda}_n|^2 + |\xi_n + \Phi_n'\hat{\lambda}_n|^2 \\ &\quad + |k_1\Phi_n'[\Phi_n\Phi_n']^{-1}\phi(X_n)|^2] \\ &\leqq - a_n k_1|\phi(X_n)|^2 + a_n^2 K[|\pi_n f_x(X_n)|^2 + |\pi_n\xi_n|^2 + |\phi(X_n)|^2],\end{aligned} \tag{2.12}$$

and we can conclude, as in the proof of Theorem 2.1, that $X_n \to B$ w.p.1.

Similarly, we can get

$$\begin{aligned}f(X_{n+1}) - f(X_n) &\leqq - a_n f_x'(X_n)\left[f_x(X_n) + \xi_n + \Phi_n'\lambda_n + \frac{k}{2}P_x(X_n)\right] \\ &\quad + a_n^2 K[|\pi_n f_x(X_n)|^2 + |\pi_n\xi_n|^2 + |\phi(X_n)|^2] \\ &\leqq - a_n f_x'(X_n)[\pi_n f_x(X_n) + \pi_n\xi_n + k_1\Phi_n'(\Phi_n\Phi_n')^{-1}\phi(X_n)] \\ &\quad + a_n^2 K[|\pi_n f_x(X_n)|^2 + |\pi_n\xi_n|^2 + |\phi(X_n)|^2].\end{aligned} \tag{2.13}$$

An argument similar to that in Part (ii) of the proof of Theorem 2.1 yields that $X_n \to G \cap B$ as $n \to \infty$ w.p.1.

There is an obvious finite difference analogue to Algorithm 2, but we omit the details.

3. Inequality constraints. Two different types of algorithms for the inequality constrained problem will be discussed, one here and one in § 4. Generally, "small step" methods require some sort of nonsingularity or linear independence assumption on the set of gradients of the constraint functions to prevent the iterates from getting "hung up" at some nonfeasible point. The problem exists in the deterministic case as well. See, for example, Polak [8, pp. 142–143]. Assumption (A3″) below and (A3‴) in § 4 are two different types of such an assumption. In this section, we constrain the Lagrange multipliers corresponding to the inequality constraints to be nonnegative. In § 4, the signs are not constrained, but the algorithm is more involved and additional conditions are needed to assure that those multipliers are nonnegative at the limit points. In this section, the requirement that those multipliers be nonnegative forces us to use a stronger condition on the noise (A6). These problems and conditions seem to be rather natural for the stochastic algorithms.

For notational simplicity, we draw the observations from $\hat{H}(\cdot\,|x)$ rather than from $H(\cdot\,|x)$, but there is an obvious finite difference analogue. Also, we treat the pure inequality case. Define $\tilde{q}_i(x) = \max[0, q_i(x)]$, and $P(x) = \sum_i \tilde{q}_i^2(x)$. Then $P_x(x) = 2\Phi'(x)\tilde{q}x)$, where $\Phi(x)$ is the Jacobian of $q(x)$. We let the components of $q(x)$ or $\tilde{q}(x)$ range over the $q_i(x)$ or $\tilde{q}_i(x)$ for which $q_i(x) \geqq 0$. Define $\Phi_n \equiv \Phi(x_n)$.

ALGORITHM 3. We use the same iteration as in Algorithm 1, namely,

$$X_{n+1} = X_n - a_n\left[f_x(X_n) + \xi_n + \Phi_n'\lambda_n + \frac{k}{2}P_x(X_n)\right], \tag{3.1}$$

where λ_n is a λ that minimizes in

$$\min_{\text{all } \lambda_i \geqq 0} |f_x(X_n) + \xi_n + \Phi_n'\lambda_n|^2 = \min_{\text{all } \lambda_i \geqq 0} |L_x(X_n, \lambda_n) + \xi_n|^2. \tag{3.2}$$

By the Kuhn–Tucker theorem, there is a vector c_n with nonnegative components c_n^i, $i = 1, \cdots, t$, so that λ_n satisfies (note that the gradient of the constraint $-\lambda_i \leqq 0$ with respect to λ is the unit vector with $a - 1$ in the ith position)

$$\Phi_n(f_x(X_n) + \xi_n + \Phi_n'\lambda_n) - c_n = 0, \tag{3.3}$$

where $c_n^i = 0$ if $\lambda_n^i > 0$.

For each $h \in R^r$ define $\pi^+(x)h \equiv h + \Phi'(x)\lambda$, where λ minimizes in $\min_{\text{all } \lambda_i \geqq 0} |h + \Phi'(x)\lambda|$. Note that $\pi^+(x)h$ is defined analogously to $\pi(x)h$ in § 2, but that it is the "error" in the projection of h onto $K(x)$, the cone generated by the nonnegative linear combinations of the row vectors of $-\Phi(x)$. Let $\tilde{\pi}_n^+ h$ be defined as $\pi_n^+ h$, but where we use only the rows of the matrix $\tilde{\Phi}_n$, which is obtained from Φ_n by deleting all rows i of Φ_n for which $\lambda_n^i = 0$. (The indices of the deleted rows are random variables and ξ_n dependent.)

Define the set $F_\varepsilon = \{x : |\pi^+(x)f_x(x)|^2 \leqq \varepsilon\}$ and $F_0 = F$. Then $F \cap C$ is the set of feasible points satisfying the Kuhn–Tucker condition $f_x(x) + \sum_{i \text{ active}} \lambda^i q_{i,x}(x) = 0$ for some λ with all $\lambda_i \geqq 0$, and we must show that $X_n \to F \cap C$

w.p.1. $F \cap C$ is closed and is the union of closed, connected and disjoint sets $U_1, \cdots$ on each of which $f(\cdot)$ is constant taking, say, the value f_i on U_i.

We need the following assumptions.

(A3″) $\Phi'(x)\tilde{q}(x) = 0$ implies that $\tilde{q}(x) = 0$, and at any $x \in \partial C$, (the boundary of C) the gradients of the *active* constraints are linearly independent.

(A5) There are only finitely many sets $U_1, \cdots$.

(A6) There is a real $k_1 > 0$ (independent of n, ω) so that

$$f_x'(X_n)E_{\mathscr{B}_n}\tilde{\pi}_n^+(f_x(X_n) + \xi_n) \geqq k_1 f_x'(X_n)\pi_n^+ f_x(X_n).$$

Assumption (A6) would seem to be difficult to explicitly verify in general, yet it holds in most of the specific special cases which we have checked graphically (by selecting simple noise distributions and constructing the projections), and we expect that it holds in a large enough number of cases for the algorithm to be useful. Some such condition appeared in all the variants of the algorithm, when $\lambda_n^i \geqq 0$ was required.

THEOREM 3.1. *Assume the conditions in the introduction and also* (A1), (A2), (A3″), (A5) *and* (A6). *Then there is a null set N so that if* $\omega \notin N$ *and* $\sup_n |X_n(\omega)| < \infty$ *and x is a limit point of* $\{X_n(\omega)\}$, *then* $q(x) \leqq 0$ *and there is a vector* $\psi, \psi^i \geqq 0$, *with* $\psi^i = 0$ *if* $q_i(x) < 0$, *for which*

$$f_x(x) + \Phi'(x)\psi = 0. \tag{3.4}$$

Remark. A condition similar to that in the remark after the statement of Theorem 2.1 implies that $\sup_n |X_n(\omega)| < \infty$ w.p.1.

Proof. As in the proof of Theorem 2.1, we can and will suppose that $|X_n(\omega)| \leqq$ some $M < \infty$, and that the generic variable x satisfies $|x| \leqq M$. The proof is very close to that of Theorem 2.1 and will only be outlined.

Part (i). Note that

$$|\Phi_n'\lambda_n|^2 \leqq |f_x(X_n) + \xi_n|^2 \tag{3.5}$$

and that (3.2) and (3.3) and $c_n^i \geqq 0$ and $\tilde{q}_i'(X_n) \geqq 0$ imply that

$$\tilde{q}'(X_n)\Phi_n\left(f_x(X_n) + \xi_n + \Phi_n'\lambda_n + \frac{k}{2}P_x(X_n)\right) \geqq \frac{k}{2}\tilde{q}'(X_n)\Phi_n P_x(X_n)$$

$$= \frac{k}{2}|\tilde{q}'(X_n)\Phi_n|^2.$$

Substituting these estimates in the first inequality of the proof of Theorem 2.1 yields

$$\begin{aligned} P(X_{n+1}) - P(X_n) \leqq & -a_n\frac{k}{2}|\tilde{q}'(X_n)\Phi_n|^2 \\ & + a_n^2 K[|f_x(X_n)|^2 + |\xi_n|^2 + |P_x(X_n)|^2]. \end{aligned} \tag{3.6}$$

By (A3″), for each $\varepsilon > 0$, there is a $\delta_\varepsilon > 0$ so that $|\tilde{q}'(x)\Phi(x)|^2 \leqq \delta_\varepsilon$ implies that $x \in N_\varepsilon(C)$, an ε neighborhood of C. Thus using the fact that the a_n^2 terms are summable and arguing as in Part (i) of the proof of Theorem 2.1, we can conclude that the X_n must ultimately be in $N_\varepsilon(C)$, for each $\varepsilon > 0$.

Part (ii). A truncated Taylor series expansion yields

$$f(X_{n+1}) - f(X_n) \leqq -a_n f_x'(X_n)\left[f_x(X_n) + \xi_n + \Phi_n'\lambda_n + \frac{k}{2}P_x(X_n)\right] + a_n^2 K\left|f_x(X_n) + \xi_n + \Phi_n'\lambda_n + \frac{k}{2}P_x(X_n)\right|^2. \tag{3.7}$$

Using $\Phi_n'\lambda_n = \tilde{\Phi}_n'\tilde{\lambda}_n$ we find that

$$\begin{aligned} f_x'(X_n)[f_x(X_n) + \xi_n + \Phi_n'\lambda_n] &= f_x'(X_n)\tilde{\pi}_n^+(f_x(X_n) + \xi_n) \\ &\equiv E_{\mathcal{B}_n} f_x'(X_n)\tilde{\pi}_n^+(f_x(X_n) + \xi_n) + \rho_n, \end{aligned} \tag{3.8}$$

where $\{\rho_n\}$ is a sequence of orthogonal random variables and $\sum a_n\rho_n$ is a square summable convergent martingale. Substituting (3.8) and (3.5) into (3.7) and using (A6) yields

$$\begin{aligned} f(X_{n+1}) - f(X_n) \leqq &-a_n k_1 f_x'(X_n)\pi_n^+ f_x(X_n) - a_n\rho_n - a_n\frac{k}{2}f_x'(X_n)P_x(X_n) \\ &+ a_n^2 K[|f_x(X_n)|^2 + |\xi_n|^2 + |P_x(X_n)|^2]. \end{aligned}$$

The $a_n\rho_n$ and a_n^2 terms are summable, and $P_x(X_n) \to 0$ as $n \to \infty$ by Part (i).

Since the U_i are disjoint and closed, and since $f(x) = f_i$ on U_i, for each sufficiently small $\delta > 0$, there is an $\varepsilon > 0$ so that we can write $C_\varepsilon \cap F_\varepsilon = \bigcup_i U_i^\varepsilon$, where U_i^ε are closed, connected, disjoint, and $U_i^\varepsilon \supset U_i$, and the maximum variation of $f(x)$ on each U_i is less than δ. Now, we complete the proof exactly as we completed the proof in Part (ii) of Theorem 2.1; that is, substitute $C, F, C_\varepsilon, F_\varepsilon, U_i, U_i^\varepsilon$ for $B, G, B_\varepsilon, G_\varepsilon, S_i, S_i^\varepsilon$ there. Q.E.D.

4. Inequality constraints: Algorithm 4. We now consider another algorithm for minimizing $f(x)$, $x \in C \cap B$. The inequalities are handled by converting them to equalities, via the common technique of adding a slack variable. Let $z \equiv (z^1, \cdots, z^t)$, denote a vector of t nonnegative real variables and define $\phi_i(z) = q_i(x) + z^i$, $i = 1, \cdots, t$, with $\phi_{t+1}(\cdot), \cdots, \phi_{t+s}(\cdot)$ denoting the original s equality constraints. For $i > t$, we may write either $\phi_i(x)$ or $\phi_i(x, z)$. The function $\Phi(\cdot)$, the Jacobian of $\phi(x, z)$ with respect to x, (and $\Phi_n \equiv \Phi(X_n)$) is defined exactly as in § 2. Note that $\phi(\cdot)$ is a $(t + s) \times r$ matrix. For notational simplicity, we draw the random variables from $\hat{H}(\cdot|x)$, rather than from $H(\cdot|x)$, although, here too, there is an obvious finite difference analogue. Define $P(x, z) = \sum_{i=1}^{s+t}\phi_i^2(x, z)$, and let $\{v_n\}, \{w_n\}$ be sequences of positive real numbers tending to zero.

ALGORITHM 4. We iterate on both variables x and z. Assume X_0, Z_0 are given. The iterates $\{X_n\}$ and, in certain cases, the iterates $\{Z_n\}$, are computed exactly as $\{X_n\}$ would be in Algorithm 1. $\{\lambda_n\}$ will be defined below. Define (the observation vector is $Y_n \equiv f_x(X_n) + \xi_n$, as in § 2)

$$X_{n+1} = X_n - a_n\left[f_x(X_n) + \xi_n + \Phi_n'\lambda_n + \frac{k}{2}P_x(X_n, Z_n)\right]. \tag{4.1}$$

If $Z_n^i > v_n$, define (λ_n^i is the ith component of λ_n)

$$Z_{n+1}^i = \max [0, Z_n^i - a_n(\lambda_n^i + k\phi_i(X_n, Z_n))] \tag{4.2a}$$

and define α_i^n by $Z_{n+1}^i - Z_n^i \equiv \alpha_i^n[-a_n(\lambda_n^i + k\phi_i(X_n, Z_n))]$. If $Z_n^i \leqq v_n$, use (4.2b, c, d)

$$Z_{n+1}^i = \max [0, Z_n^i - a_n(\lambda_n^i + k\phi_i(X_n, Z_n))] \quad \text{if} \quad \phi_i(X_n, Z_n) \leqq 0 \quad \text{and} \quad \lambda_n^i + k\phi_i(X_n, Z_n) \leqq 0, \tag{4.2b}$$

$$Z_{n+1}^i = Z_n^i \quad \text{if} \quad \phi_i(X_n, Z_n) \leqq 0 \quad \text{and} \quad \lambda_n^i + k\phi_i(X_n, Z_n) > 0, \tag{4.2c}$$

$$Z_{n+1}^i = Z_n^i + a_n w_n \quad \text{if} \quad \phi_i(X_n, Z_n) > 0. \tag{4.2d}$$

Define the $(t + s) \times t$ matrices $I(z, v)$ (resp., J) as follows. All entries are zero except that the (i, i)th entry ($i \leqq t$) takes the value 1 if $z^i > v$ (resp., always takes the value 1). Denote $I_n \equiv I(z_n, v_n)$ and $\tilde{\Phi}(x, z, v) = [\Phi(x), I(z, v)]$. $\tilde{\Phi}(x, z, v)$ (a $(t + s) \times (t + r)$ matrix) is the Jacobian of $\phi(x, z)$ where we use $\partial\phi_i(x, z)/\partial z^i = 1$ only if $z^i > v$ (i.e., we exclude the $\partial\phi_i(x, z)/\partial z^i$ term for the "v active constraints": q^i is said to be a v active constraint at x if, "loosely speaking", $z^i \leqq v$.

Define $\begin{bmatrix} f_x(x) \\ 0 \end{bmatrix} = \tilde{f}_x(x)$, $\begin{bmatrix} \xi_n \\ 0 \end{bmatrix} = \tilde{\xi}_n$, where the zeros are t-dimensional vectors. The multiplier is chosen by forcing it to satisfy an orthogonality relationship like (2.3). In particular, we let λ_n be a λ minimizing the norm of the estimated gradient of a particular Lagrangian, namely, a minimizer in

$$|\tilde{f}_x(X_n) + \tilde{\xi}_n + \tilde{\Phi}_n'\lambda|^2 = |f_x(X_n) + \xi_n + \Phi_n'\lambda|^2 + \sum_{i: z^i > v_n} (\lambda^i)^2. \tag{4.3}$$

The minimization of (4.3) yields (2.3), (2.4) with $\tilde{\Phi}_n$, $\tilde{f}_x(X_n)$, $\tilde{\xi}_n$ replacing Φ_n, $f_x(X_n)$, ξ_n there. The choice of (4.3) as the quantity to minimize is motivated by the fact that if x is a constrained minimum of $f(x)$, then the Kuhn–Tucker condition is

$$f_x(x) + \sum_{t \geqq i \text{ active}} \lambda^i q_{i,x}(x) + \sum_{i=t+1}^{t+s} \lambda^i \phi_{i,x}(x) = 0, \qquad \lambda^i \geqq 0 \quad \text{for} \quad i \leqq t.$$

and so in (4.3), we seek to penalize the use of the "v_n inactive" constraints.

For any vector $\tilde{h}$ define $\tilde{\pi}(x, z, v)\tilde{h}$ as $\pi(x)h$ was defined in § 2, where $\tilde{\Phi}(x, z, v)$, $\tilde{h}$ replace $\Phi(x)$, h there, and write $\tilde{\pi}_n\tilde{h} \equiv \tilde{\pi}(X_n, Z_n, v_n)\tilde{h}$. If λ_n is to minimize (4.3), then it must satisfy

$$0 = \tilde{\Phi}_n[\tilde{f}_x(X_n) + \tilde{\xi}_n + \tilde{\Phi}_n'\tilde{\lambda}_n], \tag{4.4}$$

and we choose

$$\lambda_n = -[\tilde{\Phi}_n\tilde{\Phi}_n']^{-1}\tilde{\Phi}_n\tilde{f}_x(X_n) - [\tilde{\Phi}_n\tilde{\Phi}_n']^{-1}\tilde{\Phi}_n\tilde{\xi}_n \equiv \bar{\lambda}_n + \hat{\lambda}_n. \tag{4.5}$$

For each real $\varepsilon > 0$, define the sets $C_\varepsilon^+, F_\varepsilon^+$ and let $C_0^+ = C^+$, $F_0^+ = F^+$: when q_i is used, then $i \leqq t$)

$$C_\varepsilon^+ = \{(x, z) : z^i \geqq 0, \quad \text{all} \quad i, |\phi(x, z)|^2 \leqq \varepsilon\},$$

$$F_\varepsilon^+ = \left\{(x, z) : \min_\lambda \left(\left| f_x(x) + \sum_i \lambda^i q_{i,x}(x) + \sum_{i=t+1}^{t+s} \lambda^i \phi_{i,x}(x) \right|^2 + \sum_{q_i(x) < 0} (\lambda^i)^2 \right) \leqq \varepsilon \right\}:$$

Note that in C^+, $q(x) = -z$, so x specifies z, and we can unambiguously speak of x in C^+. The points satisfying the Kuhn–Tucker necessary condition are the points in $C^+ \cap F^+$ for which there are nonnegative minimizing λ^i, $i \leqq t$. For $x \in C^+ \cap F^+$, define $\bar{\lambda}(x)$ (with components $\bar{\lambda}^i(x)$) to be the multiplier attaining the minimum in the definition of F^+ (if it is not unique, use the one determined by the appropriate pseudoinverse: i.e., the one of minimum norm). Now $C^+ \cap F^+$ is closed and can be written as the union of a collection of disjoint closed and connected sets $T_1, \cdots$ on each of which $f(\cdot)$ is constant taking, say, the value f_i on T_i. We need the following assumptions:

(A3‴) The rows of $\Phi(x)$ are linearly independent for each $x \in R^r$.

(A7) For a real number $l > 0$ for which $\sup_n E|\xi_n|^l < \infty$, $\sum_n (a_n/v_n)^l < \infty$.

(A8) $\sum_n a_n w_n = \infty$.

(A9) For each T_j, if, for some $i \leqq t$, we have $\bar{\lambda}^i(x) < 0$ (resp., > 0) at a point $x \in T_j$, then $\bar{\lambda}^i(x) < 0$ (resp., > 0) at all $x \in T_j$.

(A10) There are a finite number of sets T_j.

(A11) Suppose that for some i, j, $\bar{\lambda}^i(\bar{x}) < 0$ on some T_j, and let $\{\bar{x}_n, \bar{z}_n\}$ denote an arbitrary sequence in the δ-neighborhood $N_\delta(T_j) \equiv \{y : |y - u| < \delta$, some $u \in T_j\}$, which tends to $(\bar{x}, \bar{z})$ where $\bar{z} = -q(\bar{x})$. Suppose that there are positive real numbers $\tilde{v}, \delta_1, \delta$, so that ψ_n^i, the ith element of the minimizing ψ in

$$\min_{\psi} (|f_x(\bar{x}_n) + \Phi'(\bar{x}_n)\psi|^2 + \sum_{z_n^i > v_n} (\psi^j)^2) \qquad v_n \leqq \tilde{v},$$

is less than $-\delta_1$ for any such $\{\bar{x}_n, \bar{z}_n\}$ sequence.

Assumption (A9) is apparently not restrictive in applications. It basically eliminates the possibility that a T_i contains both saddle-like points together with other points which are neither saddles nor local maxima nor local minima. For an example of the type of situation excluded by (A9), consider $x = (x^1, x^2)$, $f(x) = x^1 x^2$, $q_1(x) = x^2$. On the boundary determined by $q_1 = 0$ (denote it as T_1), $f(x) = 0$ and $f'_x(x) = (0, x^1)$, $q'_{1,x}(x) = (0, 1)$, and $\bar{\lambda}(x) = -x^1$. If we add smooth constraints which bound $|x^1|$ and x^2 from below, the algorithm can be shown to converge without (A9), for this type of problem. We strongly suspect that (A9) can be dispensed with, but cannot prove it, as yet.

Condition (A11) may seem a little strange. If $\bar{\lambda}^i < 0$, $x \in T_j$, we will require that the $\bar{\lambda}_n^i < 0$ for large n and X_n near T_j. But the "discontinuous" term $\sum_{Z_n^i > v_n} (\lambda^j)^2$ creates discontinuities in the $\bar{\lambda}_n^i$, as the Z_n^i vary above and below v_n. If there is only one active constraint in T_j, the condition is no restriction, nor is it a restriction if the $q_{i,x}(\cdot)$ for all "nearly active" constraints are constant or nearly orthogonal (as for example, if all q_j were of the form $q_j(x) = \alpha_{jl} x_l + \beta_{jl}$). Assumption (A11) is implied by (A9) if the signs of the $\bar{\lambda}^i(x)$ (when >0 or <0) are the signs of $-q'_{i,x}(x) f_x(x)$. The condition can be weakened by using $\phi_i(x, z) = q_i(x) + bz^i$ for small b ($i \leqq t$) in lieu of $q_i(x) + z^i$. We then need to multiply the $\sum (\lambda^j)^2$ term by b, and the "ones" in I_n and J by b, and multiply the λ_n^i in (4.2a, b) by b. The smaller b is, the less restrictive is the corresponding condition (A11). The condition is needed only to show that $\psi^i \geqq 0$, $i \leqq t$, in Theorem 4.1.

THEOREM 4.1. *Assume* (A1)–(A2), (A3‴), (A7)–(A11). *Then there is a null set* N *so that if* $\omega \notin N$, *and* $\sup_n |X_n(\omega)| < \infty$ *and* x *is any limit of a convergent subsequence*

of $\{X_n(\omega)\}$, *then* $\phi(x) = 0$, $i > t$, $q_i(x) \leqq 0$, $i \leqq t$, *and there is a vector (perhaps depending on* x*)* $\psi = (\psi^1, \cdots, \psi^{t+s})$, $\psi_i \geqq 0$, $i = 1, \cdots, t$, *for which*

$$f_x(x) + \sum_{i, q_i(x)=0} \psi^i q_{i,x}(x) + \sum_{i=t+1}^{t+s} \psi^i \phi_{i,x}(x) = 0. \tag{4.6}$$

(Equation (4.6) is the Kuhn–Tucker necessary condition for constained optimality.)

Remark. A condition analogous to that mentioned after the statement of Theorem 2.1 yields that $\sup_n |X_n| < \infty$ w.p.1.

Proof. As in the proof of Theorem 2.1, we can and will suppose that $\sup_n |X_n| \leqq M < \infty$ w.p.1 for some real M. Also, for notational simplicity, we ignore the constraints $\phi_{t+1}(\cdot), \cdots, \phi_{t+s}(\cdot)$. The general proof is almost the same as the one given below.

Part (i). Letting $P_{x,z}(x, z)$ denote the gradient of $P(x, z)$ with respect to (x, z), we get $P_{x,z}(x, z) = k[\Phi(x), J]'\phi(x, z)$. Let α_n denote the diagonal matrix with entries α_n^i. If (4.2a) is used to calculate Z_{n+1}^i and is untruncated, then $\alpha_n^i = 1$. Using (4.1)–(4.2), a Taylor series expansion and majorization of some of the second order terms yields

$$\begin{aligned} &P(X_{n+1}, Z_{n+1}) - P(X_n, Z_n) \\ &\quad \leqq -a_n P'_{x,z}(X_n, Z_n)\left[\tilde{f}_x(X_n) + \tilde{\xi}_n + \begin{pmatrix} \Phi'_n \\ \alpha_n I'_n \end{pmatrix} \lambda_n + k\begin{pmatrix} \Phi'_n \\ \alpha_n I'_n \end{pmatrix} \Phi(X_n, Z_n)\right] \\ &\quad\quad - a_n k \sum_{(4.2b)} \phi_i(X_n, Z_n)[\lambda_n^i + k\phi(X_n, Z_n)] + a_n w_n \sum_{(4.2d)} \phi_i(X_n, Z_n) \\ &\quad\quad + a_n^2 K[|f_x(X_n)|^2 + |\xi_n|^2 + |\phi(X_n, Z_n)|^2 + w_n^2], \end{aligned} \tag{4.7}$$

where (4.2b) or (4.2d) implies that the summation is over those i for which (4.2b) or (4.2d) are used at iteration n. If (4.2a) is used and truncated, then (here we have $Z_n^i > v_n$)

$$a_n[\lambda_n^i + k(q_i(X_n) + Z_n^i)] \geqq Z_n^i \quad \text{or} \quad v_n(1 - a_n k) \leqq a_n \bar{\lambda}_n^i + a_n \hat{\lambda}_n^i + a_n^k q_i(X_n). \tag{4.8}$$

By $\sup_n |X_n| \leqq M$, $q_i(X_n)$ is bounded. The fact that the minimum (4.3) is $\leqq |f_x(X_n) + \xi_n|^2$ and the definitions of $\hat{\lambda}_n^i, \bar{\lambda}_n^i$ imply that there is a real K for which $|\bar{\lambda}_n^i| \leqq K|f_x(X_n)|, |\hat{\lambda}_n^i| \leqq K|\xi_n|$. These facts, together with (4.8) and (A7) and the Borel–Cantelli lemma, imply that (4.2a) will be truncated only finitely often w.p.1. Neglecting this "finitely often occurring" event will not affect the rest of the proof, and we will do so. Then setting $\alpha_n =$ identity and using $JI'_n = I_n I'_n$ and (4.4), (4.5), (4.7) we get

$$\begin{aligned} P(X_{n+1}, Z_{n+1}) - P(X_n, Z_n) \leqq\ & -a_n K|\tilde{\Phi}'(X_n, Z_n, v_n)\phi(X_n, Z_n)|^2 \\ & + a_n w_n \sum_{(4.2d)} \phi_i(X_n, Z_n) \\ & + a_n^2 K[|f_x(X_n)|^2 + |\xi_n|^2 \\ & + |\phi(X_n, Z_n)|^2 + |w_n^2|]. \end{aligned} \tag{4.9}$$

Using the fact that $w_n \to 0$, an analogy of the argument of Part (i) (but (A3‴) instead of (A3)) of the proof of Theorem 2.1, yields that $(X_n, Z_n) \in C_\varepsilon^+$ infinitely often w.p.1 for each $\varepsilon > 0$, and that $\phi(X_n, Z_n) \to 0$ w.p.1 as $n \to \infty$. Since $Z_n^i \geqq 0$, this implies that $\phi_i(x, z) = 0$, $q_i(x) \leqq 0$ for any limit point (x, z) of $\{X_n(\omega), Z_n(\omega)\}$ (for ω not in some null set). Note that if $X_{n_j}(\omega) \to x$, then $Z_{n_j}^i(\omega) \to -q_i(x)$.

Part (ii). Now we evaluate $f(X_{n+1}) - f(X_n)$:

$$f(X_{n+1}) - f(X_n) \leqq -a_n f_x'(X_n)\left[f_x(X_n) + \xi_n + \Phi_n'\lambda_n + \frac{k}{2}P_x(X_n)\right] + a_n^2 K[|f_x(X_n)|^2 + |\xi_n|^2 + |\phi(X_n, Z_n)|^2]. \tag{4.10}$$

It is helpful to rewrite (4.10) as

$$f(X_{n+1}) - f(X_n) \leqq -a_n \tilde{f}_x'(X_n)\left[\tilde{f}_x(X_n) + \tilde{\xi}_n + \tilde{\Phi}_n'\lambda_n + \frac{k}{2}P_{x,z}(X_n)\right] + a_n^2 K[|f_x(X_n)|^2 + |\xi_n|^2 + |\phi(X_n, Z_n)|^2]. \tag{4.11}$$

Equivalently, using (4.5) (see above (4.4) for the definition of $\tilde{\pi}$, $\tilde{\pi}_n$) we get

$$\begin{aligned} f(X_{n+1}) - f(X_n) \leqq & -a_n \tilde{f}_x'(X_n)[\tilde{\pi}_n \tilde{f}_x(X_n) + \tilde{\pi}_n \tilde{\xi}_n] \\ & -a_n \frac{k}{2} f_x'(X_n)P_x(X_n, Z_n) + a_n^2 K[|f_x(X_n)|^2 + |\xi_n|^2 \\ & + |\phi(X_n, Z_n)|^2]. \end{aligned} \tag{4.12}$$

The first term on the right-hand side of (4.12) can be written as

$$-a_n|\tilde{\pi}_n \tilde{f}_x(X_n)|^2 - a_n \tilde{f}_x'(X_n)\pi_n \tilde{\xi}_n. \tag{4.13}$$

Recall that

$$|\tilde{\pi}_n \tilde{f}_x(X_n)|^2 = \min_\lambda \left(|f_x(X_n) + \sum_i \lambda^i q_{i,x}(X_n)|^2 + \sum_{Z_n^i > v_n} (\lambda^i)^2\right) \tag{4.14}$$

and that λ_n is a minimizing λ in (4.14). Let $\{\bar{x}_n, \bar{z}_n\}$ denote any sequence with all $\bar{z}_n^i \geqq 0$ and limit $\bar{x}, \bar{z}$ and $q^i(\bar{x}) + \bar{z}^i = 0$, $i \leqq t$. Note that

$$\tilde{\pi}(\bar{x}_n, \bar{z}_n, v_n)\tilde{f}_x(\bar{x}_n) \to 0 \Rightarrow (\bar{x}, \bar{z}) \in F^+. \tag{4.15}$$

Equations (4.12), (4.13) and (4.15) and the summability of both $\sum_n a_n \tilde{f}_x'(X_n)\tilde{\pi}_n \tilde{\xi}_n$, $\sum_n a_n^2|\xi_n|^2$ and $P_x(X_n, z_n) \to 0$ and $\sum_n a_n \to \infty$ imply that $(X_n, Z_n) \in F_\varepsilon^+$ infinitely often w.p.1 for each $\varepsilon > 0$ (for otherwise the sum over n of the right-hand side of (4.12) would tend to ∞).

Thus $\{X_n, Z_n\} \in F_\varepsilon^+ \cap C_\varepsilon^+$ infinitely often w.p.1. Also $F_\varepsilon^+ \cap C_\varepsilon^+$ tends to the closed set $F^+ \cap C^+$ as $\varepsilon \to 0$. Given small $\delta > 0$, there is an $\varepsilon > 0$ so that we can write $C_\varepsilon^+ \cap F_\varepsilon^+ = \cup T_i^\varepsilon$, where T_i^ε is closed and connected, contains T_i and the $\{T_i^\varepsilon\}$ are disjoint. We can suppose that the maximum variation of $f(x)$ on each T_i^ε is less than δ and that if $f_i \neq f_j$, then $|f_i - f_j| \geqq 3\delta$.

Using the technique of Part (ii) of the proof of Theorem 2.1, we can show that $\{f(X_n)\}$ converges and that (X_n, Z_n) is in $F_\varepsilon^+ \cap C_\varepsilon^+$ for large n w.p.1 for any fixed $\varepsilon > 0$. Furthermore (again using the same idea as in the proof of Theorem 2.1) we can show that $(X_n, Z_n) \to F^+ \cap C^+$ as $n \to \infty$ w.p.1, which is the desired result (4.6), except for the nonnegativity of the ψ^i, $i \leqq t$.

Part (iii). We can suppose that $f_x(x) \neq 0$ at the limit points, for otherwise we can take $\psi^i = 0$ for all i. Thus we only need to consider limits points x ((x, z), resp.) on the boundary of C (C^+, resp.). By (A9), and the fact that the T_j are closed (and we can suppose bounded, since $\sup_n |X_n(\omega)| < \infty$), for any i, j, either $\bar{\lambda}^i(x) \equiv 0$ on T_j or $\inf_{x \in T_j} \bar{\lambda}^i(x) > 0$ or $\sup_{x \in T_j} \bar{\lambda}^i(x) = l_{ij} < 0$. *Fix* T_j, and suppose that $l_{ij} > 0$. This implies that $q_i(\cdot)$ is active on T_j.

By Parts (i) and (ii), for each $\varepsilon > 0$, $\{X_n, Z_n\}$ are ultimately in $F_\varepsilon^+ \cap C_\varepsilon^+$; hence for any $\delta > 0$, the sequence is ultimately in $\bigcup_i N_\delta(T_k)$, and we can suppose that δ is small enough so that the $\{\bar{N}_\delta(T_k)\}$ are disjoint and $\bar{N}_\delta(T_k) \supset T_k$. So for any small $\delta > 0$, the tail of $\{X_n, Z_n\}$ is (w.p.1) contained in one of (which one depends on ω) the $\{\bar{N}_\delta(T_k)\}$. Let $\{X_n, Z_n\} \to T_j$ then (for $\omega \notin$ a null set) (A11) implies that $\bar{\lambda}_n^i \leqq -\bar{l} < 0$ for some real $\bar{l}_i > 0$, and all large n, and $\phi(X_n, Z_n) \to 0$ as $n \to \infty$. For this $\{X_n, Z_n\}$ sequence, (4.16a) holds for large n when $Z_n^i > v$.

$$Z_{n+1}^i \geqq Z_n^i + a_n \bar{l}_i/2 - a_n \hat{\lambda}_n^i. \tag{4.16a}$$

If $Z_n^i \leqq v_n$, then Z_n^i cannot decrease and may increase, but v_n decreases; ultimately $v_n > Z_n^i$. Combining (4.2b, c) we get (for large n)

$$Z_{n+1}^i \geqq Z_n^i + [a_n \bar{l}_i/2 + a_n[\hat{\lambda}_n^i + k\phi_i(X_n, Z_n)]I_{\{\lambda_n^i + k\phi_i(X_n, Z_n) \leqq 0\}}. \tag{4.16b}$$

If (4.2d) holds.

$$Z_{n+1}^i = Z_n^i + a_n w_n. \tag{4.16c}$$

In (4.16a, b), we can replace $\bar{l}_i/2$ by w_n. (4.16a, b, c), the fact that $\sum_n a_n \hat{\lambda}_n^i$ is a square summable (hence convergent) martingale, and (A8) imply that $Z_n^i \to \infty$, contradicting the fact that $q_i(\cdot)$ is active in T_j. Thus all $\bar{\lambda}^i(x) \geqq 0$ on T_j, if $\{X_n, Z_n\} \in N_\delta(T_j)$ infinitely often, for each $\delta > 0$, as desired. □ Q.E.D.

Note added in proof. The conditions requiring square summability of $\{a_n\}$ and orthogonality of $\{\xi_n\}$ have been considerably weakened. See *General convergence theorems for stochastic approximation via weak convergence* by the first author, to appear. Also, numerical experiments indicate that the algorithms work reasonably well, with appropriate parameter selections.

REFERENCES

[1] H. J. KUSHNER, *Stochastic approximation algorithms for constrained optimization problems*, Ann. Statist., 2 (1974), pp. 713–723.

[2] H. J. KUSHNER AND H. T. GAVIN, *Stochastic approximation-like algorithms for constrained systems: Algorithms and numerical results*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 349–357.

[3] H. J. KUSHNER AND H. E. SANVICENTE, *Penalty function methods for constrained stochastic approximation*, J. Math. Anal. Appl., 46 (1974), pp. 499–512.

[4] A. MIELE, E. G. CRAGG, R. R. IYER AND A. V. LEVY, *Use of augmented penalty function in mathematical programming problems, Part I*, J. Optimization Theory Appl., 8 (1971), pp. 115–130.
[5] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.
[6] D. P. BERTSEKAS, *Combined primal-dual and penalty methods for constrained minimization*, this Journal, 13 (1975), pp. 521–544.
[7] R. T. ROCKAFELLAR, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.
[8] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York, 1971.

J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

LAGRANGE DUALITY THEORY FOR CONVEX CONTROL PROBLEMS*

WILLIAM W. HAGER† AND SANJOY K. MITTER‡

Abstract. The Lagrange dual of control problems with linear dynamics, convex cost and convex inequality state and control constraints is analyzed. If an interior point assumption is satisfied, then the existence of a solution to the dual problem is proved; if there exists a solution to the primal problem, then a complementary slackness condition is satisfied. A necessary and sufficient condition for feasible solutions in the primal and dual problems to be optimal is also given. The dual variables p and v corresponding to the system dynamics and state constraints are proved to be of bounded variation while the multiplier corresponding to the control constraints is proved to lie in $\mathscr{L}^1$. Finally, a control and state minimum principle is proved. If the cost function is differentiable and the state constraints have two derivatives, then the state minimum principle implies that a linear combination of p and v satisfy the conventional adjoint condition for state constrained control problems.

1. Introduction. The Lagrange dual of the following control problem is studied:

$$\inf c(x, u)$$

$$\text{subject to } \dot{x}(t) = A(t)x(t) + B(t)u(t), \qquad x(0) = x_0,$$

$$K_c(u(t), t) \leqq 0, \qquad K_s(x(t), t) \leqq 0,$$

where $c(\cdot, \cdot)$, $K_c(\cdot, t)$ and $K_s(\cdot, t)$ are all convex. Rockafellar [7] has derived duality results for convex state constrained control problems using Fenchel duality theory. The development in this paper goes beyond Rockafellar's results since the constraints are given explicitly by inequalities above, and hence the multipliers associated with the constraints can be characterized. Also, a slightly different form of the dual problem, the Lagrange dual, is studied herein; and the matrix $B(t)$ above, which Rockafellar assumes is the identity matrix in his development, is introduced. The theory in this paper provides the foundation for an analysis of the numerical solution of the dual problem by the Ritz method in [1]. The control problem stated above involves no constraints on $x(0)$ and $x(1)$ except for the condition $x(0) = x_0$; however, convex inequality and linear equality endpoint constraints could have been included with very little change in the analysis. To keep the presentation simpler, these constraints are not explicitly treated; however, notice that the state constrained problem explicitly involves endpoint restrictions because of the condition $K_s(x(t), t) \leqq 0$ for all $t \in [0, 1]$.

In §§2 and 3, the principal result based on the Hahn–Banach theorem, proves that the dual problem has an optimal solution if there exists an interior point for the constraint set (i.e., the Slater condition holds); if the primal problem has an optimal solution, then a complementary slackness condition holds. The optimal multipliers $\hat{p}$ and $\hat{v}$ corresponding to the system dynamics and state

* Received by the editors August 22, 1974, and in revised form May 15, 1975.

† Department of Mathematics, University of South Florida, Tampa. Florida 33620.

‡ Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

The second author was supported in part by the National Science Foundation under Grant GK 41647, by the U.S. Air Force under Grant AFOSR 72 2273, and by NASA under Grant NASA/AMES NGL 22-009-124.

constraints are shown to have bounded variation while the multiplier $\hat{w}$ corresponding to the control constraints lies in L^1. Also a necessary and sufficient condition for the optimality of solutions to the primal and the dual problem is given.

Section 4 then proves that a minimum principle holds, and while $(\hat{p}, \hat{v})$ are only of bounded variation, the combination $\hat{q}(t) = K_s(\hat{x}(t), t)_x^T \hat{v}(t) - \hat{p}(t)$ is absolutely continuous where $\hat{x}$ solves the primal problem; furthermore $\hat{q}$ satisfies the conventional adjoint equation for state constrained control problems. This result has important consequences for the solution of the dual problem using the Ritz method in [1] since the convergence rate of the discrete approximation depends upon the smoothness of the dual variables; hence if the dual problem is reformulated in terms of q rather than p, then a superior convergence rate is achieved.

The Appendix contains several technical lemmas concerning the regularity of the dual variables.

Notation. The following notation is used for spaces of real-valued functions on [0, 1]:

$\mathcal{A}$	absolutely continuous functions,
$\mathcal{BV}$	functions of bounded variation continuous from the left on [0, 1),
$\mathcal{NBV}$	functions of bounded variation continuous from the left on [0, 1), and normalized so that $f(1) = 0$,
$\mathcal{C}$	continuous functions,
$\mathcal{L}^p$	functions with $\int_0^1 \lvert f(t)\rvert^p \, dt < \infty$,
$\mathcal{L}^\infty$	functions essentially bounded and measurable.

If $\mathcal{W}$ is any of the spaces above, the notation $x \in \mathcal{W}(R^n)$ means that x is a vector-valued function with n components and each component lies in $\mathcal{W}$.

If $y \in R^m$, then define $\lvert y\rvert = \sum_{k=1}^m \lvert y_k\rvert$ and denote the supremum norm of a vector-valued function by $\|f\| = \sup_{t\in[0,1]} \lvert f(t)\rvert$.

If $x, y \in R^m$, the inner product $(\cdot\,, \cdot)$ is defined by $(x, y) = \sum_{k=1}^m x_k y_k$. If $f \in \mathcal{L}^p$, $g \in \mathcal{L}^q$, where $\mathcal{L}^q$ is the dual of $\mathcal{L}^p$, $v \in \mathcal{BV}$, and $h \in \mathcal{C}$, then define:

$$\langle f, g\rangle = \int_0^1 (f(t), g(t))\, dt, \qquad [v, h] = \int_0^1 h(t)\, dv(t).$$

The *complement* and *closure* of a set are denoted A^c and $\bar{A}$, respectively.

2. Duality theory. The following *control problem* is considered:

$$\text{(P)} \qquad \begin{aligned} &\inf c(x, u) \\ &\text{subject to } c(x, u) = \int_0^1 h(x(t), u(t), t)\, dt, \\ &\dot{x}(t) = A(t)x(t) + B(t)u(t), \qquad x(0) = x_0, \\ &K_c(u(t), t) \leqq 0, \qquad K_s(x(t), t) \leqq 0 \quad \text{for all } t \in [0, 1], \\ &x \in \mathcal{A}(R^n), \qquad u \in \mathcal{L}^\infty(R^m), \end{aligned}$$

where h, K_c and K_s have range in R, R^{m_c} and R^{m_s}, respectively, and the matrices A and B are of the appropriate dimensions. Note that in the control problem

above, the controls lie in $\mathscr{L}^\infty$. The next section will treat the case where the controls lie in $\mathscr{L}^1$. The *dual function* L is given by

$$
\begin{aligned}
&L(p, w, v) = \inf\{c(x, u) + \langle p, \dot{x} - Ax - Bu\rangle + \langle w, K_c(u)\rangle + [v, K_s(x)]\} \\
&\text{subject to } x(0) = x_0, \qquad x \in \mathscr{A}(R^n), \quad u \in \mathscr{L}^\infty(R^m).
\end{aligned} \tag{1}
$$

The dual problem corresponding to (P)

$$
\begin{aligned}
&\sup L(p, w, v) \\
&\text{subject to } p \in \mathscr{BV}(R^n), \quad v \in \mathscr{NBV}(R^{m_s}), \quad w \in \mathscr{L}^1(R^{m_c}), \quad w \geqq 0, \\
&v \text{ nondecreasing.}
\end{aligned} \tag{D}
$$

In order that all the terms in (P) and (1) above make sense, assumptions must be made concerning the functions appearing in these problems. Theorem 1 will require the following, continuity, convexity and Slater conditions:

(C) $h(\cdot, \cdot, t)$, $K_s(\cdot, t)$ and $K_c(\cdot, t)$ are convex for $t \in [0, 1]$, $A(\cdot)$ and $B(\cdot)$ have components in $\mathscr{L}^1$ and $h(\cdot, \cdot, \cdot)$, $K_s(\cdot, \cdot)$ and $K_c(\cdot, \cdot)$ are all continuous.

(SL) There exists a control $\bar{u} \in \mathscr{C}(R^m)$ and a corresponding trajectory $\bar{x}$ such that $(K_c(\bar{u}(t), t))_j < a < 0$ and $(K_s(\bar{x}(t), t))_j < a < 0$ for some "a", for all $t \in [0, 1]$ and for all components of K_c and K_s.

Proposition 1 below, the *weak duality theorem*, is easily verified. This is followed by the principal theorem, or *strong duality* result.

PROPOSITION 1. *$c(x, u) \leqq L(p, w, v)$ whenever (x, u) are feasible in* (P) *and (p, w, v) are feasible in* (D).

THEOREM 1. *Suppose* (C) *and* (SL) *hold and the optimal value, $\hat{c}$, of* (P) *is finite. Then there exist $(\hat{p}, \hat{w}, \hat{v})$ that are optimal in* (D) *with $L(\hat{p}, \hat{w}, \hat{v}) = \hat{c}$. Furthermore, if $(\tilde{p}, \tilde{w}, \tilde{v})$ and $(\tilde{x}, \tilde{u})$ are feasible in* (D) *and* (P), *respectively, then a necessary and sufficient condition for $(\tilde{p}, \tilde{w}, \tilde{v})$ and $(\tilde{x}, \tilde{u})$ to be optimal solutions to the dual and primal problems is that $(\tilde{x}, \tilde{u})$ achieve the minimum in* (1) *for $(p, w, v) = (\tilde{p}, \tilde{w}, \tilde{v})$ and the complementary slackness conditions $\langle \tilde{w}, K_c(\tilde{u})\rangle = [\tilde{v}, K_s(\tilde{x})] = 0$ hold.*

Observe that the condition $\langle \tilde{w}, K_c(\tilde{u})\rangle = [\tilde{v}, K_s(\tilde{x})] = 0$ implies that $K_c(\tilde{u}(t), t)_j = 0$ whenever $\tilde{w}(t)_j > 0$ a.e. and $\tilde{v}_j$ is constant on every interval where $K_s(\tilde{x}(t), t) < 0$. Also notice that the sufficiency condition follows immediately from complementary slackness, feasibility of $(\tilde{x}, \tilde{u})$ and $(\tilde{p}, \tilde{w}, \tilde{v})$, the optimality of $(\tilde{x}, \tilde{u})$ in (1) for $(p, w, v) = (\tilde{p}, \tilde{w}, \tilde{v})$ and Proposition 1; that is, $c(\tilde{x}, \tilde{u}) = L(\tilde{p}, \tilde{w}, \tilde{v})$, and this can only happen when $(\tilde{x}, \tilde{u})$ and $(\tilde{p}, \tilde{w}, \tilde{v})$ are optimal in (P) and (D), respectively. On the other hand, if $(\tilde{p}, \tilde{w}, \tilde{v})$ and $(\tilde{x}, \tilde{u})$ are optimal in (D) and (P) and it can be proved that the optimal value of the primal and dual problems are equal, then $c(\tilde{x}, \tilde{u}) = L(\tilde{p}, \tilde{w}, \tilde{v}) \leqq c(\tilde{x}, \tilde{u}) + \langle \tilde{w}, K_c(\tilde{u})\rangle + [\tilde{v}, K_s(\tilde{x})]$. Since $K_c(\tilde{u}(t), t) \leqq 0$, $\tilde{w}(t) \geqq 0$, $K_s(\tilde{x}(t), t) \leqq 0$ and $\tilde{v}$ is nondecreasing, $\langle \tilde{w}, K_c(\tilde{u})\rangle = 0$, $[\tilde{v}, K_s(\tilde{x})] = 0$ and $(\tilde{x}, \tilde{u})$ achieve the minimum in (1) for $(p, w, v) = (\tilde{p}, \tilde{w}, \tilde{v})$. Thus the proof of Theorem 1 will be complete if it can be shown that the optimal value of the dual problem and the primal problem are equal whenever (SL) and (C) hold and the value of the primal problem is finite.

Rather than prove directly that the optimal value of the primal and dual

problem are equal, we first consider a slightly more general problem:

$$\text{(P')} \quad \begin{aligned} &\inf f(x, u) \\ &\text{subject to } \dot{x}(t) = A(t)x(t) + B(t)u(t), \qquad x(0) \in X_0, \quad K_s(x(t), t) \leqq 0, \\ &u(t) \in U(t) \quad \text{for all } t \in [0, 1], \quad x \in \mathscr{A}(R^n), \quad u \in \mathscr{L}^\infty(R^m), \end{aligned}$$

where f is a functional defined on $\mathscr{A}(R^n) \times \mathscr{L}^\infty(R^m)$. The corresponding dual function is

$$\begin{aligned} L'(p, v) &= \inf \{f(x, u) + \langle p, \dot{x} - Ax - Bu \rangle + [v, K_s(x)]\} \\ &\text{subject to } x \in \mathscr{A}(R^n), \quad u \in \mathscr{L}^\infty(R^m), \quad x(0) \in X_0, \quad u(t) \in U(t) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for all } t \in [0, 1]. \end{aligned} \tag{2}$$

The dual problem is

$$\text{(D')} \quad \begin{aligned} &\sup L'(p, v) \\ &\text{subject to } p \in \mathscr{BV}(R^n), \quad v \in \mathscr{NBV}(R^{m_s}), \quad v \text{ nondecreasing.} \end{aligned}$$

Define $X = \{x \in \mathscr{A}(R^n)$: $K_s(x(t), t) \leqq 0$ for all $t \in [0, 1]\}$ and $U = \{u \in \mathscr{L}^\infty(R^m)$; $u(t) \in U(t)$ for all $t \in [0, 1]\}$, and make the following assumptions analogous to those above for problem (P).

(C') $f(\cdot, \cdot)$, $K_s(\cdot, t)$, $U(t)$ and X_0 are convex for all $t \in [0, 1]$, $K_s(\cdot, \cdot)$ is continuous, and both $A(\cdot)$ and $B(\cdot)$ have components in $\mathscr{L}^1$.

(SL') There exists a control $\bar{u} \in \mathscr{C}(R^m)$, a corresponding trajectory $\bar{x}$ and constants M, ρ, $\alpha > 0$ such that $\bar{u} \in U$, $\bar{x}(0) \in X_0$, $K_s(\bar{x}(t), t)_j < -\alpha < 0$ for all components of K_s, and $f(x, \bar{u}) < M$ whenever $\|x - \bar{x}\| \leqq \rho$.

LEMMA 1. *Suppose* (C') *and* (SL') *hold and* $\hat{c}$, *the optimal value of* (P'), *is finite. Then there exist* $(\hat{p}, \hat{v})$ *that are optimal in* (D') *and* $L'(\hat{p}, \hat{v}) = \hat{c}$. *If* $(\hat{x}, \hat{u})$ *are optimal in* (P'), *then* $[\hat{v}, K_s(\hat{x})] = 0$ *and hence* $(\hat{x}, \hat{u})$ *achieve the minimum in* (2) *for* $(\hat{p}, \hat{v})$.

Proof. Lemma 1 follows from an application of the Hahn–Banach theorem to the following two sets:

$$\begin{aligned} Y &= \{(a, b, c) : a \in R^1, b \in \mathscr{L}^1(R^n), c \in \mathscr{C}(R^{m_s}), a \leqq \hat{c}, b = 0, c \leqq 0\}, \\ Z &= \{(a, b, c) : a \in R^1, b \in \mathscr{L}^1(R^n), c \in \mathscr{C}(R^{m_s}) \text{ and there exists} \\ &\quad x \in \mathscr{A}(R^n) \text{ and } u \in U \text{ with } x(0) \in X_0, a \geqq f(x, u), \\ &\quad b(t) = \dot{x}(t) - A(t)x(t) - B(t)u(t), c(t) \geqq K_s(x(t), t) \text{ for all} \\ &\quad t \in [0, 1]\}. \end{aligned}$$

From the development of duality in the literature, it is obvious that two sets like Y and Z must be constructed, and the hyperplane separating the sets will define the optimal dual multipliers. Note though that the choice of the convex sets that are to be separated is a very delicate question since one set must have nonempty *interior* which is *disjoint* from the other set before the Hahn–Banach theorem can be employed. Also the sets must be chosen so that the dual multipliers are in "reasonable" spaces if the duality principle corresponding to the

sets is to generate a *numerically tractable problem.* It will be seen that Y and Z do indeed satisfy all these conditions and lead to the duality principle stated in the lemma.

The reader can readily verify that the convexity conditions in (C') imply that Y and Z are convex, the assumption (SL') implies that Z has an interior point, and the fact that $\hat{c}$ is the optimal value in (P') implies that Y and the interior of Z are disjoint. Thus by the Hahn–Banach theorem [4], there exists a hyperplane separating Z and Y, i.e., there exists $r \in R^1$, $p \in \mathscr{L}^\infty(R^n)$, $v \in \mathscr{NBV}(R^{m_s})$ such that

$$(r, a_1) + \langle p, b_1 \rangle + [v, c_1] \geqq (r, a_2) + \langle p, b_2 \rangle + [v, c_2] \tag{3}$$

for all $(a_1, b_1, c_1) \in \bar{Z}$, $(a_2, b_2, c_2) \in \bar{Y}$. By choosing particular points in Y and Z, properties of the separating hyperplane will be exhibited:

(a) $r \geqq 0$. Substitute $a_2 = \hat{c} - 1$, $a_1 = f(\bar{x}, \bar{u})$, $b_1 = b_2 = c_1 = c_2 = 0$ in (3) where $(\bar{x}, \bar{u})$ was given in (SL').

(b) *v is monotone nondecreasing.* For notational convenience, v is assumed scalar-valued, although for vector-valued functions the proof is identical.

Given $t, s, d \in [0, 1)$, $t < s$, $d < |s - t|$, let c_d denote the continuous piecewise linear function that is -1 on $[t, s-d]$, zero on $[0, t-d]$ and $[s, 1]$, and linear on $[t-d, t]$ and $[s-d, s]$. Now, $[v, c_d] = v(t) - v(s-d) + z_d$ where

$$|z_d| \leqq |TV(t, v) - TV(t-d, v)| + |TV(s, v) - TV(s-d, v)|$$

and $TV(t, v)$ is the total variation of v on $[0, t]$. Since v is continuous from the left on $[0, 1)$, then $TV(\cdot, v)$ is continuous from the left at t and s (see [6]), and hence $\lim_{d \to 0} |z_d| = 0$ and $\lim_{d \to 0} [v, c_d] = v(t) - v(s)$. Substituting $(\hat{c}, 0, 0) \in \bar{Z}$ and $(\hat{c}, 0, c_d) \in Y$ into (3) and letting $d \to 0$, we obtain $v(t) \leqq v(s)$. The right endpoint, $t = 1$, is treated similarly.

(c) *If $(\hat{x}, \hat{u})$ are optimal in* (P'), *then* $[v, K_s(\hat{x})] = 0$. Substitute $a_1 = a_2 = \hat{c}$, $b_1 = b_2 = c_2 = 0$, and $c_1(t) = K_s(\hat{x}(t), t)$ in (3). Then $[v, K_s(\hat{x})] \geqq 0$ and (c) follows from (b). Hence the complementary slackness condition in the lemma holds.

(d) $r > 0$. Suppose $r = 0$. Substituting $b_1 = b_2 = c_2 = 0$ and $c_1(t) = K_s(\bar{x}(t), t)$ in (3) yields $[v, K_s(\bar{x})] \geqq 0$. Since $K_s(\bar{x}(t), t)_j < -a < 0$, (b) implies that $v = 0$. Substituting $b_1 = -p$ and $b_2 = 0$ in (3) yields $-\langle p, p \rangle \geqq 0$. Hence, $p = 0$ a.e. This is impossible since r, p, v cannot all vanish so that $r > 0$ and (3) can be normalized with $r = 1$.

(e) $L'(p, v) = \hat{c}$. Substituting $a_1 = c(x, u)$, $b_1 = \dot{x} - Ax - Bu$, $c_1 = K_s(x)$, $a_2 = \hat{c}$, $b_2 = c_2 = 0$ in (3) and recalling that $r = 1$ from (d) yields $L'(p, v) \geqq \hat{c}$. However, by weak duality, $L'(p, v) \leqq \hat{c}$ and hence $L'(p, v) = \hat{c}$. Note that $p \in \mathscr{L}^\infty(R^n)$, but the lemma claims that $L'(p, v) = \hat{c}$ where $p \in \mathscr{BV}$.

(f) $p = \tilde{p}$ a.e. where $\tilde{p} \in \mathscr{BV}$. This proof is more technical than (a) to (e) and appears in Lemma A. 1 of the Appendix, so the proof of Lemma 1 is complete since $L'(p, v) = L'(\tilde{p}, v)$. □

Proof of Theorem 1. In the problem (P) with explicit control constraints, proceed exactly as in the proof of Lemma 1. A fourth component $d \in \mathscr{C}(R^{m_c})$ is added to the sets Y and Z, where $d \leqq 0$ in Y and $d(t) \geqq K_c(u(t), t)$ in Z. (Note that $d \in \mathscr{C}$ and not $d \in \mathscr{L}^\infty$—if d were chosen in $\mathscr{L}^\infty$, then the Hahn–Banach theorem

would produce a multiplier in the dual of $\mathscr{L}^\infty$ which is a miserable space. By choosing $d \in \mathscr{C}$, the dual multiplier lies in $\mathscr{NBV}$— in fact, it is seen below that the multiplier is also absolutely continuous.)

Continuing as in Lemma 1, we find the Hahn–Banach theorem yields

$$c(x, u)+\langle p, \dot{x}-Ax-Bu\rangle+[v, K_s(x)]+[z, K_c(u)] \geqq \hat{c} \tag{4}$$

for all $x \in \mathscr{A}(R^n)$ with $x(0)=x_0$ and $u \in \mathscr{C}(R^m)$ where $v \in \mathscr{NBV}(R^{m_s})$, $z \in \mathscr{NBV}(R^{m_c})$, and both v and z are nondecreasing. Note that to obtain an optimal solution to (D), it must be shown that: (i) z is absolutely continuous so that $[z, K_c(u)]=\langle w, K_c(u)\rangle$ where $w=\dot{z}$ and (ii) expression (4) holds for all $u \in \mathscr{L}^\infty(R^m)$, not just for $u \in \mathscr{C}(R^m)$. Combining these properties with weak duality, Proposition 1, implies that $L(p, w, v)=\hat{c}$.

First it is proved that the infimum of the left side of (4) over $x \in \mathscr{A}(R^n)$ and $u \in \mathscr{C}(R^m)$ actually equals $\hat{c}$. Let $\{u^k\}$ be a minimizing sequence for (P) and let $\{x^k\}$ be the corresponding trajectories. The sequence $\{u^k\}$ lies in $\mathscr{L}^\infty$; however, in Lemma A.2 of the Appendix, it is shown that the convexity of K_c and the existence of an interior point for the constraint $K_c(u(t), t) \leqq 0$ (given in (SL)) imply that for any $\varepsilon>0$, there exists $y_\varepsilon^k \in \mathscr{C}(R^m)$ satisfying $K_c(y_\varepsilon^k(t), t) \leqq 0$, $|y_\varepsilon^k(t)-u^k(t)| \leqq \varepsilon$ except on a set of measure less than ε, and $\|y_\varepsilon^k\| \leqq \|\bar{u}\|+\|u^k\|$, where $\bar{u}$ is the interior control given in (SL). Thus, by the continuity of $h(\cdot, \cdot, \cdot)$, the integrand of the cost functional of (P), it follows that $\lim_{\varepsilon \to 0} c(x^k, y_\varepsilon^k)=c(x^k, u^k)$ and $\lim_{\varepsilon \to 0}\langle p, \dot{x}^k-Ax^k-By_\varepsilon^k\rangle=0$. Now given $\delta>0$, there exist k' such that $|c(x^{k'}, u^{k'})-\hat{c}|<\delta/3$ and ε' such that $|c(x^{k'}, y_{\varepsilon'}^{k'})-c(x^{k'}, u^{k'})|<\delta/3$ and $|\langle p, \dot{x}^{k'}-Ax^{k'}-By_{\varepsilon'}^{k'}\rangle|<\delta/3$. Since $[z, K_c(y_{\varepsilon'}^{k'})] \leqq 0$ and $[v, K_s(x^{k'})] \leqq 0$, then the left side of (4) evaluated at $x=x^{k'}$ and $u=y_{\varepsilon'}^{k'}$ is within δ of $\hat{c}$, and hence the infimum of the left side over (x, u) satisfying $x \in \mathscr{A}(R^n)$, $u \in \mathscr{C}(R^m)$ and $x(0)=x_0$ equals $\hat{c}$ as claimed.

The proof that $z \in \mathscr{A}(R^{m_c})$ is now summarized, and the details can be found in Lemma A.3 of the Appendix.

Define $g(x, u)=c(x, u)+\langle p, \dot{x}-Ax-Bu\rangle+[v, K_s(x)]$. Using the construction of the previous paragraph, there exists a sequence (x^k, y^k) satisfying $g(x^k, y^k) \to \hat{c}$, $[z, K_c(y^k)] \leqq 0$, and $y^k \in \mathscr{C}(R^m)$. It is possible to express $z=r+s$ where $r \in \mathscr{A}(R^{m_c})$, $s \in \mathscr{BV}(R^{m_c})$, $\dot{s}=0$ a.e., $s(0)=0$ and s is nondecreasing (see Rudin [8, p. 166]). In Lemma A.3 of the Appendix, it is shown that a sequence $\{\delta^k\} \subset \mathscr{C}(R^{m_c})$ can be constructed with $\delta^k=0$ except on a set E of small measure on which is concentrated the variation on s, $\delta^k=\bar{u}-y^k$ just inside E, and hence $[s, K_c(y^k+\delta^k)] \leqq as(1)/2$ where $a<0$ was given in (SL). Since s is nondecreasing, then $s(1) \geqq 0$, and unless $s=0$, (4) will be contradicted since $g(x^k, y^k+\delta^k)+as(1)/2$ will be less than $\hat{c}$ for k sufficiently large. Hence $z=r \in \mathscr{A}(R^{m_c})$.

To complete the proof, it must be shown that (4) holds for $u \in \mathscr{L}^\infty(R^m)$, not just $u \in \mathscr{C}(R^m)$. By Lusin's theorem [8, p. 53], any $u \in \mathscr{L}^\infty(R^m)$ can be approximated by $y_\varepsilon \in \mathscr{C}(R^m)$ satisfying $y_\varepsilon=u$ except on a set of measure less than ε and $\|y_\varepsilon\| \leqq \|u\|$. Since (4) holds for y_ε, the continuity condition (C) implies that (4) holds for $u \in \mathscr{L}^\infty(R^m)$. Thus $L(p, w, v)=\hat{c}$ as desired and the complementary slackness conditions follow as in Lemma 1, property (c). □

Notice that the duality results above were derived by separating the sets Y and Z with a hyperplane, and exploiting the separation condition (4) above to

push the dual variables into successively smaller spaces. An immediate question is whether the spaces exhibited above are the smallest possible. A more recent paper [11] will show that for a strictly convex, quadratic cost control problem with linear state and control constraints satisfying an independence condition, there exists an optimal control u^*, a corresponding trajectory x^* and dual multipliers p^*, w^* and v^* such that $(\dot{x}^*, u^*, p^*, w^*, v^*)$ are all Lipschitz continuous when K_s has a Lipschitz continuous partial derivative in t and A, B, K_c, and h are Lipschitz continuous in t. Furthermore, if no state constraints are present, then $\ddot{p}^*$ is Lipschitz continuous when the data defining (P) is sufficiently smooth. Below it is shown that when state constraints are present, a linear combination, q^*, of p^* and v^* has increased smoothness, and in [11] the Lipschitz continuity of $\dot{q}^*$ is proved. Hence $(\dot{x}^*, \dot{q}^*, u^*, w^*, v^*)$ have derivatives in L^∞. Also by an example given in [11], it is seen that $(\dot{x}^*, \dot{q}^*, u^*, w^*, v^*)$ may be discontinuous when K_s does not possess a Lipschitz continuous partial derivative t.

3. Extension of duality theory to controls in $\mathscr{L}^1$. Let $(\bar{\mathrm{P}})$ denote the control problem with constraint $u \in \mathscr{L}^1(R^m)$ instead of $u \in \mathscr{L}^\infty(R^m)$. It is assumed both that the components of $B(\cdot)$ lie in $\mathscr{L}^\infty$ so that the differential equation $\dot{x} = Ax + Bu$ makes sense, and the integral in the cost functional is defined for $x \in \mathscr{A}(R^n)$ and $u \in \mathscr{L}^1(R^m)$ (i.e., the integrand is in $\mathscr{L}^1$).

THEOREM 2. *Suppose* (C) *and* (SL) *hold and the optimal value $\bar{c}$ of* $(\bar{\mathrm{P}})$ *is finite. Then there exist $(\hat{p}, \hat{w}, \hat{v})$ that are optimal in* (D) *with $L(\hat{p}, \hat{w}, \hat{v}) = \bar{c}$. If $(\hat{x}, \hat{u})$ are optimal in* $(\bar{\mathrm{P}})$, *then the complementary slackness condition of Theorem* 1 *holds.*

Note that in defining the dual problem (D), we still restrict $u \in \mathscr{L}^\infty(R^m)$ in the minimization of (1).

Proof. Let $\hat{c}$ denote the optimal value of (P). Since $\hat{c} \geqq \bar{c} > -\infty$, then Theorem 1 implies the existence of (p, w, v) with

$$c(x, u) + \langle p, \dot{x} - Ax - Bu \rangle + \langle w, K_c(u) \rangle + [v, K_s(x)] \geqq \hat{c} \geqq \bar{c} \tag{5}$$

for all (x, u) satisfying $x \in \mathscr{A}(R^n)$, $x(0) = x_0$ and $u \in \mathscr{L}^\infty(R^m)$. It is now shown that $\hat{c} = \bar{c}$. Suppose for the moment that there exists an optimal solution $(\hat{x}, \hat{u})$ to $(\bar{\mathrm{P}})$. Define the following control u_k and set S_k:

$$u_k = \begin{cases} \hat{u}(t) & \text{when } |\hat{u}(t)| \leqq k, \\ \bar{u}(t) & \text{when } |\hat{u}(t)| > k, \end{cases} \qquad S_k = \{t : \hat{u}(t) \neq u_k(t)\},$$

where $\bar{u}$ was given in (SL).

Since $w \geqq 0$, v is nondecreasing, and $K_c(u_k(t), t) \leqq 0$ and $K_s(\hat{x}(t), t) \leqq 0$ for $t \in [0, 1]$, then inserting $(x, u) = (\hat{x}, u_k)$ into (5) yields $c(\hat{x}, u_k) + \langle p, B(\hat{u} - u_k) \rangle \geqq \hat{c} \geqq \bar{c}$. Since the components of $B(\cdot)$ and $p(\cdot)$ lie in $\mathscr{L}^\infty$, $u \in \mathscr{L}^1(R^m)$, $\hat{u} = u_k$ except on S_k, and $\mu(S_k) \to 0$ as $k \to \infty$, where $\mu(\cdot)$ denotes Lebesque measure, then $0 = \lim_{k \to \infty} \langle p, B(\hat{u} - u_k) \rangle$. Similarly $c(\hat{x}, \hat{u}) - c(\hat{x}, u_k) = \int_{S_k} \{h(\hat{x}(t), \hat{u}(t), t) - h(\hat{x}(t), \bar{u}(t), t)\}\, dt$ and both $h(\hat{x}(\cdot), \hat{u}(\cdot), \cdot)$ and $h(\hat{x}(\cdot), \bar{u}(\cdot), \cdot)$ lie in $\mathscr{L}^1$, so $c(\hat{x}, \hat{u}) = \lim_{k \to \infty} c(\hat{x}, u_k)$. Thus $\hat{c} = \bar{c}$ since the left side of (5) evaluated at $x = \hat{x}$ and $u = u_k$ converges to $\bar{c}$. Since $L(p, w, v) = \hat{c}$, $L(p, w, v) = \bar{c}$.

If there does not exist an optimal solution to (P), then by choosing a minimizing sequence and approximating each element of the minimizing sequence as above, it can be proved that $L(p, w, v) = \hat{c} = \bar{c}$.

Now the complementary slackness condition is verified. Again by the inequality (5) above, substituting $u = u_k$, $x = \hat{x}$ yields:

$$(6) \qquad \int_{S_k^c} (w(t), K_c(\hat{u}(t), t))\, dt \geqq \langle w, K_c(u_k)\rangle \geqq \bar{c} - c(\hat{x}, u_k) - \langle p, B(\hat{u} - u_k)\rangle.$$

As shown above, the right side of (6) converges to zero as $k \to \infty$. Since $\lim_{k\to\infty} \mu(S_k^c) = 1$, $E_j = \{t : w(t)_j > 0,\ K_c(\hat{u}(t), t)_j < 0\}$ has no measure, and the complementary slackness condition in the control constraint must hold. A similar proof confirms the complementary slackness condition in the state constraint. □

4. Minimum principles. In order to solve the dual problem numerically, the x and u that achieve the infimum in (1) must be characterized. This leads to a *minimum principle* and an *adjoint condition.* Theorem 3 below proves that the minimization over u in (1) can be taken under the integral sign.

THEOREM 3. *Suppose* (C) *and* (SL) *hold,* (p, w, v) *is feasible in* (D) *with* $L(p, w, v) > -\infty$, *and* $x^* \in \mathcal{A}(R^n)$ *and* $u^* \in \mathcal{L}^\infty(R^m)$ *achieve the minimum in* (1) *corresponding to* (p, w, v). *Then the minimum of* $f(u, t) = h(x^*(t), u, t) - (p(t), B(t)u) + (w(t), K_c(u, t))$ *occurs at* $u = u^*(t)$ *for almost every* $t \in [0, 1]$. *Similarly, if* $L'(p, v) > -\infty$, *the cost functional in* (P′) *is given by* $c(\cdot, \cdot)$, $U(t) = \{b \in R^m : K_c(b, t) \leqq 0\}$, *and* $x^* \in \mathcal{A}(R^n)$ *and* $u^* \in \mathcal{L}^\infty(R^m)$ *achieve the minimum in* (2) *corresponding to* (p, v), *then the minimum of* $\{h(x^*(t), u, t) - (p(t), B(t)u)\}$ *over* $u \in U(t)$ *occurs at* $u = u^*(t)$ *for almost every* $t \in [0, 1]$.

Proof. Only the first minimum principle above will be proved since the second is similar. Let $\bar{c} = L(p, w, v)$ where by definition

$$(7) \qquad \begin{aligned} L(p, w, v) = \inf \Bigg[\int_0^1 & \{h(x(t), u(t), t) + (p(t), \dot{x}(t) - A(t)x(t) - B(t)u(t)) \\ & + (w(t), K_c(u(t), t))\}\, dt + [v, K_s(x)] \Bigg] \\ & \text{subject to } x \in \mathcal{A}(R^n), \quad u \in \mathcal{L}^\infty(R^m), \quad x(0) = x_0. \end{aligned}$$

Let E denote the intersection of the Lebesgue points of each term in the integrand of (7) evaluated at (x^*, u^*) and suppose $f(z, s) < f(u^*(s), s)$ for some $s \in E$ and $z \in R^m$. Let Δ denote a ball of diameter δ centered at s, $I(\Delta, u)$ the integral in (7) evaluated at $x = x^*$ over the ball Δ, and $J(u(\cdot))$ the integrand in (7) evaluated at $x = x^*$. Since s is a Lebesgue point of $J(u^*(\cdot))$, $I(\Delta, u^*) = J(u^*(s))\delta + o(\delta)$. Define v_δ to be a control that agrees with u^* outside Δ and equals z inside Δ. It is easy to see that $I(\Delta, v_\delta) = J(z)\delta + o(\delta)$, and since $f(z, s) < f(u^*(s), s)$, $J(z) < J(u^*(s))$ and $I(\Delta, v_\delta) < I(\Delta, u^*)$ for δ sufficiently small. This violates the optimality of (x^*, u^*) in (7) so that the minimum principle holds on E. Since E has full measure, the proof is complete. □

Note that Theorem 3 holds for all (p, w, v) that are feasible in the dual problem, while the standard necessary conditions only hold for some (p, w, v). Also observe that it is not possible to carry out the minimization over x under the

integral sign in (1) due to the presence of the $\dot{x}$ term. The following lemma will be needed before the adjoint conditions can be derived.

LEMMA 2. *Suppose* (C) *and* (SL) *hold,* (p, w, v) *is feasible in* (D) *with* $L(p, w, v) > -\infty$, (x^*, u^*) *achieves the minimum in* (1) *for* (p, w, v), $K_s(\cdot, \cdot)$ *is twice continuously differentiable, and* $G(t)$ *denotes the gradient of* $K_s(\cdot, t)$ *evaluated at* $x^*(t)$. *Then if* q *is defined by* $q(1) = 0$, $q(t) = G(t)^T v(t) - p(t)$ *for* $t \in (0, 1)$, *and* $q(0) = q(0^+)$, *then* $q \in \mathscr{A}(R^n)$. *If* K_s *is affine, then the existence of* (x^*, u^*) *is not required.*

Proof. By the definition of L,

$$L(p, w, v) \leqq c(x, u) + \langle p, \dot{x} - Ax - Bu \rangle + \langle w, K_c(u) \rangle + [v, K_s(x)] \tag{8}$$

for all $x \in \mathscr{A}(R^n)$ with $x(0) = x_0$ and $u \in \mathscr{L}^\infty(R^m)$. Each term on the right side of (8) is convex and furthermore the $[v, K_s(x)]$-term is differentiable in x. Recall the following standard necessary condition: Suppose v^* solves the problem: minimize $f(v) + g(v)$ subject to $v \in F$ where f, g and F are all convex and f is differentiable. Then v^* satisfies $g(v^*) \leqq g(v) + (d/dv) f'(v^*)\,(v - v^*)$ for all $v \in F$ (see [3]). Applying this result to the right side of (8) we get

$$\begin{aligned} c(x^*, u^*) + \langle p, \dot{x}^* - Ax^* - Bu^* \rangle + \langle w, K_c(u^*) \rangle \\ \leqq c(x, u) + \langle p, \dot{x} - Ax - Bu \rangle + [v, G(x - x^*)] + \langle w, K_c(u) \rangle \end{aligned} \tag{9}$$

for all $x \in \mathscr{A}(R^n)$ with $x(0) = x_0$ and $u \in \mathscr{L}^\infty(R^m)$. Observe that equality holds in (9) for $x = x^*$ and $u = u^*$.

Since p is continuous from the left on $[0, 1)$, the integration by parts formula of Dunford and Schwartz [4, p. 154] gives

$$\oint_0^{1} (p(t), \dot{x}(t) - \dot{x}^*(t))\, dt = (p(1^-), x(1) - x^*(1)) - \int_{0^+}^{1^-} [x(t) - x^*(t)]^T\, dp(t). \tag{10}$$

The boundary term at $t = 0$ vanishes since $x(0) = x^*(0) = x_0$. Since K_s has two continuous derivatives, then the gradient of $K_s(\cdot, t)$ is absolutely continuous, and hence $G(\cdot)$ is absolutely continuous. Thus the following relation holds:

$$\int_{0^+}^{1^-} x(t)^T G(t)^T\, dv = \int_{0^+}^{1^-} x(t)^T\, d(G(t)^T v) - \int_{0^+}^{1^-} v(t)^T \dot{G}(t) x(t)\, dt. \tag{11}$$

Since v is normalized with $v(1) = 0$ and since $x(0) = x^*(0) = x_0$,

$$\begin{aligned} \int_0^{1} (x(t) - x^*(t))^T G(t)^T\, dv = \int_{0^+}^{1^-} (x(t) - x^*(t))^T G(t)^T\, dv \\ - (x(1) - x^*(1))^T G(1)^T v(1^-). \end{aligned} \tag{12}$$

Combining (9), (10), (11) and (12) we find

$$\begin{aligned} c(x, u) - \langle p, Ax + Bu \rangle - \langle v, \dot{G}x \rangle + \int_{0^+}^{1^-} x(t)^T\, dq(t) + \langle w, K_c(u) \rangle \\ - (q(1^-), x(1)) \geqq \bar{c} > -\infty \end{aligned} \tag{13}$$

for all (x, u) satisfying $x \in \mathscr{A}(R^n)$, $u \in \mathscr{L}^\infty(R^m)$ and $x(0) = x_0$, where $\bar{c} > -\infty$ is a constant depending on x^*, u^*, p, w and v. Again equality holds in (13) for $x = x^*$ and $u = u^*$. If K_s is affine, then (13) holds without even assuming the existence of (x^*, u^*), and $\bar{c}$ only depends on $L(p, w, v)$.

Now it is shown that $q(1^-)=0$. Define the continuous function $g(\delta, \varepsilon, t)$ as follows: $g(\delta, \varepsilon, \cdot)$ is linear on $[1-\varepsilon, 1]$ and satisfies $g(\delta, \varepsilon, t)=0$ for $t\in[0, 1-\varepsilon]$ and $g(\delta, \varepsilon, 1)=\delta q(1^-)$. Inserting $x(t)=x_0+g(\delta, \varepsilon, t)$ into (13) and letting $\varepsilon\to 0$ and $\delta\to+\infty$, we get a contradiction since the left side of (13) deverges to $-\infty$ due to the presence of the boundary term in (13).

Now consider the absolute continuity of q. It is possible to express $q=r+s$, where $r\in\mathscr{A}(R^n)$, $s\in\mathscr{BV}(R^n)$, $s(0)=0$ and $\dot{s}=0$ a.e. (see Rudin [8, p. 166]). If $E=\{t : \dot{s}(t)\neq 0\}$, then Lemma A.4 in the Appendix proves that unless $s=0$, a sequence $\{x_k\}\subset\mathscr{A}(R^n)$ can be chosen such that x_k agrees with $\bar{x}$ just outside of E and $[s, x_k]\to-\infty$. This will violate (13), and hence $s=0$ and $q=r\in\mathscr{A}(R^n)$. □

THEOREM 4. *Suppose* (C) *and* (SL) *hold,* (p, w, v) *is feasible in* (D) *with* $L(p, w, v)>-\infty$, $x^*\in\mathscr{A}(R^n)$ *and* $u^*\in\mathscr{L}^\infty(R^m)$ *achieve the minimum in* (1) *corresponding to* (p, w, v) *and* $K_s(\cdot, \cdot)$ *is twice continuously differentiable. Then the minimum of*

$$f(x, t)=h(x, u^*(t), t)+(\dot{q}(t)+A^T(t)\,q(t)-(\dot{G}(t)^T+A(t)^TG(t)^T)v(t), x)$$

occurs at $x=x^*(t)$ *for almost every* $t\in[0, 1]$, *where* G *and* q *were defined in Lemma 2. If* $h(\cdot, u, t)$ *is differentiable, then the adjoint equation holds:* $q(1)=0$ *and*

$$\dot{q}(t)=-A(t)^Tq(t)-h(x^*(t), u^*(t), t)_x+(\dot{G}(t)^T+A(t)^TG(t)^T)v(t) \quad a.e.\ t \tag{14}$$

Proof. In Lemma 2 it was observed that $q\in\mathscr{A}(R^n)$ so that $[q, x]=\langle\dot{q}, x\rangle$. From (13),

$$\int_0^1 \{h(x(t), u^*(t), t)-(p(t), A(t)x(t))-(v(t), \dot{G}(t)x(t))+(\dot{q}(t), x(t))\}\,dt\geqq\tilde{c} \tag{15}$$

for all $x\in\mathscr{A}(R^n)$ with $x(0)=x_0$, where $\tilde{c}>-\infty$ is a constant depending only on x^*, u^*, p, w and v. As noted after (13), equality holds in (15) for $x=x^*$. As in Theorem 3, we wish to say that $x^*(t)$ yields the pointwise minimum for the integrand. There is one technical point, though, since in Theorem 3, u was contained in $\mathscr{L}^\infty$, while in (15), x lies in $\mathscr{A}$. However, if $z\in R^n$ yields a better minimum for the integrand of (15) at the Lebesgue point $t=s$, then by [10, p. 9] there exists an infinitely differentiable function ϕ_δ^ε that equals 1 on $[s-\delta, s+\delta]$ and equals 0 on $[s+\delta+\varepsilon, 1]$ and $[0, s-\delta-\varepsilon]$. Thus the function $x_\delta^\varepsilon=z\phi_\delta^\varepsilon+(1-\phi_\delta^\varepsilon)x^*$ is absolutely continuous and equals z near s and x^* away from s. Letting first $\varepsilon\to 0$ and then $\delta\to 0$ again violates the optimality of x^*. The adjoint equation is obtained simply by setting the derivative of $f(\cdot, t)$ to zero at $x=x^*(t)$. □

The condition (14) above is the familiar adjoint equation for state constrained problems given in [5] and [2]. These standard necessary conditions only assert that (14) holds for some (p, w, v) where (x^*, u^*) is optimal in (P), while Theorem 4 holds for all (p, w, v) feasible in (D). Using the minimum principles, Theorems 3 and 4 above, the evaluation of $L(p, w, v)$ is reduced to the solution of a sequence of math programming problems for each $t\in[0, 1]$. In certain cases, such as problems with quadratic cost and linear constraints, the minimum principles permit the explicit determination of the (x, u) achieving the minimum in (1) in terms of (p, w, v). The numerical solution of the dual problem using the Ritz method is analyzed in [1].

A combined state and control minimum principle can be proved, and the proof is similar to Theorems 3 and 4 above.

THEOREM 5. *Suppose* (C) *and* (SL) *hold,* (p, w, v) *is feasible in* (D) *with* $L(p, w, v) > -\infty$, $K_s(\cdot, \cdot)$ *is twice continuously differentiable and* $x^* \in \mathcal{A}(R^n)$ *and* $u^* \in \mathcal{L}^\infty(R^m)$ *achieve the minimum in* (1) *corresponding to* (p, w, v). *Then the minimum of* $f(x, u, t)$ *defined below occurs at* $x = x^*(t)$ *and* $u = u^*(t)$ *for a. e.* t:

$$
\begin{aligned}
f(x, u, t) = {} & h(x, u, t) + (q(t) - G(t)^T v(t), B(t)u) + (w(t), K_c(u, t)) \\
& + (\dot{q}(t) + A(t)^T q(t) - (\dot{G}(t)^T + A(t)^T G(t)^T) v(t), x)
\end{aligned} \tag{16}
$$

Appendix. Regularity of the dual variables.

LEMMA A.1. *Suppose* (C′) *and* (SL′) *are satisfied, the optimal value* $\hat{c}$ *of* (P′) *is finite, and* $L'(p, v) = \hat{c}$ *where* $p \in \mathcal{L}^\infty(R^n)$ *and* $v \in \mathcal{BV}(R^{m_s})$. *Then* $p = \tilde{p}$ *a.e. where* $\tilde{p} \in \mathcal{BV}(R^n)$.

Proof. For notational convenience, p is assumed scalar-valued (the proof below could be applied to each component of p separately to demonstrate the result for vector-valued functions). Let R denote the set of Lebesgue points and suppose that p has infinite variation on this set. It is now shown that this to a contradiction.

Given a constant b, there exists $0 = t_0 < t_1 \cdots < t_N$ such that

$$
\sum_{1 \leqq j \leqq N,\, j \text{ odd}} |p(t_{j-1}) - p(t_j)| > b \tag{A.1}
$$

and either $p(t_{j+1}) < p(t_j) > p(t_{j-1})$ for j even or the reverse inequalities hold. For the construction given below, it is assumed that the former holds. Let $\alpha, \rho, M > 0$ be as given in (SL′), and define the function $x_\varepsilon(t)$ as follows: $x_\varepsilon(\cdot)$ is the continuous, piecewise linear function that is zero for j odd and $-\rho$ for j even on the interval $[t_j + \varepsilon, t_{j+1} - \varepsilon]$, linear on the interval $[t_j - \varepsilon, t_j + \varepsilon]$ for all j, and zero at $t = 0$. Notice that as $\varepsilon \to 0$, $\dot{x}_\varepsilon \to -\rho \sum_{j=0}^{N} (-1)^j \delta(t - t_j)$, where $\delta(\cdot)$ is the delta function, and since $\{t_j\}$ are Lebesgue points of p and $p(t_{j+1}) < p(t_j) > p(t_{j-1})$ for j even, then $\lim_{\varepsilon \to 0} \langle p, \dot{x}_\varepsilon \rangle = \rho \sum_{1 \leqq j \leqq N,\, j \text{ odd}} p(t_j) - p(t_{j-1}) < -\rho b$.

From the definition of L',

$$
f(\bar{x} + x_\varepsilon, \bar{u}) + \langle p, \dot{x}_\varepsilon + \bar{x} - A(\dot{x}_\varepsilon + \bar{x}) - B\bar{u} \rangle + [v, K_s(\bar{x} + x_\varepsilon)] \geqq \hat{c}, \tag{A.2}
$$

where $(\bar{x}, \bar{u})$ was given in (SL′). Also by (SL′), $f(\bar{x} + x_\varepsilon, \bar{u}) < M$, and hence all the terms in (A.2) are bound uniformly in b and ε except for the $\langle p, \dot{x}_\varepsilon \rangle$-term which becomes less than $-\rho b$ for ε sufficiently small. Thus if b were chosen sufficiently large, this would lead to a contradiction in (A.2), and hence the total variation of p on R is finite.

Since R has full measure (see [8, p. 158]), for all $t \in R^c$, there exists a sequence $\{t_j\} \subset R$ such that $t_j \to t^-$. Because p has finite variation on R, $\lim_{j \to \infty} p(t_j)$ exists, and it is possible to define a function $\tilde{p}(t)$ that equals $[p(t)]$ for $t \in R$ and equals $[\lim_{j \to \infty} p(t_j)]$ if $t \notin R$ where $\{t_j\} \subset R$ and $t_j \to t^-$. Thus $\tilde{p}(t) = p(t)$ a.e., and $\tilde{p}$ has the same variation on $[0, 1]$ as p has on R. □

The following theorem essentially proves that if the set $U = \{u(\cdot) \in \mathcal{L}^\infty(R^m) : K(u(t), t) \leq 0\}$ has an interior, then any $u(\cdot) \in U$ can be approximated arbitrarily closely in the $\mathcal{L}^p$-norm by a continuous function in U.

LEMMA A.2. *Suppose* $K: R^m \times [0,1] \to R^n$ *is continuous,* $K(\cdot, t)$ *is convex for* $t \in [0,1]$, *and there exist* $\bar{u} \in \mathscr{C}(R^m)$ *and* $a < 0$ *such that* $K(\bar{u}(t), t)_j < a$ *for all* $t \in [0,1]$ *and* $j = 1, 2, \cdots, n$. *Then given* $u(\cdot) \in U$ *and* $\varepsilon > 0$, *there exists* $v \in U \cap \mathscr{C}(R^m)$ *such that* $|u(t) - v(t)| < \varepsilon$ *except on a set of measure less than* ε *and* $\|v\| \leqq \|\bar{u}\| + \|u\|$.

Proof. Let $w = b\bar{u} + (1-b)u$, where $1 > b > 0$ is small enough that $\|u - w\| \leqq \varepsilon$. By the convexity of $K(\cdot, t)$, $K(w(t), t)_i \leqq ba < 0$ for $j = 1, \cdots, n$, and by Lusin's theorem [8, p. 53], there exists $y \in \mathscr{C}(R^m)$ with $y = w$ on a closed set E satisfying $\mu(E^c) \leqq \varepsilon$, where $\mu(\cdot)$ denotes Lebesgue measure and furthermore, $\|y\| \leqq \|w\|$. Since $K(y(\cdot), \cdot)$ is uniformly continuous on $[0,1]$, there exists a constant $\delta > 0$ such that if $|t - s| < \delta$, then $|K(y(t), t) - K(y(s), s)| < b|a|$. Outer regularity of the Lebesgue measure implies the existence of an open set D containing E with $\mu(D - E) < \delta$. Also D can be chosen so that no point of D is more than δ away from a point of E (for example, construct open balls of diameter δ about each point of E, choose a finite subcover $\{B_j\}$ of the balls, construct an open set $B \supset E$ with $\mu(B - E) < \delta$, and define $D = (\cup B_j) \cap B$).

Since $K(y(t), t)_j = K(w(t), t)_j \leqq ba < 0$ on E and any point of D is at most δ away from a point of E, $K(y(t), t) \leqq 0$ for $t \in D$. From Urysohn's lemma, there exists $g \in \mathscr{C}(R^1)$ with the support of g contained in D, $g(t) = 1$ for $t \in E$, and $\|g\| = 1$ (we use the notation of Rudin [8] to denote a function satisfying these conditions: $E < g < D$). Define $v = gy + (1-g)\bar{u}$. For $t \in D$, v is on the line segment between two functions that satisfy the constraint $K(\cdot, t) \leqq 0$, and since $K(\cdot, t)$ is convex, $K(v(t), t) \leqq 0$. On the other hand, $v = \bar{u}$ on D^c so $v \in U$. By construction, $v(t) = y(t) = w(t)$ for $t \in E$ so $|u(t) - v(t)| \leqq \varepsilon$ except on a set E^c of measure less than ε. Also $|v(t)| \leqq g(t)|y(t)| + (1-g(t))|\bar{u}(t)| \leqq g(t)|w(t)| + (1-g(t))|\bar{u}(t)| \leqq [g(t)(b-1)+1]|\bar{u}(t)| + (1-b)g(t)|u(t)|$, and since $0 < b < 1$ and $\|g\| = 1$, the bound $\|v\| \leqq \|\bar{u}\| + \|u\|$ is immediate. □

LEMMA A.3. *Suppose* (C) *and* (SL) *hold; then the function* $z \in \mathscr{BV}$ *in* (4) *is absolutely continuous.*

Proof. To keep notation simple, K_c is assumed to have range in R^1—the proof for vector-valued functions is identical, but it is necessary to introduce extra subscripts. Let $g(x, u)$ denote the first three terms in (4) and let $F = \{(x, u) : x \in \mathscr{A}(R^n), \quad x(0) = x_0, \quad u \in \mathscr{C}(R^m)\}$. As shown after (4), $\hat{c} = \inf\{g(x, u) + [z, K_c(u)] : (x, u) \in F\}$, and there exists a sequence $(x^k, u^k) \in F$ such that $g(x^k, u^k) \to \hat{c}$ and $K_c(u^k(t), t) \leqq 0$ for $t \in [0,1]$. Also recall that z was nondecreasing.

Rudin [8, p. 166] proves that $z = r + s$, where $r \in \mathscr{A}$, $s \in \mathscr{BV}$, $\dot{s} = 0$ a.e., s is nondecreasing and $s(0) = 0$. We now suppose that $s \neq 0$ or equivalently $s(1) > 0$, and show that this leads to a contradiction. As noted above, it is possible to find $(x, u) \in F$ such that $K_c(u(t), t) \leqq 0$ for $t \in [0,1]$, and $g(x, u) < \hat{c} + |a|s(1)/8$ where "a" was given in (SL). Since s is nondecreasing and $s(0) = 0$, the total variation of s is $s(1)$. First a summary of the proof is given.

Since $\dot{s} = 0$ a.e., it will be shown that a closed set E can be constructed that is a union of a finite number of intervals with the variation of s concentrated on E, and $\mu(E)$, the measure of E, is very small. Then a function $v \in \mathscr{C}(R^m)$ is constructed that satisfies $\|v\| \leqq \|\bar{u}\| + \|u\|$, $K_c(v(t), t) \leqq 0$ for $t \in [0,1]$, and v agrees with $\bar{u}$, the interior control given in (SL), on E and agrees with u just outside of E. Thus

$\int_E K_c(v(t), t)\, ds(t) = \int_E K_c(\bar{u}(t), t)\, ds(t) < a \int_E ds(t) < (a/2)s(1)$ where the last inequality follows since E captures almost all the variation of s. Also $|\int_{\bar{E}} K_c(v(t), t)\, dr(t)| < |a|s(1)/8$ since $\mu(E)$ is chosen small, and $\int_{E^c} K_c(v(t), t)\, dz(t) \leqq 0$ since z is nondecreasing and $K_c(v(t), t) \leqq 0$. Combining these inequalities, we see that $[z, K_c(v)] \leqq s(1)(a/2 + |a|/8) = 3as(1)/8$. If $\mu(E)$ is chosen small enough so that $g(x, v) < \hat{c} + |a|s(1)/4$, then $g(x, v) + [z, K_c(v)] \leqq \hat{c} + |a|s(1)/4 + as(1)3/8 = \hat{c} + as(1)/8$. Since $a, -s(1) < 0$, this contradicts the optimality of $\hat{c}$; hence $s = 0$ and $z = r \in \mathscr{A}$.

Now E and v will be constructed. Begin by choosing a closed set $H \subset [0, 1]$ with $\dot{s} = 0$ on H and $\mu(H^c) < \varepsilon$. For each $h \in H$, construct an open ball D^h of radius $2r^h$ where r^h is choosen sufficiently small so that $|s(t) - s(T)| \leqq \varepsilon |t - T|$ whenever $t, T \in \overline{D^h}$. Since $\dot{s}(h) = 0$, this construction is possible. Let B^h be the open ball centered at h of radius r^h. Since H is compact, a finite subcover of these balls $\{B_j\}$ can be chosen of radii $\{r_j\}$. Define $B = \cup B_j$ and $D = \cup D_j$; since s is monotone and $|s(t) - s(T)| \leqq \varepsilon |t - T|$ whenever $t, T \in \bar{D}_j$ for some j, then the total variation of s on $\bar{D}$ is at most ε, and hence the variation of s on D^c is at least $s(1) - \varepsilon$. Also since $H \subset B \subset D$, $\mu(D^c) \leqq \mu(B^c) \leqq \mu(H^c) < \varepsilon$. By Urysohn's lemma, there exists $\xi \in \mathscr{C}(R^1)$ satisfying $D^c < \xi < \bar{B}^c$. Defining $v = (1 - \xi)u + \xi\bar{u}$, we see that $v = u$ on B, $v = \bar{u}$ on D^c, $\|v\| \leqq \|u\| + \|\bar{u}\|$; also since $K_c(\cdot, t)$ is convex, $K_c(u(t), t) \leqq 0$ and $K_c(\bar{u}(t), t) \leqq 0$, then $K_c(v(t), t) \leqq 0$. Choosing $E = D^c$ and returning to the summary above, we notice that for ε sufficiently small, all the statements in the summary hold. □

LEMMA A.4. Suppose (C), (SL) and inequality (13) hold. Then $q \in \mathscr{A}(R^n)$.

Proof. Again to keep notation simple, assume q is scalar-valued—the argument below can be applied to each component of q separately to treat vector-valued functions. Since $q = G^T v - p$ on $(0, 1)$ and G is absolutely continuous while v and p lie in $\mathscr{BV}$, then q is continuous from the left on $[0, 1)$ (see the definition of $\mathscr{BV}$) and by Lemma 2, q is continuous from the left on $[0, 1]$ since $q(1^-) = 0 = q(1)$. As in Lemma A.3, we can express $q = r + s$, where $r \in \mathscr{A}$, $s \in \mathscr{BV}$, $s(0) = 0$, and $s = 0$ almost everywhere. Let us suppose that $s(t) \neq 0$ for some $t \in [0, 1)$—it is shown that (13) is violated and hence $s = 0$.

Since s is continuous from the left, then the total variation of s on $[0, t]$ is a continuous function of t from the left, and it is possible to choose $t' < t$ such that the variation of s on $[t', t]$ is less than δ.

Using the construction of Lemma A.3 on $[0, t']$, one generates sets $B \subset D \subset [0, t']$ such that the variation of s on $\bar{D}$ is at most ε and $\mu([0, t'] - B) < \varepsilon$. (Since the construction of Lemma A.3 was only valid for a monotone function, this last step actually requires that we first express $s = s_1 + s_2$, where s_1 and $-s_2$ are both nondecreasing (see Natanson [6]) and $\dot{s}_1 = \dot{s}_2 = 0$ a.e. (see Rudin [8, p. 166]). Then using Lemma A.3, sets D_1 and D_2 are constructed that capture only $\varepsilon/4$ of the variation of s_1 and s_2, respectively, and that satisfy $\mu([0, t'] - D_1), \mu([0, t'] - D_2) < \varepsilon/4$. Then define $D = D_1 \cap D_2$.

Now choose $\varepsilon < |t - t'|$ and define J_ρ to be an open ball centered at t' of diameter ρ. Again construct $\xi \in \mathscr{C}(R^1)$ satisfying $([0, t'] - D) < \xi < ([0, t'] - B) \cup J_\varepsilon$ and define $x_N = (1 - \xi)\bar{x} + \xi N$, where $\bar{x}$ was given in (SL) and $N \in R^1$ with

$\operatorname{sgn}(N) = -\operatorname{sgn}(s(t))$. (We had to introduce the set J_ε since $[0, t'] - \bar{B}$ is not an open set on $[0, 1]$ as required by Urysohn's lemma.)

Since $x_N = \bar{x}$ on $[t, 1]$,

$$\int_0^1 x_N(\sigma)\, dq(\sigma) = \int_0^t x_N(\sigma)\, ds(\sigma) + \int_t^1 \bar{x}(\sigma)\, ds(\sigma) + \int_0^1 x_N(\sigma)\, dr(\sigma)$$

$$\leqq Ns(t) + 1 + \|\bar{x}\|(TV(s) + TV(r)), \tag{A.3}$$

where $TV(s)$ is the total variation of s on $[0, 1)$ and the last inequality follows from the following relations whenever ε and δ are chosen sufficiently small:

$$\int_0^{t'} x_N(\sigma)\, ds(\sigma) = \int_{\bar{D}} x_N(\sigma)\, ds(\sigma) + \int_{[0, t'] - D} x_N(\sigma)\, ds(\sigma)$$

$$= N \int_0^{t'} ds(\sigma) + \int_{\bar{D}} (x_N(\sigma) - N)\, ds(\sigma)$$

$$\leqq Ns(t') + \varepsilon(\|\bar{x}\| + |N|)$$

$$\leqq -|N|(|s(t)| - \delta) + \varepsilon(\|\bar{x}\| + |N|),$$

$$\int_{t'}^{t} x_N(\sigma)\, ds(\sigma) \leqq \delta(|N| + \|\bar{x}\|),$$

If $g(x)$ denotes the first three terms of (13) evaluated at $u = u^*$, then for ε and δ sufficiently small, $g(x_N)$ is close to $g(\bar{x})$. However, this combined with the inequality (A.3) and the fact that $Ns(t) < 0$ violates (13) for N large; i.e., $\inf\{g(x) + \int_0^1 x(t)^T\, dq(t) : x \in \mathcal{A}(R^n),\ x(0) = x_0\}$ is no longer finite. Hence $s = 0$ and q is absolutely continuous. □

REFERENCES

[1] W. W. HAGER, *The Ritz–Trefftz method for optimal control problems with state and control constraints*, SIAM J. Numer. Anal., 12 (1975), pp. 854–867.

[2] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[3] J. L. LIONS, *Contrôle Optimal des Systèmes Gouvernés par des Équations aux Derivées Partielles*, Dunod, Paris, 1968.

[4] N. DUNFORD AND T. SCHWARTZ, *Linear Operators*, Interscience, New York, 1963.

[5] L. W. NEUSTADT, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–59.

[6] I. P. NATANSON, *The Theory of Functions of a Real Variable*, Frederick Ungar, New York, 1964.

[7] R. T. ROCKAFELLAR, *State constraints in convex control problems of Bolza*, this Journal, 10 (1972), pp. 691–715.

[8] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

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

[10] J. BARROS-NETO, *An Introduction to the Theory of Distributions*, Marcel Dekker, New York, 1973.

[11] W. W. HAGER, *Quadratic program stability and optimal control regularity*, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

A PERTURBATION PROBLEM FROM CONTROL EXHIBITING ON AND OFF THE BOUNDARY BEHAVIOR*

DENNIS D. BERKEY AND MARVIN I. FREEDMAN†

Abstract. We consider a regular perturbation problem for a system arising from a control problem in which the optimal control moves on and off the boundary of the control region in a continuous way. By using a nonlinear change of time scales, we are able to give conditions under which our technique yields a solution of the perturbed system.

1. Introduction. Recent attempts to study perturbed systems of differential equations of the sort that arise in control have led to a number of interesting papers among which are those of O'Malley [6] and [7], Sannuti [8], Haddad and Kokotovic [4] and Kokotovic and Yackel [5]. These papers are mainly concerned with singularly perturbed problems where a small parameter ε enters in such a way that the order of the system drops on setting $\varepsilon = 0$. In the regularly perturbed case, some recent work by Freedman and Kaplan ([1] and [2]) has treated problems where the optimal control is of the bang-bang type.

In this paper, again in a regularly perturbed context, the authors carry out a perturbation analysis for a system of perturbed differential equations which is characteristic of the situation that arises when the optimal control moves on and off the boundary of the control region in a continuous way. For simplicity, we present only the case where the control makes one passage from the interior of the control region to its boundary. It should be clear though how our method extends to the situation in which a finite number of such passages occur.

In particular, we deal with the system described on the interval $[0, T]$ by

$$\dot{x} = f(x, u, \varepsilon), \tag{1a}$$

$$\dot{\lambda} = g(x, \lambda, u, \varepsilon), \tag{1b}$$

together with the boundary conditions

$$x(0, \varepsilon) = a(\varepsilon), \tag{1c}$$

$$\lambda(T, \varepsilon) = b(\varepsilon). \tag{1d}$$

In the above, ε is a small nonnegative real parameter and x, λ, f and g take values in R^n. The function $u(t, \varepsilon)$ is taken to be scalar-valued. Explicitly we assume that there exists a scalar-valued "Hamiltonian" function $H(x, \lambda, u, \varepsilon)$ such that $u(t, \varepsilon)$ satisfies

$$H_u(x, \lambda, u, \varepsilon) = 0 \tag{1e}$$

for all $t \in [0, T]$ for which $|u(t, \varepsilon)| < 1$ while

$$|u(t, \varepsilon)| = 1 \tag{1f}$$

otherwise.

* Received by the editors April 10, 1975, and in revised form November 24, 1975.
† Department of Mathematics, Boston University, Boston, Massachusetts 02215.

If we view (1a)–(1d) as the Euler–Lagrange equations corresponding to some free endpoint trajectory optimization problem with $u(t, \varepsilon)$ as the optimal control taking values in $[-1, 1]$, then (1e) arises as a consequence of the well-known fact that

$$u(t, \varepsilon) = \min_{v \in [-1,1]} H(x, \lambda, v, \varepsilon).$$

Thus the $H_u = 0$ condition holds while $u(t, \varepsilon)$ is in the interior of the control region. Let us call a solution of (1a)–(1f) the solution of the *full system.*

After motivating our considerations with the example of § 2, we will phrase an appropriate perturbation problem for the system (1) and develop formal procedures for computing the asymptotic series expansions of the variables involved in § 3. In § 4 we pursue the conditions for solvability given in § 3 and develop some general conditions under which they will be satisfied. In § 5 we establish the uniform validity of the asymptotic expansions developed in § 3. In § 6 we apply our techniques to a legitimate optimal control problem in which the controller is not a priori known.

2. The heuristics of the method. To illustrate the problem to be treated here, let us suppose that somehow or other one knows that the optimal control for a certain control problem takes the form

$$u(t, \varepsilon) = \min(t + \varepsilon, 1), \qquad t \in [0, 2].$$

Then one cannot write a uniformly valid expansion

$$u(t, \varepsilon) = u_0(t) + u_1(t)\varepsilon + O(\varepsilon^2),$$

as $u(t, \varepsilon)$ is not even once differentiable with respect to ε. Also if the state variable x (or costate variable λ) satisfies a differential equation involving $u(t, \varepsilon)$, then a uniformly valid expansion

$$x(t, \varepsilon) = x_0(t) + x_1(t)\varepsilon + O(\varepsilon^2)$$

is also impossible. For example, let us assume that

$$\dot{x}(t, \varepsilon) = u(t, \varepsilon), \qquad x(0, \varepsilon) = 0,$$

for $t \in [0, 2]$. Then requiring that x be continuous on $[0, 2]$ gives that

$$x(t, \varepsilon) = \begin{cases} \frac{1}{2}t^2 + t\varepsilon & \text{for } 0 \leqq t \leqq 1 - \varepsilon, \\ t - \frac{1}{2}(\varepsilon^2 - 2\varepsilon + 1) & \text{for } 1 - \varepsilon < t \leqq 2, \end{cases}$$

which is certainly not smooth in ε.

However, analytic expressions for x and u may be achieved by introducing a "new clock function" τ as follows. We make the nonlinear change of variables

$$t = h(\tau, \varepsilon) = \tau(2 - \tau)(1 - \varepsilon) + \tau(\tau - 1).$$

Then writing $X(\tau, \varepsilon) = x(h(\tau, \varepsilon), \varepsilon)$ we have

$$X(\tau, \varepsilon) = \begin{cases} \frac{1}{2}\tau^4\varepsilon^2 - \tau^3\varepsilon(2\varepsilon - 1) + \tau^2(\frac{1}{2} - 2\varepsilon + 3\varepsilon^2) + \tau(\varepsilon - 2\varepsilon^2) & \text{for } 0 \leqq \tau \leqq 1, \\ \tau^2\varepsilon + \tau(1 - 2\varepsilon) - \frac{1}{2}\varepsilon^2 + \varepsilon - \frac{1}{2} & \text{for } 1 < \tau \leqq 2, \end{cases}$$

and $X(\tau, \varepsilon)$ is now analytic in ε. In fact, letting

$$(X_0(\tau), X_1(\tau), X_2(\tau)) = \begin{cases} (\frac{1}{2}\tau^2, \tau^3 - 2\tau^2 + \tau, \frac{1}{2}\tau^4 - 2\tau^3 + 3\tau^2 - 2\tau) & \text{if } \tau \leqq 1, \\ (\tau - \frac{1}{2}, \tau^2 - 2\tau + 1, -\frac{1}{2}) & \text{if } \tau > 1, \end{cases}$$

we get that

$$X(\tau, \varepsilon) = X_0(\tau) + X_1(\tau)\varepsilon + X_2(\tau)\varepsilon^2.$$

Furthermore, writing $U(\tau, \varepsilon) = u(h(\tau, \varepsilon), \varepsilon)$ gives that

$$U(\tau, \varepsilon) = \min((\tau - 1)^2\varepsilon + \tau, 1),$$

which is analytic in ε for ε sufficiently small. Indeed, letting ε be so small that $(\tau - 1)^2\varepsilon + \tau < 1$ if $\tau < 1$ ($\varepsilon = 1$ will suffice) from the above, we have for

$$(U_0(\tau), U_1(\tau)) = \begin{cases} (0, 1) & \text{if } \tau \leqq 1, \\ ((\tau - 1)^2, \tau) & \text{if } \tau > 1, \end{cases}$$

that

$$U(\tau, \varepsilon) = U_0(\tau) + U_1(\tau)\varepsilon.$$

Notice that while producing the desired analytic nature of X and U, the transformation $t = h(\tau, \varepsilon)$ has had the effect of "freezing" that smallest value of the variable τ for which $U(\tau, \varepsilon) = 1$, denoted by $\tau(\varepsilon)$, at $\tau(\varepsilon) = 1$. Thus our idea is to view the control problem in terms of a "new clock function" τ chosen so that a perturbation analysis is possible. We remark that the clock function idea has previously been utilized in [1] and [2] where the authors were studying bang-bang behavior.

3. Our system in the new clock variable. In this section, we show how one may construct a formal asymptotic expansion of the solution to the system (1). The expansion will be valid in some suitably small neighborhood of the solution of the following system.

$$\dot{x}_0 = (x_0, u_0, 0), \tag{2a}$$

$$\dot{\lambda}_0 = g(x_0, \lambda_0, u_0, 0), \tag{2b}$$

$$x_0(0) = a_0, \tag{2c}$$

$$\lambda_0(T) = b_0, \tag{2d}$$

$$H(x_0, \lambda_0, u_0) = \min_{|v| \leqq 1} H(x_0, \lambda_0, v). \tag{2e}$$

We will refer to (2) as the reduced system associated with the system (1). Throughout what follows we shall hypothesize:

H1. that the reduced system (2) has a solution x_0, λ_0, u_0, all continuous on $[0, T]$;

H2. that $|u_0(t)| < 1$ for $0 \leqq t < t_0$ and $u_0(t) = 1$ for $t_0 \leqq t \leqq T$;

H3. that $u_0'(t_0^-)$, the left-hand derivative of u_0 at t_0, is positive;

H4. that $H_{uu}(x_0(t), \lambda_0(t), u_0(t), 0) > 0$ for $0 \leqq t \leqq t_0$;

H5. that $H_u(x_0(t), \lambda_0(t), u_0(t), 0) < 0$ for $t_0 < t \leqq T$;

H6. that there exists an $\varepsilon_0 > 0$ and an integer $K \geqq 1$ such that f and g are $(K+1)$ times continuously differentiable and that $H(x, \lambda, u, \varepsilon)$ is $(K+2)$ times continuously differentiable with respect to x, λ, u, and ε for all $0 \leqq \varepsilon \leqq \varepsilon_0$ and (x, λ, u) in a neighborhood of the reduced solution;

H7. that $a(\varepsilon)$ and $b(\varepsilon)$ are $(K+1)$ times continuously differentiable with respect to ε for $0 \leqq \varepsilon \leqq \varepsilon_0$, and that $a(0) = a_0$ and $b(0) = b_0$.

An observation concerning these hypotheses is in order. Besides hypothesizing enough differentiability to enable us to carry out our perturbation analysis, our assumptions focus attention on the situation in which the optimal controller for the reduced system takes its initial value in the interior of the control region and then moves continuously to the boundary where it remains. This is the statement of H2 and may be viewed as the result of the control variable constraint $|u| \leqq 1$. Moreover for small ε, we want the optimal controller for the perturbed system (1) to exhibit the same type of behavior, i.e., that $u(t, \varepsilon)$ will remain on the boundary $u(t, \varepsilon) = 1$ once it has arrived. This is quaranteed by H3 and H5.

For the moment, let us assume that the full system (1) possesses a solution $x(t, \varepsilon)$, $\lambda(t, \varepsilon)$ and $u(t, \varepsilon)$. For ε sufficiently small, let $t(\varepsilon)$ denote the smallest value of $t \in [0, T]$ for which $u(t, \varepsilon) = 1$. Then for $\varepsilon = 0$, $t(\varepsilon) = t_0$. Also we temporarily assume that $t(\varepsilon)$ is a $(K+1)$ times continuously differentiable function of ε. We proceed as in § 2 to introduce a nonlinear change of variables $\tau \to t$ which will freeze $t(\varepsilon)$ at t_0 while taking $0 \to 0$ and $T \to T$. Specifically, we define

$$t = h(\tau, \varepsilon) = \frac{\tau(\tau - T)}{t_0(t_0 - T)} t(\varepsilon) + \frac{\tau(\tau - t_0)}{T - t_0}. \tag{3}$$

Then

$$h(0, \varepsilon) = 0, \quad h(T, \varepsilon) = T, \quad h(t_0, \varepsilon) = t(\varepsilon),$$

and for ε sufficiently small, $h(\tau, \varepsilon)$ is monotonically increasing for $\tau \in [0, T]$.

We next define the following new variables:

$$X(\tau, \varepsilon) = x(h(\tau, \varepsilon), \varepsilon), \quad \Lambda(\tau, \varepsilon) = \lambda(h(\tau, \varepsilon), \varepsilon), \quad U(\tau, \varepsilon) = u(h(\tau, \varepsilon), \varepsilon).$$

Then

$$X(0, \varepsilon) = x(0, \varepsilon) = a(\varepsilon), \qquad \Lambda(T, \varepsilon) = \lambda(T, \varepsilon) = b(\varepsilon),$$

and for ε sufficiently small, we have that

$$|U(\tau, \varepsilon)| < 1 \quad \text{if } 0 \leqq \tau < t_0,$$

and

$$U(t_0, \varepsilon) = 1.$$

Also we write

$$X_0(\tau) = X(\tau, 0) = x_0(\tau), \quad \Lambda_0(\tau) = \Lambda(\tau, 0) = \lambda_0(\tau), \quad U_0(\tau) = U(\tau, 0) = u_0(\tau).$$

Written in terms of the transformed variable, (1) becomes

$$\frac{d}{d\tau}X(\tau,\varepsilon)=f(X,U,\varepsilon)\left[\frac{2\tau-T}{t_0(t_0-T)}t(\varepsilon)+\frac{2\tau-t_0}{T-t_0}\right], \tag{4a}$$

$$\frac{d}{d\tau}\Lambda(\tau,\varepsilon)=g(X,\Lambda,U,\varepsilon)\left[\frac{2\tau-T}{t_0(t_0-T)}t(\varepsilon)+\frac{2\tau-t_0}{T-t_0}\right] \tag{4b}$$

with boundary conditions

$$X(0,\varepsilon)=a(\varepsilon), \tag{4c}$$

$$\Lambda(T,\varepsilon)=b(\varepsilon) \tag{4d}$$

and the additional conditions that

$$H_u(X,\Lambda,U,\varepsilon)=0 \quad \text{if } \tau\leqq t_0 \tag{4e}$$

and

$$U(\tau,\varepsilon)=1 \quad \text{if } \tau\geqq t_0. \tag{4f}$$

Before proceeding further, we note that by our hypothesis H5, condition (4f) will hold for ε sufficiently small provided $U(t_0,\varepsilon)=1$. The remainder of our paper is concerning with the execution of a perturbation analysis for the system (4) above.

In § 4, we shall rigorously establish the existence of a solution $X(\tau,\varepsilon)$, $\Lambda(\tau,\varepsilon)$, $U(\tau,\varepsilon)$ and a $(K+1)$ times continuously differentiable "switch" time $t(\varepsilon)$ satisfying (4) above. The solution will converge uniformly to $x_0(\tau)$, $\lambda_0(\tau)$, $u_0(\tau)$ as $\varepsilon\to 0^{\pm}$ while $t(\varepsilon)\to t_0$.

4. The formal expansions. In this section we proceed formally, assuming that a solution to system (4) exists and that X, Λ, U and $t(\varepsilon)$ are each $(K+1)$ times continuously differentiable with respect to ε. We therefore write

$$X(\tau,\varepsilon)=x_0(\tau)+\sum_{j=1}^{K}X_j(\tau)\varepsilon^j+O(\varepsilon^{K+1}), \tag{5a}$$

$$\Lambda(\tau,\varepsilon)=\lambda_0(\tau)+\sum_{j=1}^{K}\Lambda_j(\tau)\varepsilon^j+O(\varepsilon^{K+1}), \tag{5b}$$

$$U(\tau,\varepsilon)=u_0(\tau)+\sum_{j=1}^{K}U_j(\tau)\varepsilon^j+O(\varepsilon^{K+1}), \tag{5c}$$

$$t(\varepsilon)=t_0+\sum_{j=1}^{K}t_j\varepsilon^j+O(\varepsilon^{K+1}), \tag{5d}$$

where $O(\varepsilon^{K+1})$ holds uniformly in τ in (5a)–(5c). Next we have for ε sufficiently small, that $H_u(X,\Lambda,U,\varepsilon)=0$ on $[0,t_0]$. Differentiating j times with respect to ε, setting $\varepsilon=0$ and using the above expansions gives for each $j=1,\cdots,K$ that

$$H_{ux}(x_0,\lambda_0,u_0,0)X_j(\tau)+H_{u\lambda}(x_0,\lambda_0,u_0,0)\Lambda_j(\tau)$$
$$+H_{uu}(x_0,\lambda_0,u_0,0)U_j(\tau)+h_j(\tau)=0$$

for all $\tau \in [0, t_0]$, where we have used the notation

$$h_j(\tau) = \frac{\partial^j}{\partial \varepsilon^j} H_u(X, \Lambda, U, \varepsilon)\Big|_{\varepsilon=0}.$$

We now substitute series (5) into equations (4). By our smoothness hypotheses, the resulting equations may be differentiated j times with respect to ε, $1 \leqq j \leqq K$. Upon setting $\varepsilon = 0$, we obtain

$$\begin{aligned}\frac{d}{d\tau}X_j(\tau) &= \frac{\partial}{\partial x}f(x_0, u_0, 0)X_j(\tau) + \frac{\partial}{\partial u}f(x_0, u_0, 0)U_j(\tau) \\ &\quad + \frac{2\tau - T}{t_0(t_0 - T)}t_j f(x_0, u_0, 0) + p_j(\tau),\end{aligned} \tag{6a}$$

$$\begin{aligned}\frac{d}{d\tau}\Lambda_j(\tau) &= \frac{\partial}{\partial x}g(x_0, \lambda_0, u_0, 0)X_j(\tau) + \frac{\partial}{\partial u}g(x_0, \lambda_0, u_0, 0)U_j(\tau) \\ &\quad + \frac{\partial}{\partial \lambda}g(x_0, \lambda_0, u_0, 0)\Lambda_j(\tau) + \frac{2\tau - T}{t_o(t_0 - T)}t_j g(x_0, \lambda_0, u_0, 0) + g_j(\tau),\end{aligned} \tag{6b}$$

$$\begin{aligned}&H_{ux}(x_0, \lambda_0, u_0, 0)X_j(\tau) + H_{u\lambda}(x_0, \lambda_0, u_0, 0)\Lambda_j(\tau) \\ &\quad + H_{uu}(x_0, \lambda_0, u_0, 0)U_j(\tau) + h_j(\tau) = 0, \qquad \tau \in [0, t_0],\end{aligned} \tag{6c}$$

$$U_j(\tau) = 0, \qquad \tau \in [t_0, T], \tag{6d}$$

$$X_j(0) = a_j \tag{6e}$$

and

$$\Lambda_j(T) = b_j, \tag{6f}$$

where $p_j(\tau)$ and $g_j(\tau)$ are polynomials in $X_1(\tau), \cdots, X_{j-1}(\tau), \Lambda_1(\tau), \cdots, \Lambda_{j-1}(\tau)$, $U_1(\tau), \cdots, U_{j-1}(\tau)$, and $t_1, \cdots, t_{j-1}$ with coefficients determined by $x_0(\tau)$, $u_0(\tau)$, $\lambda_0(\tau)$ and t_0. Since u is a scalar function, we note that f_u and g_u are column vectors. The symbols a_j and b_j denote the coefficients of ε^j in the finite Taylor series expansion of $a(\varepsilon)$ and $b(\varepsilon)$, respectively.

We would like to know when the system (6) can be solved recursively for $X_j(\tau)$, $\Lambda_j(\tau)$, $U_j(\tau)$ and t_j. We therefore assume these to have been determined for all positive integers less than j. Thus $p_j(\tau)$ and $g_j(\tau)$ are known functions.

We now make the following substitutions:

$$Z(\tau) = X_j(\tau) - \frac{\tau(\tau - T)}{t_0(t_0 - T)}t_j f(x_0, u_0, 0), \tag{7a}$$

$$V(\tau) = U_j(\tau) - \frac{\tau(\tau - T)}{t_0(t_0 - T)}t_j U_0'(\tau), \tag{7b}$$

$$W(\tau) = \Lambda_j(\tau) - \frac{\tau(\tau - T)}{t_0(t_0 - T)}t_j g(x_0, \lambda_0, u_0, 0). \tag{7c}$$

After we solve (6c) for U_j by hypothesis H4 and use substitutions (7), our system (6) becomes

$$\frac{d}{d\tau}Z(\tau)=f_x(\tau)Z(\tau)+f_u(\tau)V(\tau)+p_j(\tau), \tag{8a}$$

$$\frac{d}{d\tau}W(\tau)=g_\lambda(\tau)W(\tau)+g_x(\tau)Z(\tau)+g_u(\tau)V(\tau)+g_j(\tau), \tag{8b}$$

with the additional conditions

$$V(\tau)=-H_{uu}^{-1}(\tau)[H_{ux}(\tau)Z(\tau)+H_{u\lambda}(\tau)W(\tau)+h_j(\tau)], \qquad \tau\in[0,\ t_0], \tag{8c}$$

$$V(\tau)=0 \quad \text{for } \tau\in[t_0,\ T], \tag{8d}$$

and boundary conditions

$$Z(0)=X_j(0)=a_j, \tag{8e}$$

$$W(T)=\Lambda_j(T)=b_j. \tag{8f}$$

Taking the limit as τ approaches t_0 from the left in (7b), we obtain the additional internal boundary condition that

$$V(t_0^-)=-t_j u_0'(t_0^-). \tag{9}$$

In writing system (8), we have used the notation

$$f_x(\tau)=\frac{\partial f}{\partial x}(x_0(\tau),\ u_0(\tau),\ 0)$$

with the other derivatives being similarly abbreviated.

Our system on $[0,\ t_0]$ can now be written as

$$\frac{dZ}{d\tau}=[f_x(\tau)-H_{uu}^{-1}(\tau)f_u(\tau)H_{ux}(\tau)]Z(\tau) - H_{uu}^{-1}(\tau)f_u(\tau)H_{u\lambda}(\tau)W(\tau)+\bar{p}_j(\tau)$$

and

$$\frac{dW}{d\tau}=[g_x(\tau)-H_{uu}^{-1}(\tau)g_u(\tau)H_{ux}(\tau)]Z(\tau) +[g_\lambda(\tau)-g_u(\tau)H_{uu}^{-1}(\tau)H_{u\lambda}(\tau)]W(\tau)+\bar{q}_j(\tau),$$

where $\bar{p}_j(\tau)=p_j(\tau)-H_{uu}^{-1}(\tau)h_j(\tau)$ and $\bar{q}_j(\tau)=q_j(\tau)-H_{uu}^{-1}(\tau)h_j(\tau)$, while on $(t_0,\ T]$ our system is

$$\frac{dZ}{d\tau}=f_x(\tau)Z(\tau)+p_j(\tau),$$

$$\frac{dW}{d\tau}=g_x(\tau)Z(\tau)+g_\lambda(\tau)W(\tau)+q_j(\tau).$$

We now define the matrix functions F_j and G_j, $j = 1, 2$, by

$$F_1(\tau) = \begin{cases} f_x(\tau) - H_{uu}^{-1} f_u(\tau) H_{ux}(\tau) & \text{if } 0 \leqq \tau \leqq t_0, \\ f_x(\tau) & \text{if } t_0 < \tau \leqq T, \end{cases}$$

$$F_2(\tau) = \begin{cases} -H_{uu}^{-1}(\tau) f_u(\tau) H_{u\lambda}(\tau) & \text{if } 0 \leqq \tau \leqq t_0, \\ 0 & \text{if } t_0 < \tau \leqq T, \end{cases}$$

$$G_1(\tau) = \begin{cases} g_x(\tau) - H_{uu}^{-1}(\tau) g_u(\tau) H_{ux}(\tau) & \text{if } 0 \leqq \tau \leqq t_0, \\ g_x(\tau) & \text{if } t_0 < \tau \leqq T, \end{cases}$$

and

$$G_2(\tau) = \begin{cases} g_\lambda(\tau) - g_u(\tau) H_{uu}^{-1}(\tau) H_{u\lambda}(\tau) & \text{if } 0 \leqq \tau \leqq t_0, \\ g_\lambda(\tau) & \text{if } t_0 < \tau \leqq T. \end{cases}$$

Finally, let $P(\tau)$ and $Q(\tau)$ be defined by

$$P(\tau) = \begin{cases} \bar{p}_j(\tau) & \text{if } 0 \leqq \tau \leqq t_0, \\ p_j(\tau) & \text{if } t_0 < \tau \leqq T, \end{cases} \qquad Q(\tau) = \begin{cases} \bar{q}_j(\tau) & \text{if } 0 \leq \tau \leqq t_0, \\ q_j(\tau) & \text{if } t_0 < \tau \leqq T. \end{cases}$$

Then we may write system (8) on all of $[0, T]$ as

$$\frac{dZ}{d\tau} = F_1(\tau) Z(\tau) + F_2(\tau) W(\tau) + P(\tau), \tag{10a}$$

$$\frac{dW}{d\tau} = G_1(\tau) Z(\tau) + G_2(\tau) W(\tau) + Q(\tau) \tag{10b}$$

with boundary conditions

$$Z(0) = a_j, \tag{10c}$$

$$W(T) = b_j. \tag{10d}$$

Now system (10) is a linear inhomogeneous system of two vector differential equations with mixed boundary conditions. From the basic theory of systems of linear differential equations it follows that system (10) will have a unique solution if and only if the corresponding homogeneous system with boundary conditions $Z(0) = W(T) = 0$ admits only the trivial solution. If a solution to (10) does exist, then $V(t)$ may be obtained for $\tau \in [0, t_0)$ from (8c), and then t_j is determined by (9) by virtue of hypothesis H3. Finally, X_j, U_j and Λ_j can then be obtained from substitutions (7). We summarize our discussion on solvability with the following theorem.

THEOREM 1. *Let hypotheses* H1–H7 *hold. Then the system* (6) *can be solved recursively for* X_j, Λ_j, U_j *and* t_j *if the homogeneous linear system*

$$\frac{dZ}{d\tau} = F_1(\tau) Z(\tau) + F_2(\tau) W(\tau), \tag{11a}$$

$$\frac{dW}{d\tau} = G_1(\tau) Z(\tau) + G_2(\tau) W(\tau) \tag{11b}$$

with boundary conditions

$$Z(0) = 0 = W(T)$$

has only the trivial solution.

If the hypotheses of Theorem 1 hold, we shall say that the system (4) is *formally solvable.* While Theorem 1 is stated in the general setting of a two-point boundary value problem, we indicate in the following section conditions under which it is applicable to more specific setting of our control problem.

5. Conditions sufficient for solvability. For the purposes of this section only, we shall make the additional assumption that the functions $f(x, u, \varepsilon)$ and $g(x, \lambda, u, \varepsilon)$ of system (1) satisfy

$$f(x, u, \varepsilon) = H_\lambda(x, \lambda, u, \varepsilon),$$
$$g(x, \lambda, u, \varepsilon) = -H_x^T(x, \lambda, u, \varepsilon).$$

These additional assumptions amount to assuring that the function $H(x, \lambda, u, \varepsilon)$ which plays the part of a scalar-valued Hamiltonian is related to f and g via the Euler–Lagrange equations. In every application of our perturbation analysis to problems arising from optimal control, this will, of course, be the case.

In this section we develop conditions under which our system (4) will be formally solvable. We begin with an observation concerning system (11) on $[t_0, T]$.

Let $\Psi(\tau)$ denote the fundamental matrix solution for the equation

$$\frac{dZ}{d\tau} = f_x(\tau)Z(\tau)$$

for which $\Psi(t_0) = I$. Then on $[t_0, T]$ we may write $Z(\tau) = \Psi(\tau)Z(t_0)$. Since $f_x = H_{\lambda x} = -g_\lambda^T$, the fundamental matrix solution taking the value I at t_0 for

$$\frac{dW}{d\tau} = g_\lambda(\tau)W(\tau)$$

is $[\Psi^T]^{-1}$. From (11b) we have that

$$W(\tau) = [\Psi(\tau)^T]^{-1}\left\{W(t_0) + \int_{t_0}^{\tau} \Psi(s)^T g_x(s)\Psi(s)Z(t_0)\, ds\right\}.$$

Applying boundary condition (11c) we get that

$$0 = W(T) = [\Psi(T)^T]^{-1}\left\{W(t_0) + \int_{t_0}^{T} \Psi(s)^T g_x(s)\Psi(s)\, ds\, Z(t_0)\right\}.$$

Since $\Psi(T)$ is nonsingular, we have established the following.

LEMMA 1. *Any solution* (Z, W) *of system* (11a)–(11c) *must satisfy* $W(t_0) = SZ(t_0)$, *where* S *is the matrix defined by*

$$S = -\int_{t_0}^{T} \Psi(s)^T g_x(s)\Psi(s)\, ds.$$

Using Lemma 1 we will establish our main result on solvability which is the following.

THEOREM 2. *If the matrix* $\begin{bmatrix} H_{xx}(t) & H_{xu}(t) \\ H_{ux}(t) & H_{uu}(t) \end{bmatrix}$ *is positive definite for all* $t \in [0, t_0]$ *and* $H_{xx}(t)$ *is positive semidefinite for all* $t \in (t_0, T]$, *then system* (4) *is formally solvable.* (Here $H_{xx}(t)$ abbreviates $H_{xx}(x_0(t), \lambda_0(t), u_0(t), 0)$, etc.)

Proof. Our analysis here will center around the linear quadratic regulator problem given by the state equation

$$\frac{dz}{dt} = f_x(t)z(t) + v(t)f_u(t), \tag{12a}$$

the initial condition

$$z(0) = 0, \tag{12b}$$

and the performance index

$$J = \frac{1}{2}z(t_0)^T S z(t_0) + \frac{1}{2}\int_0^{t_0} [z^T, v] \begin{bmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{bmatrix} \begin{bmatrix} z \\ v \end{bmatrix} dt. \tag{12c}$$

S is as defined in Lemma 1, and the scalar v is to be chosen so as to minimize J. It is well known that a unique minimizing solution of the above problem exists in the case where S is symmetric and positive definite. We begin by verifying that S is indeed positive definite. Recall that $S = -\int_{t_0}^{T} \Psi(s)^T g_x(s)\Psi(s)\, ds$, where Ψ is the fundamental matrix solution of

$$\dot{z} = f_x z,$$

for which $\Psi(t_0) = I$. Now since $g_x = -H_{xx}$, we can write

$$S = \int_{t_0}^{T} \Psi(s)^T H_{xx} \Psi(s)\, ds,$$

from which the symmetry of S is immediate. Since from our hypotheses $H_{xx}(t_0)$ must be positive definite, we have for any nonzero vector $x \in R^n$ that

$$\langle \Psi(t_0)^T H_{xx}(t_0)\Psi(t_0)x, x\rangle = \langle H_{xx}(t_0)\Psi(t_0)x, \Psi(t_0)x\rangle > 0,$$

and since H_{xx} is positive semidefinite on $(t_0, T]$, we have

$$\langle \Psi(t)^T H_{xx}(t)\Psi(t)x, x\rangle = \langle H_{xx}(t)\Psi(t)x, \Psi(t)x\rangle \geqq 0$$

for each $t \in (t_0, T]$. Hence *as the definite integral of a positive semidefinite matrix function which is positive definite for at least one* $t \in [t_0, T]$, S *is positive definite.*

Thus we are assured that there is a unique $v(t)$ and corresponding $z(t)$ satisfying (12a)–(12c). However, direct inspection of (12a)–(12c) shows that the choice $v(t) = 0$, $z(t) = 0$ for $t \in [0, t_0]$ will certainly minimize J. On the other hand, we may define a Hamiltonian and write down the Euler–Lagrange equations for system (12) in the usual way. Let us do so.

The Hamiltonian for problem (12) is given by

$$\hat{H}(z, w, v) = \tfrac{1}{2}[z^T, w] \begin{bmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{bmatrix} \begin{bmatrix} z \\ w \end{bmatrix} + w^T(f_x z + f_u v) \tag{13}$$

where $w(t)$ is a vector in R^n for each $t \in [0, t_0]$. By necessary conditions for the existence of an optimal solution, we have that

$$\dot{w}^T = -\frac{\partial H}{\partial z} \tag{14}$$

and

$$\frac{\partial \hat{H}}{\partial v} = 0 \quad \text{on } [0, t_0]. \tag{15}$$

From (13) and (15) we obtain

$$v = -H_{uu}^{-1}(w^T f_u + H_{ux} z). \tag{16}$$

From (12a) and (16), we have that

$$\begin{aligned} \dot{z} &= f_x z - H_{uu}^{-1}(w^T f_u + H_{ux} z) f_u \\ &= [f_x - H_{uu}^{-1} f_u H_{ux}] z - H_{uu}^{-1} f_u w^T f_u. \end{aligned}$$

Now since in this section we assume $f = H_\lambda$, we have $f_u = H_{\lambda u}$. Thus $f_u w^T f_u = f_u w^T H_{\lambda u} = f_u H_{\lambda u}^T w = f_u H_{u\lambda} w$, so the above equation is

$$\dot{z} = [f_x - H_{uu}^{-1} f_u H_{ux}] z - H_{uu}^{-1} f_u H_{u\lambda} w. \tag{17}$$

Similarly from (13) and (14), the costate variable w satisfies

$$\dot{w}^T = -\hat{H}_z = -z^T H_{xx} - v H_{ux} - w^T f_x.$$

Using (16), the above equation becomes

$$\dot{w}^T = -z^T H_{xx} + H_{uu}^{-1}(w^T f_u + H_{ux} z) H_{ux} - w^T f_x.$$

Now in this section we assume $g = -H_x$, so $g_x = -H_{xx}$ (symmetric), $g_u = -H_{xu}$ and $g_\lambda = -H_{x\lambda}$. Using these equalities, we have

$$\dot{w}^T = z^T g_x - H_{uu}^{-1} w^T H_{u\lambda}^T g_u^T - H_{uu}^{-1} z^T H_{ux}^T g_u^T + w^T g_\lambda^T$$

so

$$\dot{w} = [g_x - H_{uu}^{-1} g_u H_{ux}] z + [g_\lambda - H_{uu}^{-1} g_u H_{\lambda u}] w. \tag{18}$$

Also for this quadratic regulator problem, the costate variable w and the state variable z must be related at t_0 by the equation

$$w(t_0) = S(z(t_0)). \tag{19}$$

Since we already know that the unique minimizing solution of (12a)–(12c) is $v \equiv 0$, $z \equiv 0$, we are assured by (19) and the linearity of (18) that $w \equiv 0$ on $[0, t_0)$. We have therefore shown that the unique solution of the differential system given by (17) and (18) with boundary conditions (12b) and (19) is the trivial solution $z \equiv w \equiv 0$ on $[0, t_0)$. But equations (17) and (18) are exactly those of system (11) of Theorem 1 on $[0, t_0)$, boundary condition (12b) corresponds to boundary condition (11c), and by Lemma 1 system (11) must satisfy (19). Hence any solution (z, w) of system (11) must satisfy $z(t) = w(t) = 0$, $t \in [0, t_0)$. Since this gives $z(t_0) = w(t_0) = 0$, by the uniqueness of solutions to the initial value problem for linear systems it now

follows that $z \equiv w \equiv 0$ on $[t_0, T]$ for any solution of system (11). Thus under the hypotheses of Theorem 2, system (11) has only the trivial solution and system (4) is therefore formally solvable. (The reader may recognize our technique here as being similar to those employed in the "inverse problem" in control. See [10] or [11].)

6. Justification of the asymptotic expansions. In this section we demonstrate rigorously that the formal procedure developed in § 3 yields a uniformly valid asymptotic expansion for the variables $X(\tau)$, $\Lambda(\tau)$, $U(\tau)$ and $t(\varepsilon)$ on $[0, T]$. Recall that the system (4) is said to be *formally solvable* if the hypotheses of Theorem 1 hold. Our result here is the following.

THEOREM 3. *Suppose that system* (4) *is formally solvable. Then there exists a small* $\varepsilon_0 > 0$ *such that for* $0 \leqq \varepsilon < \varepsilon_0$, *there is a unique* $(K+1)$ *times continuously differentiable solution* $X(\tau, \varepsilon)$, $\Lambda(\tau, \varepsilon)$, $U(\tau, \varepsilon)$ *and* $t(\varepsilon)$ *of system* (4). *Moreover, if* $X_j(\tau)$, $\Lambda_j(\tau)$, $U_j(\tau)$, t_j, $1 \leqq j \leqq K$, *denote the solution of* (6), *then the expansion* (5) *are uniformly valid on* $[0, T]$ *in* τ.

Proof. We consider first the case $K = 0$. Our method will be to find, for ε sufficiently small, continuous functions $\alpha(\tau, \varepsilon)$, $\beta(\tau, \varepsilon)$, $\gamma(\tau, \varepsilon)$ and $w(\varepsilon)$, $\tau \in [0, T]$ which satisfy

$$\begin{aligned} X(\tau, \varepsilon) &= x_0(\tau) + \varepsilon\alpha(\tau, \varepsilon), \\ \Lambda(\tau, \varepsilon) &= \lambda_0(\tau) + \varepsilon\beta(\tau, \varepsilon), \\ U(\tau, \varepsilon) &= u_0(\tau) + \varepsilon\gamma(\tau, \varepsilon), \\ t(\varepsilon) &= t_0 + \varepsilon w(\varepsilon). \end{aligned}$$

We begin by defining, for real functions r, v and s of τ,

$$\mathscr{F}(r, s, \varepsilon) = \begin{cases} \dfrac{f(x_0 + \varepsilon r, u_0 + \varepsilon s, \varepsilon) - f(x_0, u_0, 0)}{\varepsilon} & \text{if } \varepsilon > 0, \\ f_x r + f_u s + f_\varepsilon & \text{if } \varepsilon = 0, \end{cases}$$

$$\mathscr{G}(r, v, s, \varepsilon) = \begin{cases} \dfrac{g(x_0 + \varepsilon r, \lambda_0 + \varepsilon v, u_0 + \varepsilon s, \varepsilon) - g(x_0, \lambda_0, u_0, 0)}{\varepsilon} & \text{if } \varepsilon > 0, \\ g_x r + g_\lambda v + g_u s + g_\varepsilon & \text{if } \varepsilon = 0 \end{cases}$$

and

$$\mathscr{H}(r, v, s, \varepsilon) = \begin{cases} \dfrac{H_u(x_0 + \varepsilon r, \lambda_0 + \varepsilon v, u_0 + \varepsilon s, \varepsilon) - H_u(x_0, \lambda_0, u_0, 0)}{\varepsilon} & \text{if } \varepsilon > 0, \\ H_{ux} r + H_{u\lambda} v + H_{uu} s + H_{u\varepsilon} & \text{if } \varepsilon = 0, \end{cases}$$

and note that $\mathscr{F}$, $\mathscr{G}$ and $\mathscr{H}$ are continuous functions of ε.

Substituting the expressions for X, Λ, U and $t(\varepsilon)$ into equations (4) and using the above definitions we find that α, β and γ satisfy the system

$$\text{(20a)} \qquad \frac{d}{d\tau}\alpha(\tau, \varepsilon) = \mathscr{F}(\alpha, \gamma, \varepsilon) + \frac{2\tau - T}{t_0(t_0 - T)} w(\varepsilon) f(x_0 + \varepsilon\alpha, u_0 + \varepsilon\gamma, \varepsilon),$$

$$\frac{d}{d\tau}\beta(\tau,\varepsilon)=\mathcal{G}(\alpha,\beta,\gamma,\varepsilon)+\frac{2\tau-T}{t_0(t_0-T)}w(\varepsilon)g(x_0+\varepsilon\alpha,\lambda_0+\varepsilon\beta,u_0+\varepsilon\gamma,\varepsilon) \tag{20b}$$

with boundary conditions

$$\alpha(0,\varepsilon)=a^*(\varepsilon), \tag{20c}$$

$$\beta(T,\varepsilon)=b^*(\varepsilon), \tag{20d}$$

and the additional conditions that

$$\mathcal{H}(\alpha,\beta,\gamma,\varepsilon)=0 \quad \text{for } \tau\in[0,t_0], \tag{20e}$$

$$\gamma(\tau,\varepsilon)=0 \quad \text{for } \tau\in(t_0,T]. \tag{20f}$$

It will suffice now for the case $K=0$ to show that the above system possesses a continuous bounded solution $\alpha(\tau,\varepsilon)$, $\beta(\tau,\varepsilon)$, $\gamma(\tau,\varepsilon)$ and $w(\varepsilon)$ for ε sufficiently small. For the case $\varepsilon=0$, it is a simple matter to verify that the system (20) can be written as a system of the form (10). Hence under our hypothesis of solvability, our system (20) has the solution given by

$$\alpha(\tau)=X_1(\tau),$$

$$\beta(\tau)=\Lambda_1(\tau),$$

$$\gamma(\tau)=U_1(\tau)=\begin{cases}-H_{uu}^{-1}[H_{ux}X_1+H_{u\lambda}\Lambda_1+H_{u\varepsilon}] & \text{if } \tau\leqq t_0,\\ 0 & \text{if } \tau>t_0,\end{cases}$$

and $w(0)=t_1$.

We therefore consider the case $\varepsilon\neq 0$ and view our differential equations as defining a mapping between appropriate Banach spaces so as to be able to invoke the implicit function theorem (see, as a reference, [3]). Toward this objective, we let θ_1 denote some bounded open neighborhood of the origin in $C[0,T]^n\times C[0,T]^n\times C[0,t_0]$ and let θ_2 denote a bounded open neighborhood of 0 in R. Let $\varepsilon\geqq 0$, $(r(\sigma),v(\sigma),s(\sigma))\in\theta_1$ and $w(\varepsilon)\in\theta_2$. Extend each $s\in C[0,t_0]$ to $[0,T]$ by defining

$$s^*(\tau)=\begin{cases}s(\tau) & \text{if } 0\leqq\tau\leqq t_0,\\ s(t_0) & \text{if } \tau>t_0,\end{cases}$$

and consider the integral equations determined by our differential system. They are

$$\alpha(\tau,\varepsilon)=a^*(\varepsilon)+\int_0^\tau \mathcal{F}(r(\sigma),s^*(\sigma))\,d\sigma$$
$$+w(\varepsilon)\int_0^\tau \frac{2\tau-T}{t_0(t_0-T)}f(x_0(\sigma)+\varepsilon r(\sigma),u_0(\sigma)+\varepsilon s^*(\sigma),\varepsilon)\,d\sigma$$

and

$$\beta(\tau,\varepsilon)=b^*(\varepsilon)-\int_0^T \mathcal{G}(r(\sigma),v(\sigma),s^*(\sigma)\,d\sigma$$
$$-w(\varepsilon)\int_0^T \frac{2\tau-T}{t_0(t_0-T)}g(x_0(\sigma)$$
$$+\varepsilon r(\sigma),\lambda_0(\sigma)+\varepsilon v(\sigma),u_0(\sigma)+\varepsilon s^*(\sigma),\varepsilon)\,d\sigma.$$

We also let

$$\zeta(\tau, \varepsilon) = \mathcal{H}(r(\tau), v(\tau), s(\tau), \varepsilon), \qquad \tau \leqq t_0,$$

and

$$\nu = \gamma(t_0, \varepsilon).$$

Now let $\mathcal{B}$ denote the Banach space $C[0, T]^n \times C[0, T]^n \times C[0, t_0] \times R$. It is not difficult to verify that the preceding equations define a continuous Fréchet differentiable mapping from an open neighborhood θ of the origin in $\mathcal{B} \times [0, \varepsilon_0]$ into $\mathcal{B}$ given by

$$\Phi(r, v, s, w, \varepsilon) = (\alpha, \beta, \zeta, \nu).$$

Note that for ε_0 sufficiently small, θ may be selected to include $(X_1, \Lambda_1, U_1, t_1, 0)$. We have already established that

$$\Phi(X_1, \Lambda_1, U_1, t_1, 0) = (X_1, \Lambda_1, 0, 0).$$

Let I denote the mapping of $\mathcal{B} \times [0, \varepsilon_0] \to \mathcal{B}$ defined by

$$I(r, v, s, w, \varepsilon) = (r, v, 0, 0)$$

and consider $\Psi : \theta \to \mathcal{B}$ defined by $\Psi = \Phi - I$. It follows that Ψ is continuous and Fréchet differentiable and that $\Psi(X_1, \Lambda_1, U_1, t_1, 0) = 0$.

We now wish to obtain a continuous bounded solution of the equation

$$\Psi(\alpha(\tau, \varepsilon), \beta(\tau, \varepsilon), \gamma(\tau, \varepsilon), w(\varepsilon), \varepsilon) = 0.$$

By the implicit function theorem, it suffices to show that the Fréchet derivative of Ψ at $(X_1, \Lambda_1, U_1, t_1, 0)$ is a topological linear isomorphism. We denote this derivative by

$$D_{(r,v,s,w)}\Psi|_{(X_1,\Lambda_1,U_1,t_1,0)} : \mathcal{B} \to \mathcal{B}.$$

We consider first the linear map

$$D_{(r,v,s,w)}\Phi : (\eta_1, \eta_2, \eta_3, \eta_4) \to (\mu_1, \mu_2, \mu_3, \mu_4).$$

One can check from our integral equations that for $\varepsilon = 0$, $r = X_1$, $v = \Lambda_1$, $s = U_1$ and $w = t_1$, this mapping is given by

$$\mu_1(\tau) = \int_0^\tau \{f_x(\sigma)\eta_1(\sigma) + f_u(\sigma)\eta_3^{**}(\sigma)\}\, d\sigma + \eta_4 \int_0^\tau \frac{2\sigma - T}{t_0(t_0 - T)} f(x_0, u_0, 0)\, d\sigma,$$

$$\mu_2(\tau) = -\int_\tau^T \{g_x(\sigma)\eta_1(\sigma) + g_\lambda(\sigma)\eta_2(\sigma) + g_u(\sigma)\eta_3^{**}(\sigma)\}\, d\sigma - \eta_4 \int_\tau^T \frac{2\sigma - T}{t_0(t_0 - T)} g(x_0, \lambda_0, u_0, 0)\, d\sigma,$$

$$\mu_3(\tau) = H_{ux}(\tau)\eta_1(\tau) + H_{u\lambda}(\tau)\eta_2(\tau) + H_{uu}(\tau)\eta_3(\tau), \qquad \tau \leqq t_0,$$

$$\mu_4(\tau) = \eta_3(t_0).$$

In the above we have used the notation

$$f^{**}(\tau)=\begin{cases} f(\tau), & \tau \leqq t_0, \\ 0, & \tau > t_0. \end{cases}$$

Thus at $(X_1, \Lambda_1, U_1, t_1, 0)$, the Fréchet derivative

$$D_{(r,v,s,w)}\Psi : (\eta_1, \eta_2, \eta_3, \eta_4) \to (\mu_1, \mu_2, \mu_3, \mu_4)$$

must be given by the equations

$$(\eta_1+\mu_1)(\tau)=\int_0^\tau \{f_x(\sigma)\eta_1(\sigma)+f_u(\sigma)\eta_3^{**}(\sigma)\}\,d\sigma + \eta_4 \int_0^\tau \frac{2\sigma-T}{t_0(t_0-T)} f(x_0, u_0, 0)\,d\sigma,$$

$$(\eta_2+\mu_2)(\tau)=-\int_\tau^T \{g_x(\sigma)\eta_1(\sigma)+g_\lambda(\sigma)\eta_2(\sigma)+g_u(\sigma)\eta_3^{**}(\sigma)\}\,d\sigma - \eta_4 \int_\tau^T \frac{2\sigma-T}{t_0(t_0-T)} g(x_0, \lambda_0, u_0, 0)\,d\sigma,$$

$$\mu_3(\tau)=H_{ux}(\tau)\eta_1(\tau)+H_{u\lambda}(\tau)\eta_2(\tau)+H_{uu}(\tau)\eta_3(\tau), \qquad \tau \leqq t_0,$$

$$\mu_4=\eta_3(t_0).$$

It is easy to see that the above system defines a continuous bounded map $(\eta_1, \eta_2, \eta_3, \eta_4) \to (\mu_1, \mu_2, \mu_3, \mu_4)$. To see that this map is invertible, we assume μ_1, μ_2, μ_3 and μ_4 given and we define

$$\gamma_j = \eta_j + \mu_j \quad \text{for } j = 1, 2 \text{ and } 4,$$

$$\gamma_3 = \eta_3 + \mu_3 - (\mu_4 + \mu_3(t_0)).$$

The reader can now verify that the functions γ_1, γ_2 and γ_3 and the constant γ_4 satisfy a differential system of the form of system (6) in § 4 with $X_j = \gamma_1$, $\Lambda_j = \gamma_2$, $U_j = \gamma_3^{**}$ and $t_j = \gamma_4$ (since $\mu_1, \cdots, \mu_4$ are known). Under our present hypothesis of formal solvability the above system has a unique solution γ_1, γ_2, γ_3 and γ_4 by virtue of Theorem 1. Since μ_j and γ_j are now known, $j = 1, \cdots; 4$, we may solve for $\eta_1, \cdots, \eta_4$. Thus the Fréchet derivative of Ψ at $(X_1, \Lambda_1, U_1, t_1, 0)$ is a topological linear isomorphism. Hence by the Banach space implicit function theorem, there exists an $\varepsilon_1 > 0$ so that for $0 \leqq \varepsilon < \varepsilon_1$, the equation $\Psi(\alpha, \beta, \gamma, w, \varepsilon) = 0$ possesses a continuous bounded solution $\alpha(\tau, \varepsilon)$, $\beta(\tau, \varepsilon)$, $\gamma(\tau, \varepsilon)$ and $w(\varepsilon)$. This completes the proof in the case $K = 0$. The case $K > 0$ proceeds similarly with obvious modifications.

7. An example. We now wish to illustrate the preceding ideas as they arise in a legitimate optimal control problem. While the following example does not illustrate all the complexities that may arise in applying our technique, it does, however, indicate the use to which the technique may be put. We consider the system governed by the scalar equation

$$\dot{x}(t) = 2u(t) \tag{21}$$

with initial condition

$$x(0)=1$$

and the problem of minimizing the performance index

$$J=\int_0^3 (x+u^2)\,dt-3x(3)$$

on the interval [0, 3] subject to the control constraint

$$|u(t)|\leqq 1.$$

The Hamiltonian for this system is

$$H(x, \lambda, u)=x+u^2+2\lambda u.$$

The condition $H_u=0$ gives $\lambda=-u$ so an application of the maximal principle together with the restriction $|u|\leqq 1$ gives that

$$u_{\text{opt}}=\min\{\max(-\lambda, -1), 1\}.$$

From the Euler–Lagrange equations, we have the equation

$$\dot{\lambda}=-\frac{\partial H}{\partial x}=-1$$

and the end condition $\lambda(3)=-3$. Thus

$$\lambda(t)=-t;$$

so from (21) we obtain

$$\dot{x}(t)=\begin{cases}2t & \text{if } 0\leqq t\leqq t_0=1,\\ 2 & \text{if } 1=t_0\leqq t\leqq 3.\end{cases} \tag{22}$$

Finally from the initial condition $x(0)=1$ and the above, we have

$$x(t)=\begin{cases}t^2+1 & \text{for } 0\leqq t\leqq t_0=1,\\ 2t & \text{for } 1=t_0\leqq t\leqq 3,\end{cases}$$

$$\lambda(t)=-t \quad \text{for all } 0\leqq t\leqq 3, \tag{23}$$

and

$$u(t)=\begin{cases}t & \text{for } 0\leqq t\leqq t_0=1,\\ 1 & \text{for } 1\leqq t_0\leqq t\leqq 3.\end{cases}$$

We now use the above and our preceding work to analyze the perturbed system given by

$$\dot{x}=\varepsilon x^3+(2+\varepsilon)u, \tag{24}$$

$$x(0)=1, \tag{25}$$

where we wish to minimize

$$J=\int_0^3 (x+\varepsilon x^2+u^2)\,dt-(3+\varepsilon)x(3) \tag{26}$$

and the restriction $|u| \leqq 1$. The Hamiltonian here is given by

$$H(x, \lambda, u, \varepsilon) = x + \varepsilon x^2 + u^2 + \lambda(\varepsilon x^3 + (2+\varepsilon)u) \tag{27}$$

from which the corresponding equation for the costate variable λ is determined as

$$\dot{\lambda} = -\frac{\partial H}{\partial x} = -1 - 2\varepsilon x - 3\lambda\varepsilon x^2 \tag{28}$$

with boundary condition

$$\lambda(3) = -(3+\varepsilon). \tag{29}$$

For $\varepsilon = 0$, the explicit solution is that given by (27) which we henceforth denote by x_0, λ_0 and u_0. Let $t(\varepsilon)$ denote the smallest value of t for which $|u(t(\varepsilon), \varepsilon)| = 1$, where $u(t, \varepsilon)$ denotes the optimal controller for the perturbed system (24)–(26) at ε. (Note that $t(0) = t_0 = 1$.) Let

$$t = h(\tau, \varepsilon) = \frac{\tau(3-\tau)}{2} t(\varepsilon) + \frac{\tau(\tau-1)}{2}.$$

This change in the time scale will freeze the control switch point $t(\varepsilon)$ at $\tau = 1$ while leaving the initial and terminal times fixed. We define

$$X(\tau, \varepsilon) = x(h(\tau, \varepsilon), \varepsilon), \quad \Lambda(\tau, \varepsilon) = \lambda(h(\tau, \varepsilon), \varepsilon), \quad U(\tau, \varepsilon) = u(h(\tau, \varepsilon), \varepsilon).$$

Thus $U(1, \varepsilon) = 1$ for ε sufficiently small and positive. Substituting the new variables into (24) and (28) we get

$$\frac{dX}{d\tau} = \left[\frac{3-2\tau}{2} t(\varepsilon) + \frac{2\tau-1}{2}\right][\varepsilon X^3 + (2+\varepsilon)U], \tag{30}$$

$$\frac{d\Lambda}{d\tau} = \left[\frac{-3+2\tau}{2} t(\varepsilon) - \frac{2\tau-1}{2}\right][1 + 2\varepsilon X + 3\varepsilon\Lambda X^2], \tag{31}$$

$$X(0, \varepsilon) = x(h, 0, \varepsilon), \varepsilon) = x(0, \varepsilon) = 1 + \varepsilon, \tag{32}$$

$$\Lambda(3, \varepsilon) = \lambda(h(3, \varepsilon), \varepsilon) = \lambda(3, \varepsilon) = -(3+\varepsilon). \tag{33}$$

We now assume that each of the variables in the transformed series has an asymptotic expansion in powers of ε of the form

$$\begin{aligned} X(\tau, \varepsilon) &= x_0(\tau) + X_1(\tau)\varepsilon + X_2(\tau)\varepsilon^2 + \cdots, \\ \Lambda(\tau, \varepsilon) &= \lambda_0(\tau) + \Lambda_1(\tau)\varepsilon + \Lambda_2(\tau)\varepsilon^2 + \cdots, \\ U(\tau, \varepsilon) &= u_0(\tau) + U_1(\tau)\varepsilon + U_2(\tau)\varepsilon^2 + \cdots, \\ t(\varepsilon) &= t_0 + t_1\varepsilon + t_2\varepsilon^2 + \cdots. \end{aligned}$$

Note that $t_0 = 1$. We next insert the assumed expansions into equations (30) and (31), differentiate the resulting equations once with respect to ε, and set $\varepsilon = 0$ to obtain

$$\frac{d}{d\tau} X_1(\tau) = \frac{3-2\tau}{2} t_1 \cdot 2U_0(\tau) + X_0^3(\tau) + 2U_1(\tau) + U_0(\tau), \tag{34}$$

$$\frac{d}{d\tau}\Lambda_1(\tau)=\frac{2\tau-3}{2}t_1-2X_0(\tau)-3\Lambda_0(\tau)X_0^2(\tau), \tag{35}$$

with boundary conditions

$$X_1(0)=1 \quad \text{and} \quad \Lambda_1(3)=-1.$$

In addition, since $U(1,\varepsilon)=U_0(1)=1$, we must have that $U_1(1)=0$. We next attempt to solve (34) and (35). First note that

$$H(X,\Lambda,U,\varepsilon)=X+\varepsilon X^2+U^2+\Lambda(\varepsilon X^2+(2+\varepsilon)U).$$

Since $|U(\tau,\varepsilon)|<1$ for $\tau<1$ and $U(\tau,\varepsilon)$ is optimal, we have for $0\leqq\tau\leqq1$ that

$$0=H_u(X,U,\Lambda,\varepsilon).$$

Differentiating once with respect to ε and setting $\varepsilon=0$, we get

$$2U_1(\tau)+2\Lambda_1(\tau)+\Lambda_0(\tau)=0$$

so

$$U_1(\tau)=-\Lambda_1(\tau)-\tfrac{1}{2}\Lambda_0(\tau), \qquad 0\leqq\tau\leqq1. \tag{36}$$

Thus for $\tau\leqq1$, (34) becomes

$$\frac{d}{d\tau}X_1(\tau)=\frac{3-2\tau}{2}t_1 2U_0(\tau)+X_0^3(\tau)-2\Lambda_1(\tau)-\Lambda_0(\tau)+U_0(\tau).$$

From (23) and (35), we have for $\tau\in[0,1]$ that

$$\frac{d}{d\tau}\Lambda_1(\tau)=3\tau^5+6\tau^3-2\tau^2+(3+t_1)\tau-(2+\tfrac{3}{2}t_1)$$

so

$$\Lambda_1(\tau)=\tfrac{1}{2}\tau^6+\tfrac{3}{2}\tau^4-\tfrac{2}{3}\tau^3+\tfrac{1}{2}(3+t_1)\tau^2-(2+\tfrac{3}{2}t_1)\tau+d_1, \qquad \tau\leqq1.$$

For $\tau>1$ we have from (23), (34) and (35) that

$$\frac{d}{d\tau}\Lambda_1(\tau)=\frac{2\tau-3}{2}t_1-4\tau+12\tau^3$$

so

$$\Lambda_1(\tau)=3\tau^4-\tfrac{1}{2}(4-t_1)\tau^2-\tfrac{3}{2}t_1\tau+d_2, \qquad \tau>1.$$

Now $\Lambda_1(3)=-1$ so we can solve for d_2. Doing so we obtain $d_2=-226$ so $\Lambda(1^+)=-t_1-225$. Thus $\Lambda_1(1^-)$ must equal $-t_1-225$ so $d_1=-\frac{1355}{6}$. We therefore have determined Λ as an explicit function of τ as follows:

$$\Lambda_1(\tau)=\begin{cases}\tfrac{1}{2}\tau^6+\tfrac{3}{2}\tau^4-\tfrac{2}{3}\tau^3+\tfrac{1}{2}(t_1+3)\tau^2-(\tfrac{3}{2}t_1+2)\tau-\frac{1355}{6}, & \tau\leqq1,\\ 3\tau^4-\tfrac{1}{2}(4-t_1)\tau^2-\tfrac{3}{2}t_1\tau-226, & \tau>1.\end{cases} \tag{37}$$

Next from (34), (23) and the above, we have for $\tau\leqq1$ that

$$\frac{dX_1}{d\tau}=\tfrac{4}{3}\tau^3-3t_1\tau^2+(6t_1+6)\tau+\tfrac{1358}{3},$$

so

$$X_1(\tau)=\tfrac{1}{3}\tau^4-t_1\tau^3+(3t_1+3)\tau^2+\tfrac{1358}{3}\tau+c_1, \qquad \tau\leqq 1 \tag{38}$$

Since $X_1(0)=1$, $c_1=1$. Now for $\tau>1$, $U_1(\tau)=0$, so we have that

$$\frac{dX_1}{d\tau}=8\tau^3-2t_1\tau+3t_1+1,$$

so

$$X_1(\tau)=2\tau^4-t_1\tau^2+(3t_1+1)\tau+c_2, \qquad \tau>1. \tag{39}$$

Requiring that $X(1^-)=X(1^+)=457+2t_1$ we get $c_2=454$ so X is determined in (38) and (39).

We are now in position to solve for t_1. Since $H_u(X, \Lambda, U, \varepsilon)=0$ for $\tau<1$, by continuity we must have $H_u(X, \Lambda, U, \varepsilon)=0$ for $\tau=1$. Hence (36) holds for $\tau=1$; i.e., we have that

$$U_1(1)=-\Lambda_1(1)-\tfrac{1}{2}\Lambda_0(1).$$

But $U_1(1)=0$ since $U_0(1)=1$ and $U(1, \varepsilon)=1$ for all ε sufficiently small. Hence

$$\Lambda_1(1)=-\tfrac{1}{2}\Lambda_0(1).$$

From (23), (37) and the above, we have that

$$-t_1-225=(-\tfrac{1}{2})(-1)$$

so

$$t_1=-\tfrac{451}{2}.$$

Thus $t(\varepsilon)=1-\frac{451}{2}\varepsilon+O(\varepsilon^2)$ and X_1, Λ_1 and U_1 are determined as explicit functions of τ from (36)–(39). We may now repeat the procedure to determine t_2, X_2, Λ_2, U_2, etc.

In conclusion the reader will see that the clock function technique serves as a mechanism to allow the treatment of problems involving unbounded controls in much the same framework as bounded controls. The introduction of the time parameter $t(\varepsilon)=t_0+\varepsilon t_1+\varepsilon^2 t_2+\cdots$ when passage to the boundary occurs, however, adds an additional interval boundary condition (see (9)) to be used at each stage of our recursive procedure to solve for X_j, Λ_j and t_j.

REFERENCES

[1] M. I. FREEDMAN AND J. KAPLAN, *Perturbation analysis of an optimal control problem involving bang–bang controls*, Tech. Rep. '74–8, Boston Univ., Boston.

[2] ———, *Use of a nonlinear clock in the perturbation analysis of time optimal control problems*, Tech. Rep. '74–12, Boston Univ., Boston.

[3] L. M. GRAVES, *Implicit functions and differential equations in general analysis*, Amer. Math. Soc. Transl., 29 (1927), pp. 514–552.

[4] A. H. HADDAD AND P. V. KOKOTOVIC, *Note on singular perturbation of linear state regulators*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 279–281.

[5] P. V. KOKOTOVIC AND R. A. YACKEL, *Singular perturbations of linear regulators: Basic theorems*, Ibid., AC-17 (1972), pp. 29–38.

[6] R. E. O'MALLEY, JR., *Singular perturbation of the time invariant linear state regulator problem*, J. Differential Equations, 12 (1972), pp. 117–128.

[7] ———, *The singularly perturbed state regulator problem*, this Journal, 10 (1972), pp. 399–413.
[8] P. SANNUTI, *A note on obtaining reduced order optimal control problems by singular perturbations*, IEEE Trans. on Automatic Control, AC-19, (1974), pp. 256–257.
[9] M. I. FREEDMAN AND B. GRANOFF, *The formal asymptotic solution of a singularly perturbed nonlinear optimal control problem*, J. Optimization Theory Appl., to appear.
[10] B. ANDERSON AND J. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, N.J., 1971.
[11] E. KREINDLER AND A. JAMESON, *Optimality of linear control systems*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 349–357.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

MONOTONE OPERATORS AND THE PROXIMAL POINT ALGORITHM*

R. TYRRELL ROCKAFELLAR†

Abstract. For the problem of minimizing a lower semicontinuous proper convex function f on a Hilbert space, the proximal point algorithm in exact form generates a sequence $\{z^k\}$ by taking z^{k+1} to be the minimizer of $f(z)+(1/2c_k)\|z-z^k\|^2$, where $c_k>0$. This algorithm is of interest for several reasons, but especially because of its role in certain computational methods based on duality, such as the Hestenes-Powell method of multipliers in nonlinear programming. It is investigated here in a more general form where the requirement for exact minimization at each iteration is weakened, and the subdifferential ∂f is replaced by an arbitrary maximal monotone operator T. Convergence is established under several criteria amenable to implementation. The rate of convergence is shown to be "typically" linear with an arbitrarily good modulus if c_k stays large enough, in fact superlinear if $c_k \to \infty$. The case of $T=\partial f$ is treated in extra detail. Application is also made to a related case corresponding to minimax problems.

1. Introduction. Let H be a real Hilbert space with inner product $\langle\cdot,\cdot\rangle$. A multifunction $T: H \to H$ is said to be a *monotone operator* if

$$\langle z-z', w-w'\rangle \geqq 0 \quad \text{whenever} \quad w\in T(z),\ w'\in T(z'). \tag{1.1}$$

It is said to be *maximal monotone* if, in addition, the graph

$$G(T)=\{(z,w)\in H\times H \mid w\in T(z)\} \tag{1.2}$$

is not properly contained in the graph of any other monotone operator $T': H \to H$.

Such operators have been studied extensively because of their role in convex analysis and certain partial differential equations. A fundamental problem is that of determining an element z such that $0\in T(z)$.

For example, if T is the subdifferential ∂f of a lower semicontinuous convex function $f: H \to (-\infty,+\infty]$, $f\not\equiv+\infty$, then T is maximal monotone (see Minty [15] or Moreau [18]), and the relation $0\in T(z)$ means that $f(z)=\min f$. The problem is then one of minimization subject to implicit constraints (the points where $f(z)=+\infty$ being effectively forbidden from the competition).

The basic case of *variational inequalities* corresponds to

$$T(z)=\begin{cases} T_0(z)+N_D(z) & \text{if } z\in D,\\ \varnothing & \text{if } z\notin D,\end{cases} \tag{1.3}$$

where D is a nonempty closed convex subset of H, $T_0: D \to H$ is single-valued, monotone and hemicontinuous (i.e. continuous along each line segment in H with respect to the weak topology), and $N_D(z)$ is the *normal cone* to D at z:

$$N_D(z)=\{w\in H \mid \langle z-u, w\rangle \geqq 0,\ \forall u\in D\}.$$

* Received by the editors July 9, 1975, and in revised form November 17, 1975.

† Department of Mathematics, University of Washington, Seattle, Washington 98195. This work was supported in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under AFOSR Grant 72-2269.

The maximal monotonicity of such a multifunction T was proved by Rockafellar [27]. The relation $0 \in T(z)$ reduces to $-T_0(z) \in N_D(z)$, or the so-called variational inequality:

$$z \in D \quad \text{and} \quad \langle z-u, T_0(z)\rangle \leqq 0 \quad \text{for all} \quad i \in D. \tag{1.4}$$

If D is a cone, this condition can be written as

$$z \in D, \quad -T_0(z) \in D^\circ \quad (\text{the polar of } D), \quad \text{and} \quad \langle z, T_0(z)\rangle = 0,$$

and the problem of finding such a z is an important instance of the well-known *complementarity problem* of mathematical programming.

Another example corresponds to minimax problems. Let H be a product of Hilbert spaces H_1 and H_2, and let $L : H \to [-\infty, +\infty]$ be such that $L(x, y)$ is convex in $x \in H_1$ and concave in $y \in H_2$. For each $z = (x, y)$, let $T_L(z)$ be the set of all $w = (v, u)$ such that

$$\begin{aligned} L(x', y) - \langle x', v\rangle + \langle y, u\rangle &\geqq L(x, y) - \langle x, v\rangle + \langle y, u\rangle \\ &\geqq L(x, y') - \langle x, v\rangle + \langle y', u\rangle \end{aligned} \tag{1.5}$$

$$\text{for all} \quad x' \in H_1, \quad y' \in H_2.$$

If L is "closed and proper" in a certain general sense, then T_L is maximal monotone; see Rockafellar [24]. The global saddle points of L (with respect to minimizing in x and maximizing in y) are the elements $z = (x, y)$ such that $0 \in T_L(z)$.

In this paper, we study a fundamental algorithm for solving $0 \in T(z)$ in the case of an arbitrary maximal monotone operator T. The algorithm is based on the fact (see Minty [14]) that for each $z \in H$ and $c > 0$ there is a unique $u \in H$ such that $z - u \in cT(u)$, or in other words,

$$z \in (I + cT)(u).$$

The operator $P = (I + cT)^{-1}$ is therefore single-valued from all of H into H. It is also *nonexpansive*:

$$\|P(z) - P(z')\| \leqq \|z - z'\|, \tag{1.6}$$

and one has $P(z) = z$ if and only if $0 \in T(z)$. P is called the *proximal mapping* associated with cT, following the terminology of Moreau [18] for the case of $T = \partial f$.

The *proximal point algorithm* generates for any starting point z^0 a sequence $\{z^k\}$ in H by the approximate rule

$$z^{k+1} \approx P_k(z^k), \quad \text{where} \quad P_k = (I + c_k T)^{-1}. \tag{1.7}$$

Here $\{c_k\}$ is some sequence of positive real numbers. In the case of $T = \partial f$, this procedure reduces to

$$z^{k+1} \approx \arg\min_z \phi_k(z), \tag{1.8}$$

where

$$\phi_k(z) = f(z) + \frac{1}{2c_k}\|z - z^k\|^2 \tag{1.9}$$

(see § 4). For T corresponding to a convex-concave function L, it becomes

$$(x^{k+1}, y^{k+1}) \approx \arg \operatorname*{minimax}_{x,y} \Lambda_k(x, y), \tag{1.10}$$

where

$$\Lambda_k(x, y) = L(x, y) + \frac{1}{2c_k}\|x - x^k\|^2 - \frac{1}{2c_k}\|y - y^k\|^2 \tag{1.11}$$

(see § 5).

Results on the convergence of the proximal point algorithm have already been obtained by Martinet for certain cases where $c_k \equiv c$. He showed in [12], [13] that if T is of the form (1.3) with D bounded, and if true equality is taken in (1.7), then z^k converges in the weak topology to a particular z^∞ such that $0 \in T(z^\infty)$. Similarly if $T = \partial f$ and the level sets

$$\{z \in H | f(z) \leqq \alpha\}, \qquad \alpha \in R,$$

are all weakly compact, in which event it is also true that $f(z^k) \downarrow f(z^\infty) = \min f$.

These results of Martinet are based on a more general theorem concerning operators V with the property

$$\|V(z) - V(z')\|^2 \leqq \|z - z'\|^2 - \|(I - V)(z) - (I - V)(z')\|^2. \tag{1.12}$$

This class includes $(I + cT)^{-1}$ (cf. Proposition 1(c) below). If $V : C \to C$ satisfies (1.12), where C is a nonempty, closed, bounded, convex subset of H, then for any starting point $z^0 \in C$ the sequence $\{z^k\}$ generated by $z^{k+1} = V(z^k)$ converges weakly to some fixed point of V. This theorem is a corollary of one of Opial [32] treating iterates of $\lambda I + (1 - \lambda)U$ when U is nonexpansive, $0 < \lambda < 1$. In fact, V satisfies (1.12) if and only if $V = \frac{1}{2}(I + U)$, where U is nonexpansive. Genel and Lindenstrauss [33] have recently furnished an example of such a mapping V for which $\{z^k\}$ does not converge strongly. However, this V does not appear to be of the form $(I + cT)^{-1}$ for $c > 0$ and T maximal monotone.

The question of whether the weak convergence established by Martinet can be improved to strong convergence thus remains open. The answer is known to be affirmative if $T = \partial f$ with f quadratic. This follows from a theorem of Krasnoselskii [10], as has been noted by Kryanev [11]. In the quadratic case, ∂f reduces to a densely defined, single-valued mapping of the form $x \to A(x) - b$, where A is a nonnegative, closed, self-adjoint linear operator. Then the relation $0 \in T(z)$ is equivalent to $A(z) = b$.

Strong convergence of the algorithm in its exact form with $z^{k+1} = P_k(z^k)$ is also assured if c_k is bounded away from zero and T is *strongly monotone* (*with modulus* $\alpha > 0$), i.e., in place of (1.1) one has

$$\langle z - z', w - w' \rangle \geqq \alpha\|z - z'\|^2 \quad \text{whenever} \quad w \in T(z), \quad w' \in T(z'). \tag{1.13}$$

Indeed, the latter condition means that $T' = T - \alpha I$ is monotone, and hence the mapping $P'_k = (I + c'_k T')^{-1}$ is nonexpansive for any $c'_k > 0$; taking $c'_k = c_k(1 + \alpha c_k)^{-1}$ one has

$$P'_k[(1 - \alpha c_k(1 + \alpha c_k)^{-1})I + c_k(1 + \alpha c_k)^{-1}T]^{-1} = [(1 + \alpha c_k)^{-1}(I + c_k T)]^{-1}$$

or

$$P_k(z) = P'_k((1 + \alpha c_k)^{-1} z) \quad \text{for all} \quad z,$$

so that the nonexpansiveness of P_k' yields

$$\|P_k(z)-P_k(z')\| \leqq (1+\alpha c_k)^{-1}\|z-z'\| \quad \text{for all} \quad z, z' \in H. \tag{1.14}$$

In particular, this implies P_k has a unique fixed point, which must then be the unique point z^∞ satisfying $0 \in T(z^\infty)$. One has

$$\|z^{k+1}-z^\infty\| = \|P_k(z^k)-P_k(z^\infty)\| \leqq (1+\alpha c_k)^{-1}\|z^k-z^\infty\| \quad \text{for all} \quad k, \tag{1.15}$$

so if $c_k \geqq c > 0$ for all k sufficiently large the sequence $\{z^k\}$ converges to the solution z^∞ of the problem, not only strongly, but at least as fast as the linear rate with coefficient $(1+\alpha c)^{-1} < 1$. If $c_k \to \infty$, the convergence is *superlinear*:

$$\lim_{k\to\infty} \frac{\|z^{k+1}-z^\infty\|}{\|z^k-z^\infty\|} = 0.$$

Unfortunately, the assumption that T is strongly monotone excludes some of the most important applications, such as to typical problems of convex programming, and it is important therefore to study convergence under weaker assumptions. Of course, from a practical point of view it is also essential to replace the equation $z^{k+1} = P_k(z^k)$ by a looser relation which is computationally implementable for a wide variety of problems.

Two general criteria for the approximate calculation of $P_k(z^k)$ are treated here:

$$\text{(A)} \qquad \|z^{k+1}-P_k(z^k)\| \leqq \varepsilon_k, \qquad \sum_{k=0}^{\infty} \varepsilon_k < \infty,$$

$$\text{(B)} \qquad \|z^{k+1}-P_k(z^k)\| \leqq \delta_k\|z^{k+1}-z^k\|, \qquad \sum_{k=0}^{\infty} \delta_k < \infty.$$

It is shown (Proposition 3) that these are implied respectively by

$$\text{(A')} \qquad \operatorname{dist}(0, S_k(z^{k+1})) \leqq \varepsilon_k/c_k, \qquad \sum_{k=0}^{\infty} \varepsilon_k < \infty,$$

and

$$\text{(B')} \qquad \operatorname{dist}(0, S_k(z^{k+1})) \leqq (\delta_k/c_k)\|z^{k+1}-z^k\|, \qquad \sum_{k=0}^{\infty} \delta_k < \infty,$$

where

$$S_k(z) = T(z) + c_k^{-1}(z-z_k). \tag{1.16}$$

(Note that these conditions are certainly satisfied if $z^{k+1} = P_k(z^k)$.)

We prove under very mild assumptions (Theorem 1) that (A) (or (A')) guarantees (for any starting point z^0) weak convergence of $\{z^k\}$ to a particular solution z^∞ to $0 \in T(z)$, even though there may be more than one solution. (In general, the set of all such points z forms a closed convex set in H, denoted by $T^{-1}(0)$.) The results of Martinet are thereby extended to a much larger class of problems, and with only $z^{k+1} \approx P_k(z^k)$.

When (B) (or (B')) is also satisfied and the multifunction T^{-1} happens to be "Lipschitz continuous at 0", we are able to show (Theorem 2) that the con-

vergence is at least at a linear rate, where the modulus can be brought arbitrarily close to zero by taking c_k large enough. If $c_k \to \infty$, one has superlinear convergence.

In other words, the same convergence properties noted above for the case of strong monotonicity are established under far weaker assumptions. A criterion for the convergence of the algorithm in a finite number of iterations is also furnished (Theorem 3).

The assumption of Lipschitz continuity of T^{-1} at 0 turns out to be very natural in applications to convex programming. It is fulfilled, for instance, under certain standard second-order conditions characterizing a "nice" optimal solution. Such applications, having many ramifications, will be discussed elsewhere [31].

There are actually three distinct types of applications of the proximal point algorithm in convex programming: (i) to $T = \partial f$, where f is the essential objective function in the problem, (ii) to $T = -\partial g$, where g is the concave objective function in the dual problem, and (iii) to the monotone operator T_L corresponding to the convex–concave Lagrangian function.

The second type of application corresponds to the "method of multipliers" of Hestenes [8] and Powell [21]. The relationship with the proximal point algorithm in this case has already been used by Rockafellar [29]. The third type of application yields a new form of the method of multipliers that seems superior, at least in some respects. Although the details will not be treated here, we mention these applications because of their role in motivating the present work.

Some implications of the theorems in this paper for the general cases of $T = \partial f$ or T corresponding to a convex–concave function L are nevertheless discussed in § 4.

Aside from the obvious case of strong monotonicity, or special results for the method of multipliers in convex programming (for a survey, see Bertsekas [5]), there are no rate-of-convergence results relating to the proximal point algorithm prior to those developed here.

For discussion of other methods for solving $0 \in T(z)$ in the case of a maximal monotone operator, we refer to Auslender [2] and Bakushinskii and Polyak [3].

2. Convergence of the general algorithm. Henceforth T is always maximal monotone. In addition to $P_k = (I + c_k T)^{-1}$, we shall make use of the mapping

$$Q_k = I - P_k = (I + (c_k T)^{-1})^{-1}. \tag{2.1}$$

Thus

$$0 \in T(z) \Leftrightarrow P_k(z) = z \Leftrightarrow Q_k(z) = 0. \tag{2.2}$$

PROPOSITION 1.

(a) $z = P_k(z) + Q_k(z)$ *and* $c_k^{-1} Q_k(z) \in T(P_k(z))$ *for all* z.

(b) $\langle P_k(z) - P_k(z'), Q_k(z) - Q_k(z') \rangle \geqq 0$ *for all* z, z'.

(c) $\|P_k(z) - P_k(z')\|^2 + \|Q_k(z) - Q_k(z')\|^2 \leqq \|z - z'\|^2$ *for all* z, z'.

Proof. Part (a) is immediate from the definitions, while (b) is a consequence of (a) and the monotonicity of T. Part (c) follows from (a) and (b) by expanding

$$\|z - z'\|^2 = \|[P_k(z) - P_k(z')] + [Q_k(z) - Q_k(z')]\|^2.$$

Part (c) of Proposition 1 states that property (1.12) holds for P_k and Q_k. If $c_k \equiv c > 0$, the mappings P_k all reduce to a single V to which the Martinet's corollary of Opial's theorem (recalled in § 1 after (1.12)) can be applied with respect to any nonempty closed bounded convex set C such that $V(C) \subset C$. Of course, if V is known to have at least one fixed point in H, then for arbitrary $z^0 \in H$ one can take C to be the closed ball of radius $\|z^0 - \bar{z}\|$ and center $\bar{z}$, where $\bar{z}$ is any fixed point.

In this way one obtains an immediate generalization of Martinet's results for the case of $T = \partial f$ or variational inequalities. *If there exists at least one z satisfying $0 \in T(z)$, then the proximal point algorithm in exact form ($z^{k+1} = P_k(z^k)$) with $c_k \equiv c$ converges weakly from any starting point z^0 to a particular z^∞ satisfying $0 \in T(z^\infty)$.* This should be compared with the still more general Theorem 1 below.

In connection with the existence of solutions to the problem we want to solve, it is worth mentioning the following result (Rockafellar [25, Prop. 2]; this is a generalization of Theorem 2.2 of Browder [7]).

PROPOSITION 2 (see [25]). *Suppose that for some $\tilde{z} \in H$ and $\rho \geqq 0$ one has*

$$\langle z - \tilde{z}, w \rangle \geqq 0 \quad \text{for all} \quad z, w \quad \text{with} \quad w \in T(z), \|z - \tilde{z}\| \geqq \rho. \tag{2.3}$$

Then there exists at least one z satisfying $0 \in T(z)$. (This condition is not only sufficient for existence, but necessary.)

The condition in Proposition 2 holds trivially for example, if the effective domain

$$D(T) = \{z \in H | T(z) \neq \varnothing\} \tag{2.4}$$

is a bounded set. A convenient, weaker condition, which is also sufficient for existence when $T = \partial f$, is the weak compactness of the level sets $\{z \in H | f(z) \leqq \beta\}$, $\beta \in R$.

The relationship between the criteria (A) and (B) on the one hand and (A′) and (B′) on the other is laid out by the next of our preliminary results.

PROPOSITION 3. *The estimate*

$$\|z^{k+1} - P_k(z^k)\| \leqq c_k \operatorname{dist}(0, S_k(z^{k+1}))$$

holds, where S_k is given by (1.16). *Thus* (A′) *implies* (A), *and* (B′) *implies* (B).

Proof. For any $w \in S_k(z^{k+1})$ we have

$$c_k w + z^k \in (I + c_k T)(z^{k+1}),$$

and hence,

$$z^{k+1} = (I + c_k T)^{-1}(c_k w + z^k) = P_k(c_k w + z^k).$$

Then by virtue of the nonexpansiveness of P_k

$$\|z^{k+1} - P_k(z^k)\| = \|P_k(c_k w + z^k) - P_k(z^k)\| \leqq c_k \|w\|.$$

Thus

$$\|z^{k+1} - P_k(z^k)\| \leqq c_k \min \{\|w\| \mid w \in S_k(z^{k+1})\}$$

as claimed.

THEOREM 1. *Let* $\{z^k\}$ *be any sequence generated by the proximal point algorithm under criterion* (A) *(or* (A′)*) with* $\{c_k\}$ *bounded away from zero. Suppose* $\{z^k\}$ *is bounded; this holds under the preceding assumption if and only if there exists at least one solution to* $0 \in T(z)$.

Then $\{z^k\}$ *converges in the weak topology to a point* z^∞ *satisfying* $0 \in T(z^\infty)$, *and*

$$0 = \lim_{k\to\infty} \|Q_k(z^k)\| = \lim_{k\to\infty} \|z^{k+1} - z^k\|. \tag{2.5}$$

Proof. First we verify the asserted sufficient condition for the boundedness of $\{z^k\}$. The necessity of the condition will follow from the last part of the theorem.

Suppose that $\bar{z}$ is a point satisfying $0 \in T(\bar{z})$. We have

$$\|z^{k+1} - \bar{z}\| - \varepsilon_k \leqq \|P_k(z^k) - \bar{z}\| = \|P_k(z^k) - P_k(\bar{z})\| \leqq \|z^k - \bar{z}\|, \tag{2.6}$$

and this furnishes the bound

$$\|z^l - \bar{z}\| \leqq \|z^0 - \bar{z}\| + \sum_{k=0}^{l-1} \varepsilon_k \leqq \|z^0 - \bar{z}\| + \alpha \quad \text{for all} \quad l.$$

Thus $\{z^k\}$ must be bounded.

For the rest of the proof, we assume that $\{z^k\}$ is any bounded sequence satisfying (A). Let $s > 0$ be such that

$$\|z^k\| \leqq s \quad \text{and} \quad \varepsilon_k < s \quad \text{for all} \quad k. \tag{2.7}$$

Then $\{z^k\}$ has at least one weak cluster point z^∞, $\|z^\infty\| \leqq s$.

Our next goal is to demonstrate that $0 \in T(z^\infty)$, but for this purpose it is helpful to show first that the argument can be reduced to the case where it is already known that $T^{-1}(0) \neq \varnothing$. Consider the multifunction T' defined by

$$T'(z) = T(z) + \partial h(z) \quad \text{for all} \quad z \in H,$$

where

$$h(z) = \begin{cases} 0 & \text{if} \quad \|z\| \leqq 2s, \\ +\infty & \text{if} \quad \|z\| > 2s, \end{cases}$$

and consequently

$$\partial h(z) = \begin{cases} \{0\} & \text{if} \quad \|z\| < 2s, \\ \{\lambda z \mid \lambda \geqq 0\} & \text{if} \quad \|z\| = 2s, \\ \varnothing & \text{if} \quad \|z\| > 2s. \end{cases}$$

Observe that ∂h is a maximal monotone operator, because h is a lower semicontinuous proper convex function; its effective domain is

$$D(\partial h) = \{z \mid \|z\| \leqq 2s\}.$$

Furthermore,

$$T'(z) = T(z) \quad \text{if} \quad \|z\| < 2s. \tag{2.8}$$

Since $\|P_k(z^k)\| < 2s$ for all k by (2.7) and condition (A), while

$$c_k^{-1}(z^k - P_k(z^k)) \in T(P_k(z^k))$$

by Proposition 1(a), we have

$$P_k(z^k) \in D(T) \cap \operatorname{int} D(\partial h) \quad \text{for all} \quad k, \tag{2.9}$$

$$P_k(z^k) \in (I + c_k T')^{-1}(z^k) \quad \text{for all} \quad k. \tag{2.10}$$

Inasmuch as $D(T) \cap \operatorname{int} D(\partial h) \neq \emptyset$ by (2.9), we know that T', as the sum of the maximal monotone operators T and ∂h, is itself maximal monotone (Rockafellar [27, Thm. 1]). Hence $P_k' = (I + c_k T')^{-1}$ is actually single-valued, and (2.10) implies

$$P_k(z^k) = P_k'(z^k) \quad \text{for all large} \quad k.$$

Thus the sequence $\{z^k\}$ can be regarded equally well as arising from the proximal point algorithm for T'. The advantage in this is that the effective domain $D(T')$ is bounded, so that $(T')^{-1}(0) \neq \emptyset$ by Proposition 2. Since $T'(z^\infty) = T(z^\infty)$ by (2.8), we could replace T by T' without loss of generality in verifying that $0 \in T(z^\infty)$.

We are therefore justified in assuming, from now on, the existence of a certain $\bar{z}$ such that $0 \in T(\bar{z})$. Applying Proposition 1(c) to $z = z^k$ and $z' = \bar{z}$, we get

$$\|P_k(z^k) - \bar{z}\|^2 + \|Q_k(z^k)\|^2 \leqq \|z^k - \bar{z}\|^2 \quad \text{for all} \quad k. \tag{2.11}$$

Hence,

$$\begin{aligned} \|Q_k(z^k)\|^2 - \|z^k - \bar{z}\|^2 + \|z^{k+1} - \bar{z}\|^2 &\leqq \|z^{k+1} - \bar{z}\|^2 - \|P_k(z^k) - \bar{z}\|^2 \\ &= \langle z^{k+1} - P_k(z^k), (z^{k+1} - \bar{z}) + (P_k(z^k) - \bar{z}) \rangle \\ &\leqq \|z^{k+1} - P_k(z^k)\| (\|z^{k+1} - \bar{z}\| + \|z^k - \bar{z}\|), \end{aligned}$$

and consequently,

$$\|Q_k(z^k)\|^2 \leqq \|z^k - \bar{z}\|^2 - \|z^{k+1} - \bar{z}\|^2 + 2\varepsilon_k(s + \|\bar{z}\|). \tag{2.12}$$

At the same time we have

$$\|z^{k+1} - \bar{z}\| \leqq \|P_k(z^k) - \bar{z}\| + \varepsilon_k \leqq \|z^k - \bar{z}\| + \varepsilon_k,$$

which because of $\sum_{k=0}^{\infty} \varepsilon_k < \infty$ implies the existence of

$$\lim_{k \to \infty} \|z^k - \bar{z}\| = \mu < \infty. \tag{2.13}$$

We can therefore take the limit on both sides of (2.12), obtaining (2.5), because

$$\|Q_k(z^k)\| = \|(z^k - z^{k+1}) + (z^{k+1} - P_k(z^k))\| \geqq \|z^{k+1} - z^k\| - \varepsilon_k.$$

It follows that

$$c_k^{-1} Q_k(z^k) \to 0 \quad \text{strongly}, \tag{2.14}$$

the numbers c_k being bounded away from zero.

Observe next that Proposition 1(a) entails

$$0 \leqq \langle z - P_k(z^k), w - c_k^{-1} Q_k(z^k) \rangle \quad \text{for all} \quad k \quad \text{if} \quad w \in T(z). \tag{2.15}$$

Since z^∞ is a weak cluster point of $\{z^k\}$ and $\|z^{k+1}-P_k(z^k)\| \to 0$, it is also a weak cluster point of $\{P_k(z^k)\}$. Then (2.14) and (2.15) yield

$$0 \leqq \langle z - z^\infty, w\rangle \quad \text{for all} \quad z, w \quad \text{satisfying} \quad w \in T(z).$$

This implies, in view of the maximality of T, that $0 \in T(z^\infty)$.

The next step is to show that there cannot be more than one weak cluster point of $\{z^k\}$. Suppose there were two: $z_1^\infty \neq z_2^\infty$. Then $0 \in T(z_i^\infty)$ for $i = 1, 2$, as just seen, so that each z_i^∞ can play the role of $\bar{z}$ in (2.13), and we get the existence of the limits

$$\lim_{k\to\infty} \|z^k - z_i^\infty\| = \mu_i < \infty \quad \text{for} \quad i = 1, 2. \tag{2.16}$$

Writing

$$\|z^k - z_2^\infty\|^2 = \|z^k - z_1^\infty\|^2 + 2\langle z^k - z_1^\infty, z_1^\infty - z_2^\infty\rangle + \|z_1^\infty - z_2^\infty\|^2,$$

we see that the limit of $\langle z^k - z_1^\infty, z_1^\infty - z_2^\infty\rangle$ must also exist and

$$2 \lim_{k\to\infty} \langle z^k - z_1^\infty, z_1^\infty - z_2^\infty\rangle = \mu_2^2 - \mu_1^2 - \|z_1^\infty - z_2^\infty\|^2.$$

But this limit cannot be different from 0, because z_1^∞ is a weak cluster point $\{z^k\}$. Therefore

$$\mu_2^2 - \mu_1^2 = \|z_1^\infty - z_2\|^2 > 0.$$

However, the same argument works with z_1^∞ and z_2^∞ reversed, so that also $\mu_1^2 - \mu_2^2 > 0$. This is a contradiction which establishes the uniqueness of z^∞.

(The uniqueness argument just given closely follows the one of Martinet [12], and it was also suggested to the author by H. Brézis.)

Counterexample. The convergence of $\{z^k\}$ in Theorem 1 may fail if instead of $\sum_{k=0}^\infty \varepsilon_k < \infty$ one has only $\varepsilon_k \to 0$, even when H is one-dimensional. This can be seen for any maximal monotone T such that the set $T^{-1}(0) = \{z \mid 0 \in T(z)\}$, which is known always to be convex, contains more than one element. Then $T^{-1}(0)$ contains a nonconvergent sequence $\{z^k\}$ with

$$\|z^{k+1} - z^k\| \to 0$$

but

$$\sum_{k=0}^{\infty} \|z^{k+1} - z^k\| = \infty.$$

We have $P_k(z^k) = z^k$ and therefore a counterexample with $\varepsilon_k = \|z^{k+1} - z^k\|$. In particular, all this holds for $T = \partial f$ if the convex function f attains its minimum nonuniquely.

3. Rate of convergence. We shall say that T^{-1} is *Lipschitz continuous at* 0 (*with modulus* $a \geqq 0$) if there is a unique solution $\bar{z}$ to $0 \in T(z)$ (i.e. $T^{-1}(0) = \{\bar{z}\}$), and for some $\tau > 0$ we have

$$\|z - \bar{z}\| \leqq a\|w\| \quad \text{whenever} \quad z \in T^{-1}(w) \quad \text{and} \quad \|w\| \leqq \tau. \tag{3.1}$$

THEOREM 2. *Let* $\{z^k\}$ *be any sequence generated by the proximal point algorithm using criterion* (B) (*or* (B')) *with* $\{c_k\}$ *nondecreasing* ($c_k \uparrow c_\infty \leqq \infty$).

Assume that $\{z^k\}$ *is bounded* (*cf. Theorem* 1) *and that* T^{-1} *is Lipschitz continuous at* 0 *with modulus* a; *let*

$$\mu_k = a/(a^2+c_k^2)^{1/2} < 1.$$

Then $\{z^k\}$ *converges strongly to* $\bar{z}$, *the unique solution to* $0 \in T(z)$. *Moreover, there is an index* $\bar{k}$ *such that*

$$\|z^{k+1}-\bar{z}\| \leqq \theta_k \|z^k-\bar{z}\| \quad \textit{for all} \quad k \geqq \bar{k}, \tag{3.2}$$

where

$$1 > \theta_k \equiv (\mu_k+\delta_k)/(1-\delta_k) \geqq 0 \quad \textit{for all} \quad k \geqq \bar{k}, \tag{3.3}$$

$$\theta_k \to \mu_\infty \quad (\textit{where} \quad \mu_\infty = 0 \quad \textit{if} \quad c_\infty = \infty). \tag{3.4}$$

Proof. The sequence $\{z^k\}$, being bounded, also satisfies criterion (A) for $\varepsilon_k = \delta_k\|z^{k+1}-z^k\|$, so the conclusions of Theorem 1 are in force. We have

$$\|Q_k(z^k)\| = \|z^k - P_k(z^k)\| \leqq \|z^k - z^{k+1}\| + \|z^{k+1} - P_k(z^k)\|,$$

so that

$$\|c_k^{-1}Q_k(z^k)\| \leqq c_k^{-1}(1+\delta_k)\|z^{k+1}-z^k\| \quad \text{for all} \quad k,$$

where $\|z^k - z^{k+1}\| \to 0$ (Theorem 1). Choose $\tilde{k}$ so that

$$c_k^{-1}(1+\delta_k)\|z^{k+1}-z^k\| < \tau \quad \text{for all} \quad k \geqq \tilde{k}. \tag{3.5}$$

Then $\|c_k^{-1}Q_k(z^k)\| \leqq \tau$ for $k \geqq \tilde{k}$. But $P_k(z^k) \in T^{-1}(c_k^{-1}Q_k(z^k))$ by Proposition 1(a). The Lipschitz condition (3.1) can therefore be invoked for $w = c_k^{-1}Q_k(z^k)$ and $z = P_k(z^k)$ if k is sufficiently large:

$$\|P_k(z^k)-\bar{z}\| \leqq a\|c_k^{-1}Q_k(z^k)\| \quad \text{for all} \quad k \geqq \tilde{k}. \tag{3.6}$$

We next apply (2.2) and Proposition 1(c) to $z = \bar{z}$ and $z' = z^k$ to obtain

$$\|\bar{z} - P_k(z^k)\|^2 + \|Q_k(z^k)\|^2 \leqq \|\bar{z} - z^k\|^2,$$

which via (3.6) yields

$$\|P_k(z^k)-\bar{z}\|^2 \leqq [(a/c_k)^2/(1+(a/c_k)^2)]\|z^k-\bar{z}\|^2,$$

or in other words

$$\|P_k(z^k)-\bar{z}\| \leqq \mu_k\|z^k-\bar{z}\| \quad \text{if} \quad k \geqq \bar{k}. \tag{3.7}$$

But

$$\|z^{k+1}-\bar{z}\| \leqq \|z^{k+1}-P_k(z^k)\| + \|P_k(z^k)-\bar{z}\|,$$

where under (B) we have

$$\|z^{k+1}-P_k(z^k)\| \leqq \delta_k\|z^{k+1}-z^k\| \leqq \delta_k\|z^{k+1}-\bar{z}\| + \delta_k\|z^k-\bar{z}\|.$$

Therefore by (3.7),

$$\|z^{k+1}-\bar{z}\| \leqq \delta_k\|z^{k+1}-\bar{z}\| + \mu_k\|z^k-\bar{z}\| + \delta_k\|z^k-\bar{z}\| \quad \text{if} \quad k \geqq \tilde{k}.$$

This inequality produces the one in (3.2) if $\bar{k} \geqq \tilde{k}$ is taken so that (3.3) holds, as is possible since $1 > \mu_k \downarrow \mu_\infty$ and $\delta_k \to 0$.

Remark 1. The proof shows that the estimates in Theorem 2 are valid for any $\bar{k}$ such that (3.3) holds and, for some $\tilde{k} \leqq \bar{k}$, also (3.5) holds. To cite a simple

specific case, let us suppose that

$$\delta_k \leqq \tfrac{1}{4} \quad \text{for all} \quad k,$$

and, as can easily be estimated explicitly for instance if the effective domain $D(T)$ is bounded, that for a certain $d>0$,

$$\|z^{k+1}-z^k\| \leqq d \quad \text{for all} \quad k.$$

It may then be seen that the estimates in Theorem 2 are valid if $\bar{k}$ is such that

$$c_k \geqq 2 \max\{a, d/\tau\} \quad \text{for all} \quad k \geqq \bar{k}.$$

Remark 2. If we replace the condition on δ_k in (B) by the assumption that (A) is satisfied and

$$\delta_k \to \delta_\infty < \tfrac{1}{2}(1-\mu_\infty), \tag{3.8}$$

then all the conclusions of Theorem 2 hold, except that

$$\theta_k \to \theta_\infty = (\mu_\infty+\delta_\infty)/(1-\delta_\infty) < 1.$$

Since

$$\mu_\infty = a/(a^2+c_\infty^2)^{1/2},$$

the inequality (3.8) holds in particular if $\delta_\infty < \frac{1}{2}$ and $c_k \uparrow \infty$.

The next two results help illuminate the Lipschitz condition in Theorem 2.

We shall say that a multifunction $S : H \to H$ is *differentiable* at a point $\bar{w}$ if $S(\bar{w})$ consists of a single element $\bar{z}$ and there is a continuous linear transformation $A : H \to H$ such that, for some $\delta>0$,

$$\emptyset \neq S(\bar{w}+w)-\bar{z}-Aw \subset o(\|w\|)B \quad \text{when} \quad \|w\| \leqq \delta,$$

where B is the closed unit ball and

$$o(\|w\|)/\|w\| \downarrow 0 \quad \text{as} \quad \|w\| \downarrow 0.$$

We then write $A = \nabla S(\bar{w})$. This coincides with the usual notion of differentiability (in the sense of Fréchet), if S is single-valued on a neighborhood of $\bar{w}$.

PROPOSITION 4. *The condition of Lipschitz continuity in Theorem* 2 *is satisfied if* T^{-1} *is differentiable at* 0. *In particular, it is satisfied if there is a* $\bar{z}$ *such that* $0 \in T(\bar{z})$ *and* T *is single-valued and continuously differentiable in a neighborhood of* $\bar{z}$, *with* $\nabla T(\bar{z})$ *invertible* (*i.e. having all of* H *as its range*).

Proof. If T^{-1} is differentiable at 0 and $A = \nabla T^{-1}(0)$, there is a unique $\bar{z}$ satisfying $0 \in T(\bar{z})$, and we have

$$T^{-1}(w)-\bar{z}-Aw \subset o(\|w\|)B \quad \text{when} \quad \|w\| \leqq \delta.$$

Thus there exist $a_0 \geqq 0$ and $\varepsilon>0$ such that

$$\|z-\bar{z}-Aw\| \leqq a_0\|w\| \quad \text{whenever} \quad z \in T^{-1}(w), \quad \|w\| \leqq \varepsilon.$$

It follows that

$$\|z-\bar{z}\| \leqq a_0\|w\|+\|A\|\cdot\|w\| \quad \text{whenever} \quad w \in T(z), \quad \|w\| \leqq \varepsilon.$$

Thus (3.1) holds for $a = a_0+\|A\|$. The second assertion then follows from the first

by way of the implicit function theorem [9]: under these assumptions T^{-1} is single-valued and continuously differentiable on a neighborhood of 0.

PROPOSITION 5. *Suppose T^{-1} is Lipschitz continuous globally, i.e. T^{-1} is everywhere single-valued and satisfies*

$$\|T^{-1}(w)-T^{-1}(w')\| \leqq a\|w-w'\| \quad \text{for all} \quad w, w',$$

where $a \geqq 0$; this is true in particular if T is strongly monotone with modulus $\alpha > 0 (a = \alpha^{-1})$. Then the explicit assumption that $\{z^k\}$ is bounded is superfluous for the conclusions of Theorem 2, and the estimate (3.2) is valid for any $\bar{k}$ large enough that (3.3) holds.

Proof. The proof of Theorem 2 works in this case with $\tilde{k} = 0$. If T is strongly monotone, we have (1.13) for some $\alpha > 0$. Then the operator $T' = T - \alpha I$ is monotone and hence $P = (I + \alpha^{-1}T')^{-1}$ is nonexpansive. But $T = \alpha P^{-1}$, so

$$T^{-1}(w) = P(\alpha^{-1}w) \quad \text{for all} \quad w,$$

and in particular from the nonexpansiveness of P:

$$\|T^{-1}(w)-T^{-1}(w')\| \leqq \alpha^{-1}\|w-w'\| \quad \text{for all} \quad w, w'. \tag{3.9}$$

Finally, we describe a very special but noteworthy case where the algorithm can converge in finitely many iterations. This result was suggested by one of Bertsekas [4] for the method of multipliers in convex programming.

THEOREM 3. *Let $\{z^k\}$ be any sequence generated by the proximal point algorithm under any of the criteria (A), (A'), (B) or (B') with $\{c_k\}$ bounded away from zero. Suppose that $\{z^k\}$ is bounded (cf. the conditions in Theorem 1) and there exists $\bar{z}$ such that $0 \in \operatorname{int} T(\bar{z})$. Then*

$$z^\infty = \bar{z} = P_k(z^k) \quad \text{for all} \quad k \quad \text{sufficiently} \quad \text{large.} \tag{3.10}$$

Hence under (A) (or (A')) one has

$$\|z^k - \bar{z}\| \leqq \varepsilon_k \quad \text{for all} \quad k \quad \text{sufficiently} \quad \text{large,}$$

while under (B) (or (B')) with $c_k \uparrow c_\infty \leqq \infty$ one has the estimates (3.2) and (3.5) for

$$\theta_k = \delta_k/(1-\delta_k) \to 0.$$

Thus in particular, the proximal point algorithm in its exact form (i.e. with $z^{k+1} = P_k(z^k)$) gives convergence to $\bar{z}$ in a finite number of iterations from any starting point z^0.

Proof. We demonstrate first that T^{-1} is single-valued and constant on a neighborhood of 0:

$$T^{-1}(w) = \bar{z} \quad \text{if} \quad \|w\| < \varepsilon. \tag{3.11}$$

Let $\varepsilon > 0$ be chosen so that $\|w\| < \varepsilon$ implies $w \in \operatorname{int} T(\bar{z})$. Taking any z, $w \in T(z)$, and w' with $\|w\| < \varepsilon$, we have

$$0 \leqq \langle z - \bar{z}, w - w' \rangle$$

by the monotonicity of T. Therefore

$$\sup \langle z - \bar{z}, w' \rangle \leqq \langle z - \bar{z}, w \rangle \quad \text{whenever} \quad w \in T(z), \quad \|w'\| < \varepsilon,$$

so that

$$\varepsilon\|z-\bar{z}\| \leqq \|z-\bar{z}\| \cdot \|w\| \quad \text{whenever} \quad w \in T(z).$$

Thus if $z \neq \bar{z}$ we have $\|w\| \geqq \varepsilon$ for all $w \in T(z)$. Stated another way, if $\|w\| < \varepsilon$ and $z \in T^{-1}(w)$, then $z = \bar{z}$, which is the same assertion as (3.11).

Our hypothesis subsumes that of Theorem 1, and hence we know as in Theorem 1 that $\|c_k^{-1}Q_k(z^k)\| \to 0$. However, $P_k(z^k) \in T^{-1}(c_k^{-1}Q_k(z^k))$ by Proposition 1(a). Therefore (3.11) implies (3.10), and everything else in Theorem 3 follows immediately, the Lipschitz condition in Theorem 2 being fulfilled with $a = 0$.

4. Application to minimization. Let $f : H \to (-\infty, +\infty]$ be a lower semicontinuous convex function which is not identically $+\infty$. Then, as noted in the introduction, the multifunction $T = \partial f$ is maximal monotone, where

$$\begin{aligned} w \in \partial f(z) &\Leftrightarrow f(z') \geqq f(z) + \langle z'-z, w\rangle \quad \text{for all} \quad z' \\ &\Leftrightarrow z \in \arg\min (f - \langle \cdot, w\rangle). \end{aligned} \tag{4.1}$$

Since in particular

$$0 \in \partial f(z) \Leftrightarrow z \in \arg\min f,$$

the proximal point algorithm for $T = \partial f$ is a method for minimizing f. We collect here some facts relevant to this special case.

Recall that a function $\phi : H \to (-\infty, +\infty]$ is said to be *strongly convex* (*with modulus* α) if $\alpha > 0$ and

$$\phi((1-\lambda)z + \lambda z') \leqq (1-\lambda)\phi(z) + \lambda\phi(z') - \tfrac{1}{2}\alpha\lambda(1-\lambda)\|z-z'\|^2 \quad \text{for all} \quad z, z' \quad \text{if} \quad 0 < \lambda < 1. \tag{4.2}$$

THEOREM 4. *Let $T = \partial f$. Then $S_k = \partial\phi_k$ in criteria* (A′) *and* (B′), *where ϕ_k is the function defined by* (1.9), *and ϕ_k is lower semicontinuous and strongly convex with modulus $1/c_k$. Furthermore, if $\{z^k\}$ is any sequence generated by the proximal point algorithm under the hypothesis of Theorem* 1 *with criterion* (A′), *then $z^k \to z^\infty$ weakly, where $f(z^\infty) = \min f$ and*

$$f(z^{k+1}) - f(z^\infty) \leqq c_k^{-1}\|z^{k+1} - z^\infty\|(\varepsilon_k + \|z^{k+1} - z^k\|) \to 0. \tag{4.3}$$

Proof. The strong convexity of ϕ follows directly from formula (1.9). Subdifferentiating both sides of this formula, we also get

$$\partial\phi_k(z) = \partial f(z) + c_k^{-1}(z - z^k) \equiv S_k(z) \quad \text{for all} \quad z.$$

(For the relevant rule of subdifferentiation, see Moreau [17] or Rockafellar [22, Thm. 3].) To establish (4.3), let w^k denote the unique element of $\partial\phi_k(z^{k+1})$ nearest the origin. (This exists, because $\partial\phi_k(z^{k+1})$ is a closed convex set which, since (A′) is supposed to hold, is nonempty.) Then

$$w^k - c_k^{-1}(z^{k+1} - z^k) \in T(z^{k+1}) = \partial f(z^{k+1}),$$

where

$$\|w^k\| \leqq \varepsilon_k / c_k \to 0. \tag{4.4}$$

Let z^∞ be the weak limit of $\{z^k\}$ (Theorem 1). Then $0 \in \partial f(z^\infty)$, and the defining inequality for subgradients yields

$$f(z^{k+1}) + \langle z^\infty - z^{k+1}, w^k - c_k^{-1}(z^{k+1} - z^k)\rangle \leqq f(z^\infty) = \min f,$$

so that

$$f(z^{k+1}) - f(z^\infty) \leqq \|z^{k+1} - z^\infty\|(\|w^k\| + c_k^{-1}\|z^{k+1} - z^k\|).$$

Applying (4.4) and (2.5), we reach the desired conclusion (4.3).

Remark 3. The quantity $\operatorname{dist}(0, \partial\phi_k(z^{k+1}))$ occurring in criteria (A′) and (B′) for $T = \partial f$ is generally convenient as a measure of how near z^{k+1} is to being a minimizer of ϕ_k. Exact minimization corresponds, of course, to $\operatorname{dist}(0, \partial\phi_k(z^{k+1})) = 0$. Many methods that might be used for minimizing ϕ_k depend on the calculation of gradients or subgradients, and one can use the estimate

$$\operatorname{dist}(0, \partial\phi_k(z^{k+1})) \leqq \|u\| \quad \text{for any} \quad u \in \partial\phi_k(z^{k+1}).$$

This is not the place to describe all of the possible structures of $\partial\phi_k$ corresponding to minimization problems of different types, but we nevertheless mention an important case. Suppose f is of the form

$$f(z) = \begin{cases} f_0(z) & \text{if} \quad z \in D, \\ +\infty & \text{if} \quad z \in D, \end{cases}$$

where D is a nonempty closed convex set and f_0 is a function which is convex on D and differentiable on a neighborhood of D. Then minimizing f on H is equivalent to minimizing f_0 on D, while minimizing ϕ_k on H is equivalent to minimizing

$$\phi_k^0(z) = f_0(z) + \tfrac{1}{2}\alpha\|z - z^k\|^2$$

on D. Furthermore,

$$\partial\phi_k(z) = \nabla\phi_k^0(z) + N_D(z),$$

where $N_D(z)$ is the normal cone to D at z, and hence $\operatorname{dist}(0, \partial\phi_k(z^{k+1}))$ is the norm of the projection of $-\nabla\phi_k^0(z^{k+1})$ on the tangent cone to D at z^{k+1} (where $z^{k+1} \in D$).

In particular, if $D = H$, i.e., f itself is differentiable on all of H, we have

$$\operatorname{dist}(0, S_k(z^{k+1})) = \|\nabla\phi_k(z^{k+1})\|$$

in (A′) and (B′).

It remains now to show how the various conditions in the hypotheses of Theorems 2 and 3 are realized in the case of $T = \partial f$.

Let f^* be the lower semicontinuous convex function conjugate to f. Thus $\partial f^* = T^{-1}$ for $T = \partial f$. (For the theory of conjugate functions, see [19], [30].)

PROPOSITION 6. *The following conditions are equivalent for* $T = \partial f$:

(a) *T is strongly monotone with modulus α,*

(b) *f is strongly convex with modulus α,*

(c) *whenever $w \in \partial f(z)$, one has for all $z' \in H$:*

$$f(z') \geqq f(z) + \langle z' - z, w\rangle + \tfrac{1}{2}\alpha\|z' - z\|^2.$$

Proof. (b)$\Rightarrow$(a). Suppose $w \in T(z)$ and $w' \in T(z')$, and let $0 < \lambda < 1$. Then

$$f((1-\lambda)z + \lambda z') \geqq f(z) + \langle [(1-\lambda)z + \lambda z'] - z, w\rangle$$
$$= f(z) + \lambda\langle z' - z, w\rangle,$$

and hence by (4.2) for f:

$$-\lambda f(z) + \lambda f(z') - \tfrac{1}{2}\alpha\lambda(1-\lambda)\|z - z'\|^2 \geqq \lambda\langle z' - z, w\rangle,$$

or equivalently,

$$\langle z - z', w\rangle \geqq \tfrac{1}{2}\alpha(1-\lambda)\|z - z'\|^2 + f(z) - f(z').$$

By symmetry it is also true that

$$\langle z' - z, w'\rangle \geqq \tfrac{1}{2}\alpha(1-\lambda)\|z' - z\|^2 + f(z') - f(z),$$

and in adding these two inequalities we obtain

$$\langle z - z', w - w'\rangle \geqq \alpha(1-\lambda)\|z' - z\|^2.$$

This holds for arbitrary $\lambda \in (0, 1)$, so it must also hold for $\lambda = 0$, which is the assertion of (a).

(a)$\Rightarrow$(c). As observed in the proof of Proposition 5, the strong monotonicity implies that T^{-1} is single-valued and satisfies the global Lipschitz condition (3.9). But $T^{-1} = \partial f^*$. In particular, therefore, ∂f^* is single-valued and continuous everywhere, from which it follows that f^* is differentiable everywhere and ∇f^* reduces to the gradient mapping of f^* (see Asplund/Rockafellar [1, p. 461]). For any w and w', we have

$$\|\nabla f^*(w + t(w' - w)) - \nabla f^*(w)\| \leqq (t/\alpha)\|w' - w\| \quad \text{for} \quad t > 0,$$

so that

$$\langle \nabla f^*(w + t(w' - w)), w' - w\rangle \leqq \langle \nabla f^*(w), w' - w\rangle + (t/\alpha)\|w' - w\|^2 \quad \text{for} \quad t > 0.$$

From this we obtain

$$f^*(w') - f^*(w) = \int_0^1 \langle \nabla f^*(w + t(w' - w)), w' - w\rangle\, dt$$
$$\leqq \langle \nabla f^*(w), w' - w\rangle + \frac{1}{2\alpha}\|w' - w\|^2.$$

Fixing arbitrary z and w with $w \in \partial f(z)$, we have $z = \nabla f^*(w)$ and $f(z) + f^*(w) = \langle z, w\rangle$. Then for any z',

$$f(z') = f^{**}(z') = \sup_{w' \in H}\{\langle z', w'\rangle - f^*(w')\}$$
$$\geqq \sup_{w' \in H}\left\{\langle z', w'\rangle - f^*(w) - \langle \nabla f^*(w), w' - w\rangle - \frac{1}{2\alpha}\|w' - w\|^2\right\}$$
$$= \sup_{w' \in H}\left\{f(z) + \langle z' - z, w'\rangle - \frac{1}{2\alpha}\|w' - w\|^2\right\}$$
$$= f(z) + \langle z' - z, w\rangle + \tfrac{1}{2}\alpha\|z' - z\|^2.$$

Thus (c) holds.

(c)$\Rightarrow$(b). Let $G=\{(z, w)|w\in\partial f(z)\}$, and for each $(z, w)\in G$ define the functions $g_{z,w}$ and $h_{z,w}$ by

$$g_{z,w}(z')=f(z)+\langle z'-z, w\rangle+\tfrac{1}{2}\alpha\|z'-z\|^2,$$

$$h_{z,w}(z')=f(z)+\langle z'-z, w\rangle.$$

Then $f\geqq g_{z,w}\geqq h_{z,w}$. It is a known fact, however, that

$$f(z')=\sup_{(z,w)\in G} h_{z,w}(z') \quad \text{for all} \quad z'$$

(Brønsted/Rockafellar [6, Thm. 2]). Hence

$$f(z')=\sup_{(z,w)\in G} g_{z,w}(z').$$

Each function $g_{z,w}$ is strongly convex with modulus α, and therefore f has this same property. This completes the proof of Proposition 6.

PROPOSITION 7. *The following conditions are equivalent for $T=\partial f$ and $\bar{z}\in H$.*

(a) *T^{-1} is Lipschitz continuous at 0, and $\bar{z}$ is the unique solution to $0\in T(z)$.*

(b) *$\bar{z}$ is the unique minimizing point for f, and*

$$\liminf_{z\to\bar{z}}\frac{f(z)-f(\bar{z})}{\|z-\bar{z}\|^2}>0.$$

(c) *$\bar{z}$ is the unique element of $\partial f^*(0)$, and*

$$\limsup_{u\to 0}[(f^*(u)-f^*(0)-\langle\bar{z}, u\rangle)/\|u\|^2]<\infty.$$

Proof. (a)$\Rightarrow$(c). Since $T^{-1}=\partial f^*$, we have

$$\|z-\bar{z}\|\leqq a\|w\| \quad \text{whenever} \quad z\in\partial f^*(w) \quad \text{and} \quad \|w\|\leqq\varepsilon. \tag{4.5}$$

This implies the boundedness of the set

$$\bigcup_{\|w\|\leqq\varepsilon}\partial f^*(w), \tag{4.6}$$

which contains $\bar{z}$; in other words, ∂f^* is locally bounded at 0, which is a point of the effective domain

$$D(\partial f^*)=\{w|\partial f^*(z)\neq\varnothing\}. \tag{4.7}$$

But ∂f^* is a maximal monotone operator, so this property necessitates $0\in$ int $D(\partial f^*)$ (see Rockafellar [25, Thm. 1]). Since

$$D(\partial f^*)\subset\operatorname{dom} f^*=\{w|f^*(w)<\infty\} \tag{4.8}$$

it follows that f^* is finite on a neighborhood of 0. This implies in turn that f^* is continuous on a neighborhood of 0 [23, Cor. 7c] and hence that for all u in some neighborhood of 0, say for $\|u\|\leqq\delta(0<\delta\leqq\varepsilon)$ we have $\partial f^*(u)$ nonempty weakly compact and

$$f^{*\prime}(w; u)=\max\{\langle z, u\rangle|z\in\partial f^*(w)\} \quad \text{for all} \quad u\in H, \tag{4.9}$$

where

$$f^{*\prime}(w; u)=\lim_{\lambda\downarrow 0}[f^*(w+\lambda u)-f^*(w)]/\lambda.$$

(Moreau, [17]). Moreover (4.5) and (4.9) give the estimate

$$f^{*\prime}(w; u) \leqq \langle \bar{z}, u\rangle + a\|w\| \cdot \|u\| \quad \text{if} \quad \|w\| \leqq \delta. \tag{4.10}$$

Observe next that if $\|u\| \leqq \delta$ and $\zeta(t) = f^*(tu)$, then ζ is a finite continuous convex function on $[0, 1]$, and hence

$$\zeta(1) = \zeta(0) + \int_0^1 \zeta'_+(t)\, dt,$$

where ζ'_+ is the right derivative of ζ [26, Cor. 24.2.1]. This formula says that

$$f^*(u) = f^*(0) + \int_0^1 f^{*\prime}(tu; u)\, dt,$$

and hence by (4.10),

$$f^*(u) \leqq f^*(0) + \langle \bar{z}, u\rangle + \tfrac{1}{2}a\|u\|^2 \quad \textit{if} \quad \|u\| \leqq \delta. \tag{4.11}$$

Therefore (c) is valid.

(c) $\Rightarrow$ (b). Under (c), we have (4.11) for some $a > 0$ and $\delta > 0$. Let

$$\xi(s) = \begin{cases} \tfrac{1}{2}as^2 & \text{if} \quad |s| \leqq \delta, \\ +\infty & \text{if} \quad |s| > \delta. \end{cases}$$

Then (4.11) can be expressed as

$$f^*(u) - f^*(0) - \langle \bar{z}, u\rangle \leqq \xi(\|u\|) \quad \text{for all} \quad u \in H, \tag{4.12}$$

where $\xi(\|u\|)$ is convex in u. Taking conjugates on both sides, we obtain

$$f(\bar{z} + v) + f^*(0) \geqq \xi^*(\|v\|) \quad \text{for all} \quad v \in H,$$

where

$$\xi^*(r) = \begin{cases} \tfrac{1}{2}a^{-1}r^2 & \text{if} \quad |r| \leqq a\delta, \\ \delta|r| - \tfrac{1}{2}a\delta^2 & \text{if} \quad |r| \geqq a\delta. \end{cases} \tag{4.13}$$

But

$$f(\bar{z}) + f^*(0) = \langle \bar{z}, 0\rangle, \tag{4.14}$$

since $\bar{z} \in \partial f^*(0)$. Therefore

$$f(z) - f(\bar{z}) \geqq \xi^*(\|z - \bar{z}\|) \quad \text{for all} \quad z \in H, \tag{4.15}$$

and in particular

$$f(z) - f(\bar{z}) \geqq \tfrac{1}{2}a^{-1}\|z - \bar{z}\|^2 \quad \text{if} \quad \|z - \bar{z}\| \leqq a\delta.$$

Thus (b) holds.

(b) $\Rightarrow$ (c). The hypothesis means that for a certain $a > 0$ and $\delta > 0$ we have

$$f(z) - f(\bar{z}) \geqq \tfrac{1}{2}a^{-1}\|z - \bar{z}\|^2 \quad \text{whenever} \quad \|z - \bar{z}\| \leqq 2a\delta. \tag{4.16}$$

We shall show first that this implies (4.15). Of course, since $\xi(s) \geqq \frac{1}{2}as^2$ for all $s \in R$ we have (taking conjugates on both sides) that $\xi^*(r) \leqq \frac{1}{2}a^{1-}r^2$ for all $r \in R$, and hence the inequality in (4.15) follows from the one in (4.16) if $\|z - \bar{z}\| \leqq 2a\delta$.

Suppose therefore that $\|z-\bar{z}\|>2a\delta$ and let $\lambda=2a\delta/\|z-\bar{z}\|<1$. The point

$$\tilde{z}=(1-\lambda)\bar{z}+\lambda z=\bar{z}+\lambda(z-\bar{z})$$

then satisfies $\|\tilde{z}-\bar{z}\|=2a\delta$, so that by (4.16),

$$f(\bar{z})+\tfrac{1}{2}a^{-1}\|\tilde{z}-\bar{z}\|^2\leqq f(\tilde{z})\leqq(1-\lambda)f(\bar{z})+\lambda f(z).$$

Thus

$$f(z)-f(\bar{z})\geqq\tfrac{1}{2}\lambda^{-1}a^{-1}\|\tilde{z}-\bar{z}\|^2=\delta\|z-\bar{z}\|\geqq\xi^*(\|z-\bar{z}\|),$$

and (4.15) is justified. We pass now to the conjugate on each side of (4.15) to obtain

$$f^*(u)+f(\bar{z})\leqq\xi(\|u\|)+\langle u,\bar{z}\rangle\quad\text{for all}\quad u\in H.$$

Making use again of (4.14) and the definition of ξ, we can rewrite this as (4.11). Hence (c) holds.

(c)$\Rightarrow$(a). Again we have (4.11) for some $a>0$ and $\delta>0$, and this can be expressed as (4.12). Consider any z and w with $w\in\partial f(z)$, or equivalently $z\in\partial f^*(w)$. We have

$$f^*(w)+\langle z,u-w\rangle\leqq f^*(u)\quad\text{for all}\quad u\in H,$$

and hence by (4.12),

$$f^*(w)+\langle z,u-w\rangle\leqq f^*(0)+\langle\bar{z},u\rangle+\xi(\|u\|)\quad\text{for all}\quad u\in H. \tag{4.17}$$

At the same time, the relation $\bar{z}\in f^*(0)$ implies

$$f^*(w)\geqq f^*(0)+\langle\bar{z},w\rangle.$$

Combined with (4.17), this yields

$$\langle\bar{z},w\rangle+\langle z,u-w\rangle\leqq\langle\bar{z},u\rangle+\xi(\|u\|)\quad\text{for all}\quad u\in H$$

or

$$\sup_{u\in H}\{\langle z-\bar{z},u\rangle-\xi(\|u\|)\}\leqq\langle z-\bar{z},w\rangle\leqq\|z-\bar{z}\|\cdot\|w\|.$$

Therefore

$$\xi^*(\|z-\bar{z}\|)\leqq\|z-\bar{z}\|\cdot\|w\|\quad\text{whenever}\quad w\in T(z), \tag{4.18}$$

where ξ^* is given by (4.13) as before. But

$$\xi^*(r)\geqq\delta|r|-\tfrac{1}{2}a\delta^2\quad\text{for all}\quad r\in R,$$

since

$$\tfrac{1}{2}a^{-1}r^2+\tfrac{1}{2}a\delta^2\geqq r\delta\quad\text{for all}\quad r\in R,\quad\delta\in R.$$

Hence (4.18) entails

$$\delta\|z-\bar{z}\|-\tfrac{1}{2}a\delta^2\leqq\|z-\bar{z}\|\cdot\|w\|.$$

If $\|w\|\leqq\tfrac{1}{2}\delta$, the latter implies $\|z-\bar{z}\|\leqq a\delta$, so that

$$\xi^*(\|z-\bar{z}\|)=\tfrac{1}{2}a^{-1}\|z-\bar{z}\|^2,$$

and the inequality in (4.18) becomes

$$\tfrac{1}{2}a^{-1}\|z-\bar{z}\|^2 \leqq \|z-\bar{z}\| \cdot \|w\|.$$

Thus (4.18) gives us

$$\|z-\bar{z}\| \leqq 2a\|w\| \quad \text{whenever} \quad \|w\| \leqq \delta/2 \quad \text{and} \quad w \in T(z),$$

and (a) is verified.

Remark 4. The proof of Proposition 7 shows that the infimum $\bar{a}$ of the numbers $a \geqq 0$ such that the Lipschitz condition in Theorem 2 holds (for $T = \partial f$) satisfies $\frac{1}{2}b^{-1} \leqq \bar{a} \leqq b^{-1}$, where

$$b = \liminf_{z \to \bar{z}} \frac{f(z)-f(\bar{z})}{\|z-\bar{z}\|^2} = \left[\limsup_{u \to 0} \frac{f^*(u)-f^*(0)-\langle \bar{z}, u\rangle}{\|u\|^2}\right]^{-1}$$

($\bar{z}$ being the unique minimizing point for f; $0^{-1} = \infty$ and $\infty^{-1} = 0$).

PROPOSITION 8. *Suppose that H is finite-dimensional and f is polyhedral convex (i.e. the epigraph of f is a polyhedral convex set). If f attains its minimum at a unique point $\bar{z}$, then $0 \in \operatorname{int} \partial f(\bar{z})$, so that Theorem 3 is applicable to $T = \partial f$. However, even if f does not attain its minimum at a unique point but merely is bounded below, the proximal point algorithm with exact minimization of ϕ_k at each step (and with c_k bounded away from zero) will converge to some minimizer of f in a finite number of iterations.*

Proof. The conjugate f^* is also polyhedral [26, p. 173]. If $\bar{z}$ is the unique minimizer of f, it is the sole element of $\partial f^*(0)$. Then f^* is differentiable at 0 [26, p. 242], hence actually affine in an open neighborhood W of 0 by polyhedral convexity, implying $\bar{z} = \nabla f^*(w)$ for all $w \in W$. Thus $w \in \partial f(\bar{z})$ for all $w \in W$.

More generally, if f is merely a polyhedral convex function which is bounded below, we still have $f^*(0) = -\inf f$ finite and attained [26, p. 268]. By Theorem 1, the proximal point algorithm with c_k bounded away from zero generates from any starting point z^0 a sequence $\{z^k\}$ such that $Q_k(z^k) \to 0$. We must show that in the case of exact minimization ($\varepsilon_k = 0$ in (A')) finite convergence is still obtained.

There is no loss of generality in supposing for convenience in the rest of the proof that $\min f = 0$, so that $f^*(0) = 0$. Let

$$M = \partial f^*(0) = \{z \mid f(z) = \min f\}$$

and

$$h(z) = \begin{cases} 0 & \text{if} \quad z \in M, \\ +\infty & \text{if} \quad z \notin M. \end{cases}$$

Then M is a polyhedral convex set, so that h is a polyhedral convex function. The conjugate h^* is then polyhedral too, and we have

$$h^*(w) = f^{*\prime}(0; w) = \lim_{\lambda \downarrow 0} [f^*(\lambda w) - f^{*\prime}(0)]/\lambda$$

[26, p. 216], since the polyhedral property of f^* implies that of $f^*(0; \cdot)$. It is clear from the latter formula that h^* coincides with f^* in some open neighborhood of 0. Moreover $c_k^{-1} Q_k(z^k)$ lies in this neighborhood for all k sufficiently large, since

$Q_k(z^k) \to 0$ and c_k is bounded away from 0. Thus

$$\partial h^*(c_k^{-1}Q_k(z^k)) = \partial f^*(c_k^{-1}Q_k(z^k)) \quad \text{for all large} \quad k.$$

Since $\partial f^* = T^{-1}$ for $T = \partial f$, we can conclude from Proposition 1(a) that

$$c_k^{-1}Q_k(z^k) \in (\partial h^*)^{-1}(P_k(z^k)) = \partial h(P_k(z^k))$$

for all k sufficiently large. This tells us that ultimately the algorithm acts on $\{z^k\}$ just as if the multifunction $T = \partial f$ were replaced by $T = \partial h$, or equivalently if f were replaced by h. But in that event $P_k(z^k)$ is just the point of M nearest to z^k.

Thus, as soon as we reach the stage where $c_k^{-1}Q_k(z^k)$ lies in the neighborhood where f^* coincides with $h^* = f^{*\prime}(0; \cdot)$ we have $z^{k+1} = P_k(z^k) \in M$. Since M consists of the fixed points of the mappings P_k, the sequence $\{z^k\}$ is constant thereafter.

Remark 5. In the case of Proposition 8, quadratic programming algorithms can be employed, at least in principle, to calculate the exact minimum of ϕ_k at each iteration. Then the exact form of the proximal point algorithm is reasonable, and according to Theorem 3 it will yield the unique minimizer $\bar{z}$ of f in a finite number of iterations. We shall show elsewhere [31] that this result, when applied to the dual of a linear programming problem, yields a fact proved by Polyak and Tretyakov [20]: when the "method of multipliers" is used on a linear programming problem with exact minimization of the augmented Lagrangian at each iteration, one has convergence to an optimal solution in a finite number of iterations.

5. Application to calculating saddle points. Let $L(x, y)$ be a convex–concave function on the Hilbert space $H_1 \times H_2$ which is closed and proper in the sense of [24], [28], and let T_L be the maximal monotone operator corresponding to L, as defined in the introduction. Then

$$(0, 0) \in T_L(x, y) \Leftrightarrow (x, y) = \arg\operatorname{minimax} L.$$

The proximal point algorithm for $T = T_L$ is thus capable of computing saddle points of L, and some of the results in the preceding section have analogues for this case.

Let us say that a function $\Lambda : H_1 \times H_2 \to [-\infty, +\infty]$ is *strongly convex–concave (with modulus α)* if $\Lambda(x, y)$ is strongly convex in x and strongly concave in y, both with modulus α.

THEOREM 5. *Let $T = T_L$. Then one has $S_k = T_{\Lambda_k}$ in criteria* (A′) *and* (B′), *where Λ_k is the function defined by* (1.11), *and Λ_k is closed, proper and strongly convex–concave with modulus $1/c_k$. Furthermore, if $\{z^k = (x^k, y^k)\}$ is any sequence generated by the proximal point algorithm under the hypothesis of Theorem* 1 *with criterion* (A′), *then $(x^k, y^k) \to (x^\infty, y^\infty)$ weakly, where (x^∞, y^∞) is a saddle point of L and*

$$\lim_{k\to\infty} L(x^k, y^k) = L(x^\infty, y^\infty) = \operatorname{minimax} L. \tag{5.1}$$

Proof. This is mostly an easy extension of the argument for $T = \partial f$ in Theorem 4, but the justification of (5.1) is trickier and deserves some attention. Since

(x^∞, y^∞) is a saddle point, we have

$$L(x^{k+1}, x^\infty) \geqq L(x^\infty, y^\infty) \geqq L(x^\infty, y^{k+1}) \quad \text{for all} \quad k. \tag{5.2}$$

Let $w^k = (v^k, u^k)$ denote the element of $S_k(x^k, y^k)$ nearest the origin. Thus $(v^k, u^k) \to (0, 0)$ strongly and

$$(v^k - c_k^{-1}(x^{k+1} - x^k), u^k - c_k^{-1}(y^{k+1} - y^k)) \in T_L(x^{k+1}, y^{k+1}). \tag{5.3}$$

The latter relation gives us

$$L(x^\infty, y^{k+1}) \geqq L(x^{k+1}, y^{k+1}) + \langle x^\infty - x^{k+1}, v^k - c_k^{-1}(x^{k+1} - x^k)\rangle,$$
$$L(x^{k+1}, y^\infty) \leqq L(x^{k+1}, y^{k+1}) - \langle y^\infty - y^{k+1}, u^k - c_k^{-1}(y^{k+1} - y^k)\rangle.$$

Combining these inequalities with (5.2), we obtain

$$-\langle y^\infty - y^{k+1}, u^k - c_k^{-1}(y^{k+1} - y^k)\rangle \geqq L(x^\infty, y^\infty) - L(x^{k+1}, y^{k+1})$$
$$\geqq \langle x^\infty - x^{k+1}, v^k - c_k^{-1}(x^{k+1} - x^k)\rangle,$$

where the outer expressions converge to 0 by virtue of the limits already mentioned and assertion (2.5) of Theorem 1.

The analogue of Proposition 6 is valid for $T = T_L$, but the other results in § 4 do not have obvious extensions to the minimax context. For Proposition 7, this is seen from the example of $L(x, y) = xy$ on $R \times R$, which has $T_L(x, y) = (y, -x)$ and therefore T_L^{-1} globally Lipschitz continuous with modulus 1.

REFERENCES

[1] E. ASPLUND, AND R. T. ROCKAFELLAR, *Gradients of convex functions*, Trans. Amer. Math. Soc., 139 (1969), pp. 443–467.

[2] A. AUSLENDER, *Problèmes de Minimax via l'Analyse Convexe et les Inégalités Variationelles: Théorie et algorithmes*, Lecture Notes in Econ. and Math. Systems, 77, Springer-Verlag, 1972.

[3] A. B. BAKUSHINSKII AND B. T. POLYAK, *On the solution of variational inequalities*, to appear.

[4] D. P. BERTSEKAS, *Necessary and sufficient conditions for a penalty method to be exact*, Math. Programming, to appear.

[5] ———, *Multiplier methods: a survey*, Automatica—J. IFAC., March (1976).

[6] A. BRØNSTED AND R. T. ROCKAFELLAR, *On the subdifferentiability of convex functions*, Proc. Amer. Math. Soc., 16 (1965), pp. 605–611.

[7] F. E. BROWDER, *Multivalued monotone nonlinear mappings and duality mappings in Banach spaces*, Trans. Amer. Math. Soc., 118 (1965), pp. 338–351.

[8] M. R. HESTENES, *Multiplier and gradient methods*, J. Optimization Theory and Appl., 4 (1969), pp. 303–320.

[9] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, (1950); English transl., Macmillan, New York, 1964.

[10] M. A. KRASNOSELSKII, *Solution of equations involving adjoint operators by successive approximations*, Uspekhi Mat. Nauk, 15 (1960), no. 3 (93), pp. 161–165.

[11] A. V. KRYANEV, *The solution of incorrectly posed problems by methods of successive approximations*. Dokl. Akad. Nauk SSSR, 210 (1973), pp. 20–22 = Soviet Math. Dokl., 14 (1973), pp. 673–676.

[12] B. MARTINET, *Regularisation d'inéquations variationelles par approximations successives*, Rev. Francaise Inf. Rech. Oper., (1970), pp. 154–159.

[13] ———, *Determination approchée d'un point fixe d'une application pseudo-contractante*, C.R. Acad. Sci. Paris, 274 (1972), pp. 163–165.

[14] G. J. MINTY, *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J., 29 (1962), pp. 341–346.

[15] ———, *On the monotonicity of the gradient of a convex function*, Pacific J. Math., 14 (1964), pp. 243–247.

[16] J. J. MOREAU, *Fonctionelles sous-différentiables*, C.R. Acad. Sci. Paris, 257 (1963), pp. 4117–4119.

[17] ———, *Sur la fonction polaire d'une fonction sémi-continue supérieurement*, C.R. Acad. Sci. Paris, 258 (1964), pp. 1128–1131.

[18] ———, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

[19] ———, *Fonctionelles Convexes*, lecture notes, Séminaire "Equations aux derivées partielles", Collège de France, Paris, 1966–67.

[20] B. T. POLYAK AND N. V. TRETYAKOV, *On an iterative method of linear programming and its economic interpretation*, Ekon. Mat. Met., 8 (1972), pp. 740–751.

[21] M. J. D. POWELL, *A method for nonlinear optimization in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.

[22] R. T. ROCKAFELLAR, *Extension of Fenchel's duality theorem*, Duke Math. J., 33 (1966), pp. 81–89.

[23] ———, *Level sets and continuity of conjugate convex functions*, Trans. Amer. Math. Soc., 123 (1966), pp. 46–63.

[24] ———, *Monotone operators associated with saddle functions and minimax problems*, Nonlinear Functional Analysis, Part 1, F. E. Browder, ed., Symposia in Pure Math., vol. 18, Amer. Math. Soc., Providence, R.I., 1970, pp. 397–407.

[25] ———, *Local boundedness of nonlinear monotone operators*, Michigan Math. J., 16 (1969), pp. 397–407.

[26] ———, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[27] ———, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc. 149 (1970), pp. 75–88.

[28] ———, *Saddle functions and convex analysis*, Differential Games and Related Topics, H. W. Kuhn and G. P. Szego, eds., North-Holland, Amsterdam, 1971, pp. 109–128.

[29] ———, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optimization Theory and Appl., 12 (1973), pp. 555–562.

[30] ———, *Conjugate Duality and Optimization*, Regional Conference Series in Applied Mathematics No. 16, Society for Industrial and Applied Mathematics, Philadelphia, 1974.

[31] ———, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. of Operations Research, 1976, to appear.

[32] Z. OPIAL, *Weak convergence of the successive approximations for nonexpansive mappings in Banach spaces*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.

[33] A. GENEL AND L. LINDENSTRAUSS, *An example concerning fixed points*, Israel J. Math., 20 (1975).

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

EXISTENCE OF AN OPTIMAL CONTROL FOR SYSTEMS WITH JUMP MARKOV DISTURBANCES*

ROBERT M. GOOR†

Abstract. We consider the question of existence of an optimal control for stochastic problems with dynamics represented by a system of ordinary differential equations perturbed by a countable state, jump Markov disturbance $r(\cdot)$, and with performance criterion in the stochastic Mayer form. We consider a class of controls which are functions of the time t and the history r_t of r. Under the expected conditions of continuity, closure and convexity, we prove the existence of an optimal control which is nonanticipative in the above sense. We proceed via the "direct method" of proof, utilizing the topology of "convergence in distribution" and applying the McShane–Warfield implicit function theorem to select a nonanticipative control.

Introduction. We consider the optimal control of stochastic systems in the form

$$\text{(A)} \qquad \dot{x}(t) = f(t, r(t), x(t), u(t)),$$

where $r(t)$ is a countable state Markov process with stationary transition probabilities, and the control $u(t)$ is to be chosen from a suitable class $\mathcal{U}$ of nonanticipative functions. The performance criterion to be minimized is the conditional expectation $J[x, u] = E\{\phi(\tau, x) | x(t_0) = x_0, r(t_0) = r_0\}$, where τ is the smaller of $\bar{\tau}$, the first time $x(t)$ reaches a target set M, and T, a fixed terminal time, and ϕ is a continuous functional on the trajectories. Control systems of this form have been studied in [10], [11], [13], [15] and [16]. In this paper, our purpose is to prove the existence of an optimal control of a certain form. To this end, it proves convenient to construct an underlying probability space Ω for the process $r(t)$. We then consider nonanticipative controls defined on $[t_0, T] \times \Omega$, i.e., controls which are functions of the time t and the past r_t of the disturbance $r(\cdot)$. Our main results say that an optimal control in this class exists under the assumptions that: the dynamics in (A) are uniformly bounded, the appropriate constraint sets are closed and the associated orientor field is convex.

The technique of proof of our existence theorems involves the so-called "direct method" of construction of the optimal control. That is, we show that a minimizing sequence has a subsequence which converges in some sense. In this case, the topology of "convergence in distribution" is appropriate, and the functional $J(\cdot)$ is shown to be lower semicontinuous with respect to this topology. A major and somewhat surprising development is the application of the McShane–Warfield implicit function theorem to select a *non*anticipative control.

We spend the first part of the paper defining our terms and deriving fundamental properties, and the remainder is devoted to the proofs of our main results, Theorems 2.1 and 2.2. Theorem 2.1 states the existence of expected time-optimal control ($\phi(t, x) = t$). Theorem 2.2 shows the existence of an optimal control in case $\phi(t, x) = \psi(x)$, where ψ is a continuous functional on the space of

* Received by the editors February 13, 1975, and in revised form July 28, 1975.
† Department of Mathematics, University of Delaware, Newark, Delaware 19711.

continuous R^n-valued functions defined on the interval $[t_0, T]$. We conclude by showing that our results apply, in particular, to the situation in which f is linear in x and u.

1. Preliminaries. We assume that the Markov process $r(t)$, defined on a fixed interval $[t_0, T]$, has states in $Z^+ = \{0, 1, 2, \cdots\}$ and transition probabilities $P_{ij}(t) = \Pr\{r(t+s) = j | r(s) = i\}$ which satisfy the "standard" conditions. That is, the $P_{ij}(t)$ are assumed continuous for $t > 0$ and:

(a) $P_{ij}(t) \geqq 0$, $t > 0$;

(b) $\sum_j P_{ij}(t) = 1$, $t > 0$;

(c) $\sum_k P_{ik}(t) P_{kj}(h) = P_{ij}(t+h)$, $t, h > 0$;

(d) $\lim_{t \to 0+} P_{ij}(t) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$

It is well known, then, that the derivatives $P'_{ij}(t)$ exist for $t \geqq 0$. We define $\lambda_{ij} = P'_{ij}(0)$, $i \neq j$ and $\lambda_i = -P'_{ii}(0)$ to be the infinitesimal parameters of the process, so that $\lambda_{ij}, \lambda_i \geqq 0$. We assume further that the process $r(t)$ is conservative and that each state is stable. Thus $0 \leqq \lambda_i < +\infty$ for each i, and $\sum_{j \neq i} \lambda_{ij} = \lambda_i$. It is well known that, with probability 1, $r(t)$ undergoes only finitely many transitions in any finite time interval $[t_0, T]$. Furthermore, the waiting time in state i is exponentially distributed with parameter λ_i, and stability implies there are no "instantaneous" states. The expression λ_{ij}/λ_i $(\lambda_i \neq 0)$ gives the probability of a transition from state i to state j, given that a transition occurs (see [7], [8]). We will assume that $r(t_0) = r_0$ is fixed throughout this paper.

For $N = 0, 1, \cdots$, we define the set $A_N \subseteq [t_0, T]^N \times (Z^+)^{N+1}$ as follows:

$A_0 = \{r_o\}$;

$A_N = \{t_1, \cdots, t_N, r_0, r_1, \cdots, r_N) | t_0 < t_1 < \cdots < t_N \leqq T, r_i \in Z^+, i = 1, \cdots, N, r_{i+1} \neq r_i\}$

for $N \geqq 1$. We set

$$\Omega = \bigcup_{N=0}^{\infty} A_N.$$

We may further define a metric ρ on Ω as follows: write $\omega_i = (t_1^i, t_2^i, \cdots, t_{N_i}^i, r_0, r_1^i, \cdots, r_{N_i}^i)$, $i = 1, 2$. Without loss of generality, assume $N_1 \leqq N_2$ and define

$$\rho(\omega_1, \omega_2) = |N_2 - N_1| + \sum_{j=1}^{N_1} (|t_j^1 - t_j^2| + |r_j^1 - r_j^2|).$$

It is clear that, with respect to the metric ρ, the space Ω is separable. It is also easy to see that Ω is locally compact and, hence open in its completion. By ([9, p. 207]), it follows that there is a metric ρ', equivalent to ρ, such that Ω is complete relative to ρ'.

Furthermore, each point $\omega = (t_1, \cdots, t_N, r_0, \cdots, r_N)$ of Ω corresponds to a sample path $r(t)$, $t_0 \leqq t \leqq T$; specifically, $r(t) = r_i$ for $t_i \leqq t < t_{i+1}$, $i = 0, \cdots, N-1$,

and $r(t) = r_N$ for $t_N \leqq t \leqq T$. With probability one, each sample path is represented as a unique point in Ω (allowing no "instantaneous" states) and thus, a probability measure μ is induced on Ω. From our previous remarks on the distributions of waiting times and the conditional transition probabilities, it is clear that μ is defined on the Borel sets $\mathcal{B}$ of Ω (with respect to ρ). Thus $(\Omega, \mathcal{B}, \mu)$ is an explicit construction of the underlying probability space governing the process $r(\cdot)$. We will regard r as a function from $[t_0, T]\times\Omega \to Z^+$, defined so that

$$r(t, t_1, \cdots, t_N, r_0, \cdots, r_N) = \begin{cases} r_i, & t_i \leqq t < t_{i+1},\ 0 \leqq i \leqq N-1, \\ r_N, & t_N \leqq t \leqq T. \end{cases}$$

Because we will investigate a class of functions which are intended to be nonanticipative, in some sense, we define an operator P_t on Ω which "projects" any ω in Ω onto that portion which is observable in the interval $[t_0, t]$. We let

$$P_t\omega = P_t(t_1, \cdots, t_N, r_0, \cdots, r_N) = \begin{cases} (r_0) & \text{if } t < t_1, \\ \omega & \text{if } t \geqq t_N, \\ (t_1, \cdots, t_j, r_0, \cdots, r_j) & \text{if } t_j \leqq t < t_{j+1} \text{ for any } j,\ 1 \leqq j \leqq N-1. \end{cases}$$

We show that $P_t\omega$ is jointly measurable on $[t_0, T]\times\Omega$. We will construct a sequence of simple functions $\{P^{(N)}\}$ so that $P_t^{(N)}\omega$ converges pointwise to $P_t\omega$. For $N \geqq 1$, define

$$t_i^N = t_0 + i(T - t_0)/N,$$

$i = 0, 1, \cdots, N$. Let t lie in $[t_0, T]$ and let $\omega = (t_1, \cdots, t_k, r_0, \cdots, r_k)$ be an element of Ω. Suppose that $t_j \leqq t < t_{j+1}$. We will generate a sequence of j elements from the set $\{t_i^N\}_{i=0}^N$ which approximates $(t_1, \cdots, t_j)$ in a suitable sense and, barring repetitions in the generated sequence, we will use it to define $P_t^{(N)}\omega$. In case of repetitions, we arbitrarily define $P_t^{(N)}\omega = (r_0)$, so that, for all t, ω and N, $P_t^{(N)}\omega \in \Omega$. Define $l(i)$, $i = 0, 1, \cdots, j$, to be the unique integer in $[0, N]$ such that $t_{l(i)}^N \leqq t_i < t_{l(i)+1}^N$. We define

$$P_t^{(N)}\omega = \begin{cases} (r_0) & \text{if } l(m) = l(n) \text{ for some } m \neq n,\ m, n \leqq j, \\ (t_{l(1)}^N, \cdots, t_{l(j)}^N, r_0, r_1, \cdots, r_j) & \text{otherwise.} \end{cases}$$

The functions $P^{(N)}: [t_0, T]\times\Omega \to \Omega$ are clearly measurable, and it is an exercise to show that $\lim_{N\to\infty} P_t^{(N)}\omega = P_t\omega$ for all (t, ω) in $[t_0, T]\times\Omega$.

We now utilize the functions $\{P_t\}$, $t \in [t_0, T]$, to define nonanticipativity. We will give two definitions, which we will show to be equivalent.

(i) We will say that a function F on $[t_0, T]\times\Omega$ is nonanticipative if there exists a subset Ω_1 of Ω, $\mu(\Omega_1) = 1$, such that for all t in $[t_0, T]$ and ω in Ω_1,

$$F(t, \omega) = F(t, P_t\omega).$$

(ii) The function F is nonanticipative if there exists a subset Ω_2 of Ω, $\mu(\Omega_2)=1$, such that, if $\omega_1, \omega_2 \in \Omega_2$ and $P_t\omega_1 = P_t\omega_2$, then

$$F(s, \omega_1) = F(s, \omega_2)$$

for $t_0 \leqq s \leqq t$.

Let us assume (i). Define $\Omega_2 = \Omega_1$ and let ω_1, ω_2 be elements of Ω_2 such that $P_t\omega_1 = P_t\omega_2$. If $t_0 \leqq s \leqq t$, then $P_s\omega_1 = P_s\omega_2$ so that, for $t_0 \leqq s \leqq t$,

$$F(s, \omega_1) = F(s, P_s\omega_1) = F(s, P_s\omega_2) = F(s, \omega_2),$$

and (ii) holds.

Now assume that (ii) holds, and let ω be an element of Ω_2. We would like to argue that $P_t(\omega) = P_t(P_t\omega)$ for all t, so that $F(s, \omega) = F(s, P_t\omega)$ for $t_0 \leqq s \leqq t$, or in particular, that $F(t, \omega) = F(t, P_t\omega)$. However, to make the above inference, we would need to know that $P_t\omega \in \Omega_2$ for all $\omega \in \Omega_2$ and $t \in [t_0, T]$, and this need not hold. Therefore we will construct $\Omega_1 \subseteq \Omega_2$ such that $\mu(\Omega_1) = 1$ and $P_t(\Omega_1) \subseteq \Omega_2$ for all t in $[t_0, T]$.

We will need the following preliminary construction. Suppose B is a Borel subset of Ω. For $t \in [t_0, T]$ and $i \in Z^+$, define $B_t^i = \{\omega \in B | r(t, \omega) = i\}$ and let $S^i = \{\omega \in \Omega | r(s, \omega) = i \text{ for } t < s \leqq t\}$. Using the Markov property ([7, p. 346, Thm. 1]) and the property of conditional independence ([7, p. 137, Thm. 2]), it can be shown that

$$\mu(S_t^i \cap P_t^{-1}(P_t(B_t^i))) = \mu(P_t^{-1}(P_t(B_t^i))) \exp[-\lambda_i(T-t)].$$

But, $S_t^i \cap P_t^{-1}(P_t(B_t^i)) = P_t(B_t^i)$ by the definition of the operator P_t.

Hence substituting and rearranging, we get

$$\mu(P_t^{-1}(P_t(B_t^i))) = \mu(P_t(B_t^i)) \exp[\lambda_i(T-t)].$$

Since $B = \bigcup_{i=0}^{\infty} B_t^i$, and the union is disjoint,

$$\mu(P_t^{-1}(P_t(B))) = \sum_{i=0}^{\infty} \mu(P_t(B_t^i)) \exp[\lambda_i(T-t)].$$

Suppose in addition that $B \subseteq P_t(\Omega)$. Then $P_t(B) = B$ implies that

$$\mu(P_t^{-1}(B)) = \sum_{i=0}^{\infty} \mu(B_t^i) \exp[\lambda_i(T-t)].$$

In this case then, $\mu(P_t^{-1}(B)) = 0$ if and only if $\mu(B) = 0$.

In particular, $P_t(\Omega) \backslash \Omega_2 \subseteq P_t(\Omega)$ and $\mu(P_t(\Omega) \backslash \Omega_2) = 0$. Therefore

$$\mu(P_t^{-1}[P_t(\Omega) \backslash \Omega_2]) = 0$$

for each t in $[t_0, T]$.

Let $\Lambda = \{s_k\}$ be a countable dense subset of $[t_0, T]$ containing T, and let $\bar{\Omega} = \bigcup_k P_{s_k}^{-1}[P_{s_k}(\Omega) \backslash \Omega_2]$. Then $\mu(\bar{\Omega}) = 0$ and $\mu(\Omega_2 \backslash \bar{\Omega}) = 1$. Define $\Omega_1 = \Omega_2 \backslash \bar{\Omega}$. Let ω be an element of Ω_1 and let t lie in $[t_0, T]$. We will show that $P_t\omega \in \Omega_2$. Given ω and t, there is an element s_k of Λ such that $s_k \geqq t$ and such that $P_{s_k}\omega = P_t\omega$. If $P_{s_k}\omega$ is an element of $\Omega \backslash \Omega_2$, then $P_{s_k}\omega \in P_{s_k}(\Omega) \backslash \Omega_2$, or $\omega \in \bar{\Omega}$. This contradicts the choice of ω, and we have shown that $P_t(\Omega_1) \subseteq \Omega_2$.

We may now follow our earlier argument on the new set Ω_1: if $\omega \in \Omega_1$, then ω and $P_t\omega$ are elements of Ω_2. Since $P_t(\omega) = P_t(P_t\omega)$,

$$F(s, \omega) = F(s, P_t\omega)$$

for $t_0 \leqq s \leqq t$. In particular, $F(t, \omega) = F(t, P_t\omega)$ for all t in $[t_0, T]$, and (i) holds.

Thus (i) and (ii) are equivalent, and we will say that F is nonanticipative if either condition applies. It is clear that if F is nonanticipative, then there is a subset Ω_1 of Ω, $\mu(\Omega_1) = 1$, such that both (i) and (ii) hold on Ω_1. In this case, we will say that F is nonanticipative on Ω_1.

We wish to relate our definition of nonanticipativity to measurability with respect to a given (increasing) family of σ-fields. For a given t in $[t_0, T]$, let Ω_t be the set $\{\omega \in \Omega | \omega = (t_1, \cdots, t_N, r_0, \cdots, r_N), t_N \leqq t\}$. As a subset of Ω, the set Ω_t has a σ-field $\mathscr{C}_t$ of Borel sets induced by restriction:

$$\mathscr{C}_t = \{B \cap \Omega_t | B \in \mathscr{B}\},$$

for each t, $t_0 \leqq t \leqq T$. Clearly $P_t(B) \in \mathscr{C}_t$ for all $B \in \mathscr{B}$. We define the family of σ-fields $\{\mathscr{B}_t\}_{t \in [t_0, T]}$ of Ω by setting $\mathscr{B}_t = \{P_t^{-1}(B) | B \in \mathscr{C}_t\}$. It is clear that $t_1 \leqq t_2$ implies $\mathscr{B}_{t_1} \subseteq \mathscr{B}_{t_2}$. Furthermore, $\mathscr{B}_{t_0} = \{\phi, \Omega\}$ and $\mathscr{B}_T = \mathscr{B}$. Let $\bar{\mathscr{B}}_t$ be the completion of $\mathscr{B}_t$.

For any real numbers $a, b, a < b$, we denote by $C^n[a, b]$ the space of continuous R^n-valued functions defined on $[a, b]$, with metric $\bar{\rho}$ defined as follows: $\bar{\rho}(x, y) = \sup_{a \leqq t \leqq b} |x(t) - y(t)| = \|x - y\|_{\sup}$. Here, $|h|$ denotes the Euclidian norm of an element h of R^n. If $F : [t_0, T] \times \Omega \to R^n$ for some n, and $t \in [t_0, T]$, we denote by F_t the restriction of F to the interval $[t_0, t]$, *i.e.*, $F_t : [t_0, t] \times \Omega \to R^n$. If $F(\cdot, \omega)$ is continuous on $[t_0, T]$, with probability 1 (w.p. 1), we may regard F as a mapping from Ω into $C^n[t_0, T]$, and F_t as a mapping from Ω into $C^n[t_0, t]$. We will require the following result.

LEMMA 1.1. *Suppose that $F : [t_0, T] \times \Omega \to R^n$ for some n and that $F(t, \omega)$ is jointly measurable and continuous in t w.p. 1. Then F is nonanticipative if and only if the map $F_t : \Omega \to C^n[t_0, t]$ is measurable with respect to the σ-field $\mathscr{B}_t$ for each t in $[t_0, T]$.*

Proof. We assume that F is nonanticipative. Let x be an arbitrary element of $C^n[t_0, t]$, let $\varepsilon > 0$ be arbitrary and write $B_\varepsilon(x) = \{y \in C^n[t_0, t] | \bar{\rho}(x, y) \leqq \varepsilon\}$. To show that F_t is a $\mathscr{B}_t$-measurable map, it is sufficient to prove that $F_t^{-1}(B_\varepsilon(x)) \in \mathscr{B}_t$.

For $t_0 \leqq s \leqq t$, let $B_s = \{\omega | |F(s, \omega) - x(s)| \leqq \varepsilon\}$, so that $B_s \in \mathscr{B}$ for each $s \in [t_0, t]$, since F is jointly measurable. Let $t_1, t_2, \cdots$, be a countable dense subset of $[t_0, t]$ and let $B = \bigcap_{k=1}^{\infty} B_{t_k}$. Clearly, $\omega \in B$ if and only if $F_t(\omega) \in B_\varepsilon(x)$, or, in other words, $B = F_t^{-1}(B_\varepsilon(x))$. By construction, $B \in \mathscr{B}$, so that F_t is $\mathscr{B}$-measurable. In addition, under the assumption of nonanticipativity, $P_t\omega_1 = P_t\omega_2$ implies (w.p. 1) $\omega_1 \in B_s$ if and only if $\omega_2 \in B_s$ for each s in $[t_0, t]$. Thus $P_t\omega_1 = P_t\omega_2$ implies $\omega_1 \in B$ if and only if $\omega_2 \in B$. It follows that $P_t^{-1}(P_t(B)) = B$ and we conclude that $B \in \mathscr{B}_t$. Thus F_t is $\mathscr{B}_t$-measurable.

We now assume that F_t is $\mathscr{B}_t$-measurable for each t in $[t_0, T]$. Then for given t in $[t_0, T]$, there is a sequence of simple functions $\{y_k\}$, $y_k : \Omega \to C^n[t_0, t]$, such that $\lim_{k \to \infty} y_k(\omega) = F_t(\omega)$ for almost all ω and each y_k is $\mathscr{B}_t$-measurable. In other words,

$$\lim_{k \to \infty} \sup_{t_0 \leqq s \leqq t} |y_k(s, \omega) - F(s, \omega)| = 0$$

w.p. 1. Now, y_k is a simple function implies there are disjoint subsets $\Omega_1^k, \Omega_2^k, \cdots, \Omega_{N_k}^k$ of Ω such that: $\Omega_i^k \in \mathcal{B}_t$ for each i, k;

$$\bigcup_{i=1}^{N_k} \Omega_i^k = \Omega;$$

and $y_k(\omega) = y_k^i$ for some element y_k^i of $C^n[t_0, t]$, and for all $\omega \in \Omega_i^k$.

We may write $\Omega_i^k = P_t^{-1}(A_i^k)$, where $A_i^k \in \mathcal{C}_t$, so that $\bigcup_{i=1}^{N_k} A_i^k = \Omega_t$ and $A_i^k \cap A_j^k = \{\varnothing\}$ for $i \neq j$. Therefore if $\omega_1, \omega_2 \in \Omega$ and $P_t\omega_1 = P_t\omega_2$, then $P_t\omega_1 \in A_i^k$, for some i, if and only if $P_t\omega_2 \in A_i^k$. In this case, $\omega_1, \omega_2 \in \Omega_i^k$. Thus $P_t\omega_1 = P_t\omega_2$ implies $y_k(\omega_1) = y_k(\omega_2) = y_k^i$. In other words, $P_t\omega_1 = P_t\omega_2$ implies $y_k(s, \omega_1) = y_k(s, \omega_2)$ for $t_0 \leqq s \leqq t$. Let X_t be the subset of Ω on which $\lim_{k\to\infty} y_k(\omega) = F_t(\omega)$; then $\mu(X_t) = 1$. Then, $\omega_1, \omega_2 \in X_t$ and $P_t\omega_1 = P_t\omega_2$ together imply that $F(s, \omega_1) = F(s, \omega_2)$ for $t_0 \leqq s \leqq t$.

Let $\{s_k\}$ be a countable dense subset of $[t_0, T]$ (including t_0 and T), and let $\bar{\Omega} = \bigcap_{k=1}^{\infty} X_{s_k}$, so that $\mu(\bar{\Omega}) = 1$. Let $t \in (t_0, T]$ and let ω_1, ω_2 be elements of $\bar{\Omega}$ such that $P_t\omega_1 = P_t\omega_2$. Then there is an $\varepsilon > 0$ so that $P_{t+\varepsilon}\omega_1 = P_{t+\varepsilon}\omega_2$. Thus there is an $s_k \geqq t$ so that $P_{s_k}\omega_1 = P_{s_k}\omega_2$. We conclude that $F(s, \omega_1) = F(s, \omega_2)$ for $t_0 \leqq s \leqq t$, and F is nonanticipative.

We are now in a position to define the optimal control problem that is the central object of investigation of this paper. We assume as given: a subset A of R^n for some n and a fixed $x_0 \in A$; a subset U of R^m for some m; a continuous function $f : [t_0, T] \times Z^+ \times A \times U \to R^n$ and a functional $\phi : [t_0, T] \times C^n[t_0, T] \to R$.

We define V to be the class of nonanticipative measurable functions $u : [t_0, T] \times \Omega \to R^m$ such that $u(t, \omega) \in U$ for all (t, ω). If a function $x(t, \omega)$, with values in R^n, is measurable on $[t_0, T] \times \Omega$ and absolutely continuous in t for almost all ω, we will denote by $\dot{x}(t, \omega)$ the time derivative of $x(t, \omega)$, for fixed ω. (In view of our construction of Ω, we will normally express explicit ω-dependence, and we will omit the argument ω only when we wish to emphasize the t-dependence of a particular process.) We will say that a pair of functions $[x, u]$, defined and measurable on $[t_0, T] \times \Omega$, is admissible if:

(1) $$u \in V;$$

(2) $$x(t, \omega) \text{ is absolutely continuous in } t \quad \text{w.p. } 1;$$

(3) $$x(t, \omega) \in A \text{ for all } t \quad \text{w.p. } 1;$$

(4) $$x(t_0, \omega) = x_0 \quad \text{for all } \omega;$$

(5) $$\dot{x}(t, \omega) = f(t, r(t, \omega), x(t, \omega), u(t, \omega)) \quad \text{for a.a. } t, \text{ w.p. } 1.$$

We will let $\mathcal{U}$ be the subclass of V consisting of u such that $[x, u]$ is admissible for some x. We will assume throughout that $\mathcal{U}$ is nonempty.

We assume that a closed "target" set M in R^n is given. For $[x, u]$ admissible, we define $\Gamma(\omega) = \{t | t_0 \leqq t \leqq T, \ x(t, \omega) \in M\} \cup \{T\}$ and we define $\tau(\omega) = \min\{t \in \Gamma(\omega)\}$. We define a cost functional $J[x, u]$, then, as the following conditional expectation:

$$J[x, u] = E\{\phi(\tau, x) | x(t_0) = x_0, r(t_0) = r_0\}.$$

We wish to minimize $J(\cdot)$ over the class of admissible pairs of functions.

We summarize our main results, for both the minimum expected time problem and for the fixed terminal time problem. Assume that: A and M are closed; U is compact; f is uniformly bounded; $\phi(t, x) = t$; and the set-valued function $f(t, r, x, U)$ is closed and convex for each (t, r, x). Then there is an optimal control for $J(\cdot)$. If the functional $J(\cdot)$ is defined in terms of $\phi(t, x) = \psi(x)$ for some continuous functional ψ on $C^n[t_0, T]$ (and M is empty), then as above, an optimal control exists.

It is important that the response x to a nonanticipative control u in $\mathcal{U}$ be nonanticipative, as well. In addition to our intuitive feeling that this be so, essential use will be made of this relation between control and response in the proof of our existence theorems. The following lemma gives general conditions guaranteeing a positive result.

LEMMA 1.2. *Assume that for any measurable function $u = u(t)$, $t_0 \leqq t \leqq T$, with values in U, for any $i \in Z^+$, $t_1 \in [t_0, T]$ and $x_1 \in A$, the differential equation $\dot{y}(t) = f(t, i, y(t), u(t))$, $y(t_1) = x_1$, has a unique solution in some open neighborhood of t_1. Then for any admissible pair $[x, u]$, the trajectory x is a nonanticipative function.*

Proof. Let $[x, u]$ be an admissible pair and suppose that u is nonanticipative on Ω_1, $u(\Omega_1) = 1$. For any $\omega \in \Omega_1$, define $I(\omega) = \{t | \omega_1 \in \Omega_1$ and $P_t\omega = P_t\omega_1$ together imply $x(s, \omega) = x(s, \omega_1)$ for $t_0 \leqq s \leqq t\}$, and let $\alpha(\omega) = \sup\{t | [t_0, t] \subseteq I(\omega)\}$. Since $t_0 \in I(\omega)$ for all ω, $\alpha(\omega)$ is well-defined. Suppose $P_{\alpha(\omega)}\omega = P_{\alpha(\omega)}\omega_1$ for some $\omega_1 \in \Omega_1$. Then $P_t\omega = P_t\omega_1$ for all $t < \alpha(\omega)$, which implies $x(s, \omega) = x(s, \omega_1)$ for all s and t such that $t_0 \leqq s \leqq t < \alpha(\omega)$. By the continuity of the sample paths, it follows that $x(\alpha(\omega), \omega) = x(\alpha(\omega), \omega_1)$. Thus $\alpha(\omega) \in I(\omega)$. Suppose, if possible, that $\alpha(\omega) < T$. Then for each $k = 1, 2, \cdots$, there exists $\omega_k \in \Omega_1$ and $t_k = \min(T, \alpha(\omega) + (1/k))$, such that $P_{t_k}\omega = P_{t_k}\omega_k$ but $x(s, \omega) \neq x(s, \omega_k)$ for some $s \in (\alpha(\omega), t_k]$. We note that for fixed ω, $u(t, \omega)$ is a measurable function of t taking values in U.

Let $i = r(\alpha(\omega), \omega)$. Then there is a point $\bar{t} > \alpha(\omega)$ such that $r(s, \omega) = i$ for $\alpha(\omega) \leqq s < \bar{t}$. We may choose $\bar{t}$ close enough to $\alpha(\omega)$ so that

$$\dot{y}(t) = f(t, i, y(t), u(t, \omega)), \qquad y(\alpha(\omega)) = x(\alpha(\omega), \omega) \tag{6}$$

has the unique solution $y_1(t)$ on $[\alpha(\omega), \bar{t})$. Then $y_1(t) = x(t, \omega)$ for $t \in [\alpha(\omega), \bar{t})$. Choose k so that $t_k \leqq \bar{t}$, and let $y_2(t)$ be the unique solution to

$$\dot{y}(t) = f(t, i, y(t), u(t, \omega_k)), \qquad y(\alpha(\omega)) = x(\alpha(\omega), \omega) \tag{7}$$

for t in an open neighborhood of $\alpha(\omega)$. Then $y_2(t) = x(t, \omega_k)$, by uniqueness, since it follows by the definition of ω_k that $x(\alpha(\omega), \omega) = x(\alpha(\omega), \omega_k)$. But since u is nonanticipative, $u(t, \omega_k) = u(t, \omega)$ for $t_0 \leqq t \leqq t_k$, and, therefore, equations (6) and (7) are identical. Since a solution to (6) is defined on $[\alpha(\omega), \bar{t})$, and (7) is the same as (6) on $[\alpha(\omega), t_k] \subseteq [\alpha(\omega), \bar{t})$, both (6) and (7) must have the same solution on $[\alpha(\omega), t_k]$. That is, $x(t, \omega) = x(t, \omega_k)$ for $\alpha(\omega) \leqq t \leqq t_k$, which contradicts the definition of ω_k. It follows that $\alpha(\omega) = T$ for each ω, or, $I(\omega) = [t_0, T]$ for all $\omega \in \Omega_1$. From the definition of $I(\omega)$, we conclude that if $\omega_1, \omega_2 \in \Omega_1$ and $P_t\omega_1 = P_t\omega_2$, then $x(s, \omega_1) = x(s, \omega_2)$ for $t_0 \leqq s \leqq t$, i.e., x is nonanticipative.

COROLLARY 1.3. *Under the hypotheses of Lemma 1.2, $[x, u]$ an admissible pair implies $\dot{x}(t, \omega)$ is nonanticipative and jointly measurable.*

Proof. This is clearly a consequence of Lemma 1.2 and the relation $\dot{x}(t, \omega) = f(t, r(t, \omega), x(t, \omega), u(t, \omega))$ for almost all t w.p. 1.

We will assume throughout this paper that the conditions of Lemma 1.2 are met.

It is useful to know that the nonanticipativity of $\dot{x}$ follows from that of x, even without the differential equation.

LEMMA 1.4. *If $x(t, \omega)$ is a nonanticipative function which is absolutely continuous in t for each $\omega \in \Omega$, then $\dot{x}(t, \omega)$ is nonanticipative.*

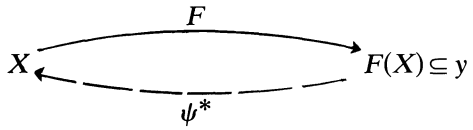
Proof. Suppose x is nonanticipative on Ω_1, $\mu(\Omega_1) = 1$. We note that if ω_1, $\omega_2 \in \Omega_1$, $t \in [t_0, T)$ and $P_t\omega_1 = P_t\omega_2$, then there is a point $\bar{t} \in (t, T)$ such that $P_{\bar{t}}\omega_1 = P_{\bar{t}}\omega_2$. In this case, for $|\Delta t| < \bar{t} - t$,

$$[x(s+\Delta t, \omega_1) - x(s, \omega_1)]/\Delta t = [x(s+\Delta t, \omega_2) - x(s, \omega_2)]/\Delta t$$

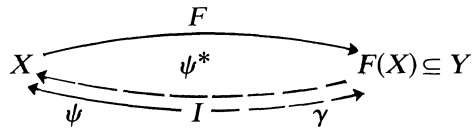
for $t_0 \leqq s \leqq t$. Thus for given s in $[t_0, t]$, the expression on the left has a limit as $\Delta t \to 0$ if and only if the expression on the right has a limit. It follows that $\dot{x}(s, \omega_1) = \dot{x}(s, \omega_2)$ for $t_0 \leqq s \leqq t$ (where $\dot{x}(s, \omega)$ may be defined to be 0 if the derivative does not exist in the usual sense).

We now return to the specifications of our control problem. As in proofs of deterministic existence theorems, we will make use of the orientor field formulation of the dynamics in (5). Let $Q(t, r, x) = f(t, r, x, U)$ for all (t, r, x) in $[t_0, T] \times Z^+ \times A$. Then $[x, u]$ is an admissible pair implies $\dot{x}(t, \omega) \in Q(t, r(t, \omega), x(t, \omega))$ for a.a. t w.p. 1. We would like to be able to solve the inverse problem: namely, if $\dot{x}(t, \omega)$, $x(t, \omega)$ are nonanticipative and measurable, and $\dot{x}(t, \omega) \in Q(t, r(t, \omega), x(t, \omega))$ for a.a. t w.p. 1, then there is a *nonanticipative* measurable control $u(t, \omega)$ so that $\dot{x}(t, \omega) = f(t, r(t, \omega), x(t, \omega), u(t, \omega))$. To this end, we include a short discussion of the McShane–Warfield implicit function theorem. (See [5] and [12].)

The McShane–Warfield theorem centers on one key result: let F be a continuous mapping from a metric space X (which is a countable union of compact subspaces) into a metric space Y, with range $F(X)$. Then, there is a



Borel measurable mapping $\psi^* : F(X) \to X$ so that $F(\psi^*(y)) = y$ for all $y \in F(X)$. This is the heart of the implicit function theorem. It follows easily from this result that if I is a measurable space, and $\gamma : I \to Y$ is a measurable map such that $\gamma(I) \subseteq F(X)$, then



there exists a measurable mapping $\psi : I \to X$ so that $F(\psi(t)) = \gamma(t)$ for almost all t in I. (In fact, define $\psi(t) = \psi^*(\gamma(t))$.) The usual application to deterministic control

theory designates the sets X, Y, I and maps F, γ as follows:

$$X=\{(t, x, u)|t\in[t_o, T], x\in A, u\in U\};$$
$$Y=\{(t, x, \dot{x})|t\in[t_0, T], x\in A, \dot{x}\in R^n\};$$
$$I=[t_0, T];$$
$$F(t, x, u)=(t, x, f(t, x, u)).$$

Finally, given a trajectory $x(t)$ satisfying $\dot{x}(t)\in f(t, x(t), U)$, γ is defined by $\gamma(t)=(t, x(t), \dot{x}(t))$. If f is continuous, so is F, and if $x(t)$ is absolutely continuous, all the hypotheses of the implicit function theorem are satisfied. Thus there is a measurable function $\psi(t)=(t, x(t), u(t))$ so that $F(\psi(t))=\gamma(t)$. Comparing components, we conclude that $\dot{x}(t)=f(t, x(t), u(t))$.

We propose to use the same theorem for our stochastic problem by redefining X, Y, I, F and γ (let us assume hereafter that A, U are closed):

$$X=\{(t, r, x, u)|t\in[t_0, T], r\in Z^+, x\in A, u\in U\};$$
$$Y=\{(t, r, x, \dot{x})|t\in[t_0, T], r\in Z^+, x\in A, \dot{x}\in R^n\};$$
$$F(t, r, x, u)=(t, r, x, f(t, r, x, u)).$$

Suppose x is a jointly measurable process on $[t_0, T]\times\Omega$, nonanticipative on a set Ω_1, with $\mu(\Omega_1)=1$, and having $x(\cdot, \omega)$ absolutely continuous for each ω, and suppose that $\dot{x}(\cdot, \cdot)$ is jointly measurable. By Lemma 1.4, $\dot{x}$ is nonanticipative. Suppose $\dot{x}(t, \omega)\in Q(t, r(t, \omega), x(t, \omega))$ for $(t, \omega)\in[t_0, T]\times\Omega_1$. Define $I=\{(t, P_t\omega)|t\in[t_0, T], \omega\in\Omega_1\}$ and define $\gamma: I\to Y$ by setting $\gamma(t, \omega)=(t, r(t, \omega), x(t, \omega), \dot{x}(t, \omega))$. It follows that: X is a σ-compact metric space; $F: X\to Y$ is continuous; I is a measurable space and γ is a measurable map so that $\gamma(I)\subseteq F(X)$. We conclude that there is a measurable function $\psi: I\to X$ such that $F\circ\psi=\gamma$. Again, comparing components, we have shown the existence of a measurable function $v(t, \omega)$ defined on I, taking values in U, such that $\dot{x}(t, \omega)=f(t, r(t,\omega), x(t, \omega), v(t, \omega))$ for $(t, \omega)\in I$. The function v may be extended to all of $[t_0, T]\times\Omega_1$ by defining $u(t, \omega)=v(t, P_t\omega)$ for all $(t, \omega)\in[t_0, T]\times\Omega_1$. It is clear that the function u is nonanticipative and that

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

on $[t_0, T]\times\Omega_1$. We summarize our findings in the following lemma.

LEMMA 1.5. *Let $x(t, \omega)$ be a jointly measurable, nonanticipative function which is absolutely continuous in t w.p.*1. *Suppose that $\dot{x}(t, \omega)$ is jointly measurable and that $\dot{x}(t, \omega)\in Q(t, r(t, \omega), x(t, \omega))$ for almost all t w.p.* 1. *Then there is a measurable, nonanticipative function $u(t, \omega)$, taking values in U, such that*

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

for almost all t w.p. 1.

We require a result concerning the existence of a measurable mapping from a given probability space into a given metric space, which gives rise to a given distribution. In particular, suppose that X is a complete, separable metric space with metric ρ' and that π is a probability measure defined on the Borel sets of X.

Let Y be a complete, separable metric space and let $\mathcal{P}$ be a probability measure on the Borel sets of Y.

LEMMA 1.6. *If π is nonatomic, there is a measurable function $h : X \to Y$ such that $\mathcal{P} = \pi h^{-1}$.*

Proof. It is well known (see [2, p. 29] or [14, p. 10]) that there is a measurable map $\alpha : [0, 1] \to Y$ such that $l\alpha^{-1} = \mathcal{P}$, where l is Lebesgue measure on the interval $[0, 1]$. We will construct a measurable map $\beta : X \to [0, 1]$ so that $l = \pi\beta^{-1}$. In this case, the function $h = \alpha \circ \beta$ gives the desired conclusion: $\mathcal{P} = \pi h^{-1}$.

Our construction of the map β parallels those in ([2, p. 29]) and ([14, p. 10]). For each $k = 1, 2, \cdots$, let $\mathcal{A}_k = \{A_{k,v}\}_{v=1}^{\infty}$ be a decomposition of X into disjoint π-continuity sets of diameter lesss than $1/k$, and let $\mathcal{I}_k = \{I_{k,v}\}_{v=1}^{\infty}$ be a decomposition of $[0, 1]$ into disjoint subintervals with lengths $l(I_{k,v}) = \pi(A_{k,v})$. We further arrange that $\mathcal{A}_{k+1}$ refines $\mathcal{A}_k$, $\mathcal{I}_{k+1}$ refines $\mathcal{I}_k$ and $A_{k+1,v} \subseteq A_{k,v'}$ implies $I_{k+1,v} \subseteq I_{k,v'}$.

We wish to show that $l(I_{k,v}) \to 0$ as $k \to \infty$, uniformly in v. Let $\varepsilon > 0$ be given. By ([2, p. 10, Thm. 1.4]), there is a compact subset K of X such that $\pi(K) > 1 - \varepsilon/2$. Suppose that for each $k = 1, 2, \cdots$, there is a v_k such that $\pi(A_{k,v_k}) \geqq \varepsilon$. Then $A_{k,v_k} \cap K$ is nonempty and $\pi(A_{k,v_k} \cap K) \geqq \varepsilon/2$. Let x_k be an element of $A_{k,v_k} \cap K$. Since K is compact, the sequence $\{x_k\}_{k=1}^{\infty}$ has a limit point $x_0 \in K$. Take $\delta > 0$ and write $N_\delta(x_0) = \{x \in X | \rho'(x, x_0) < \delta\}$. Then x_k lies in $N_\delta(x_0)$ for infinitely many values of k and, in particular, for some k satisfying $1/k < \delta$. For such a value of k, then, $A_{k,v_k} \subseteq N_\delta(x_0)$ (since diam $A_{k,v_k} < 1/k$). Therefore

$$\pi(N_\delta(x_0)) \geqq \pi(A_{k,v_k}) \geqq \varepsilon.$$

Since this relation holds for all $\delta > 0$, it follows that $\pi[\{x_0\}] \geqq \varepsilon$, which contradicts the assumption that π is nonatomic.

Therefore given $\varepsilon > 0$, there exists a number N so that $k \geqq N$ implies $\pi(A_{k,v}) < \varepsilon$ for all v, and hence $l(I_{k,v}) < \varepsilon$ for all v.

We define $\beta_k : X \to [0, 1]$ as follows. For all x in $A_{k,v}$, take $\beta_k(x)$ to be the mid-point of $I_{k,v}$. Clearly, each function β_k is Borel measurable. It is now easy to see that, for each x in X, $\{\beta_k(x)\}_{k=1}^{\infty}$ is a Cauchy sequence, for if $x \in X$, then there is a nested family $\{A_{k,v_k}\}_{k=1}^{\infty}$ such that $x \in A_{k,v_k}$. In this case, $\beta_k(x) \in I_{k,v_k}$, where $\{I_{k,v_k}\}_{k=1}^{\infty}$ is a nested family of intervals with diameters converging to 0. Write $\beta(x) = \lim_{k\to\infty} \beta_k(x)$. Then β is Borel measurable, and it is an exercise to show that $\pi\beta_k^{-1}$ converges weakly both to $\pi\beta^{-1}$ and to l. Hence $\pi\beta^{-1} = l$, and, by our previous remarks, Lemma 1.6 is proved.

In the next section, we will use the related concepts of weak convergence of a sequence of probability measures and convergence in distribution of a sequence of random elements. We refer to [2], and to pages 1–40 in particular, for definitions and theorems related to these concepts. If X is a probability space with measure μ, and y_k is a sequence of measurable maps from X into a metric space Y, it may happen that y_k converges in distribution to some random element y mapping X into Y. That is, if $\mathcal{P}_k = \mu y_k^{-1}$ and $\mathcal{P} = \mu y^{-1}$, then $\mathcal{P}_k$ converges weakly to $\mathcal{P}$. However, there are many random elements having the law $\mathcal{P}$. In particular, if $h : X \to X$ is a measurable, measure-preserving transformation, then $\bar{y}(x) = y(h(x))$ satisfies $\mathcal{P} = \mu\bar{y}^{-1}$ as well. Hence while there *are* certain properties of a sequence of random elements which are preserved by convergence in distribution,

these properties are invariant under measure-preserving transformations. For example, in general, measurability with respect to a given sub-σ-field of the Borel sets, or nonanticipativity, is *not* preserved by convergence in distribution. The following lemma establishes a result concerning convergence in distribution which is strong enough for our purposes.

Let X be a complete, separable metric space, as above, with metric ρ' and let Y be a separable Banach space with norm $\|\cdot\|$. Let μ be a nonatomic probability measure on the Borel sets $\mathcal{B}$ of X. Let i_X denote the identity of X onto itself.

LEMMA 1.7. *Let $y_k : X \to Y$, $k = 1, 2, \cdots$, be a sequence of measurable maps such that $\int_X \|y_k(x)\|\, d\mu(x) < +\infty$ for each k and such that the sequence of probability measures $\mathcal{P}_k = \mu(i_X, y_k)^{-1}$, defined on the Borel subsets of $X \times Y$, converges weakly to some probability measure $\mathcal{P}$. Then there exists a measurable map $y : X \to Y$ such that $\mathcal{P} = \mu(i_X, y)^{-1}$, i.e., (i_X, y_k) converges in distribution to (i_X, y).*

Proof. Write $p_1(E) = \mathcal{P}(E \times Y)$ and $p_2(F) = \mathcal{P}(X \times F)$, where E and F are Borel subsets of X and Y, respectively. Then $p_1(E) = \mu(E)$ for all $E \in \mathcal{B}$, by ([2, p. 20, Thm. 3.1]), since $\mathcal{P}_k(E \times Y) = \mu(E)$ for all k. We now parallel ([15, p. 10]) in constructing and utilizing a subdivision of Y into p_2-continuity sets.

We will use multi-index notation. Let $i^k = (i_1, \cdots, i_k)$ denote a k-tuple of positive integers. If $i^k = (i_1, \cdots, i_k)$ and $j^k = (j_1, \cdots, j_k)$, we will write $i^k < j^k$ if there exists an $r \leqq k$ such that $i_m = j_m$ for $m = 1, 2, \cdots, r-1$ and $i_r < j_r$. Denote the set of all k-tuples of positive integers by Λ^k. For each k, we let $\{A_{i^k}\}_{i^k \in \Lambda^k}$ be a subdivision of X into μ-continuity sets such that $\{A_{(i^k, i)}\}_{i=1}^{\infty}$ is a partition of A_{i^k} for each i^k, and the diameter of A_{i^k} is bounded by $(\frac{1}{2})^k$. Let $\{B_{j^k}\}_{j^k \in \Lambda^k}$ be a subdivision of Y into disjoint p_2-continuity sets having properties analogous to those of the $\{A_{i^k}\}$. (See [15, p. 10] for construction details.) Then the sets $S_{i^k, j^k} = A_{i^k} \times B_{j^k}$, $i^k, j^k \in \Lambda^k$, form a partition of $X \times Y$ into disjoint $\mathcal{P}$-continuity sets with diameters less than $(\frac{1}{2})^{k-1}$.

For each k, there is a sequence $y_k^m : X \to Y$ such that $y_k^m \to y_k$ in probability as $m \to \infty$, and y_k^m is constant on each set A_{i^m}. Indeed, by the assumed μ-integrability of $\|y_k(\cdot)\|$, we may define

$$y_k^m(x) = (1/\mu(A_{i^m})) \int_{A_{i^m}} y_k(\bar{x})\, d\mu(\bar{x})$$

for all $x \in A_{i^m}$, $\mu(A_{i^m}) \neq 0$, the integral defined in the sense of Bochner (see [18, p. 132]). Let $\varepsilon > 0$ be given and consider k fixed. We observe that $y_k(\cdot)$ is Bochner μ-integrable and it is an easy consequence that there is a uniformly continuous map $g_k : X \to Y$ such that $\int_X \|y_k(x) - g_k(x)\|\, d\mu(x) < \varepsilon$. Choose N so large that if x_1, x_2 satisfy $\rho'(x_1, x_2) < (\frac{1}{2})^N$, then $\|(g_k(x_1) - g_k(x_2)\| < \varepsilon$. For $m \geqq N$, then, $\|g_k(x_1) - g_k(x_2)\| < \varepsilon$ for $x_1, x_2 \in A_{i^m}$ and for any $i^m \in \Lambda^m$. Hence for $m \geqq N$ and for $x \in A_{i^m}$,

$$\begin{aligned}
\|y_k(x) - y_k^m(x)\| &\leqq \left(\frac{1}{\mu(A_{i^m})}\right) \int_{A_{im}} \|y_k(x) - y_k(\bar{x})\|\, d\mu(\bar{x}) \\
&\leqq \|y_k(x) - g_k(x)\| + \left(\frac{1}{\mu(A_{i^m})}\right) \int_{A_{im}} \|g_k(x) - g_k(\bar{x})\|\, d\mu(\bar{x}) \\
&\quad + \left(\frac{1}{\mu(A_{i^m})}\right) \int_{A_{im}} \|g_k(\bar{x}) - y_k(\bar{x})\|\, d\mu(\bar{x}),
\end{aligned}$$

the last inequality following from the triangle inequality. The middle term is bounded by ε, by the uniform continuity of g_k. Integrating both sides of the above inequality over X, and replacing the middle term on the right by ε, we get

$$\int_X \|y_k(x)-y_k^m(x)\|\,d\mu(x)\leqq 2\int_X \|y_k(x)-g_k(x)\|\,d\mu(x)+\varepsilon<3\varepsilon.$$

Hence $y_k^m\to y_k$ in $L^1(\mu)$ as $m\to\infty$. It follows that, for all $\varepsilon>0$, $\lim_{m\to\infty}\mu\{x|\|y_k^m(x)-y_k(x)\|\geqq\varepsilon\}=0$, i.e., $y_k^m\to y_k$ in probability as $m\to\infty$.

Let $\{y_k^{m_k}\}_{k=1}^{\infty}$ denote a subsequence such that $\|y_k^{m_k}(\cdot)-y_k(\cdot)\|\to 0$ in probability as $k\to\infty$. It follows that $(i_X, y_k^{m_k})$ converges in distribution to $\mathscr{P}$ as $k\to\infty$ (see [2, p. 25, Thm. 4.1]). In other words, if $\bar{\mathscr{P}}_k=\mu(i_x, y_k^{m_k})^{-1}$, then $\bar{\mathscr{P}}_k$ converges weakly to $\mathscr{P}$. Let $y^k(x)=y_k^{m_k}(x)$.

We now follow Skorokhod's method ([15, p. 10]) to construct mappings from the interval [0, 1] into $X\times Y$ which give rise to the laws $\bar{\mathscr{P}}_k, \mathscr{P}$. For $i^m, j^m\in\Lambda^m$, let $(\bar{x}_{i^m}, \bar{y}_{j^m})$ be points of $X\times Y$ such that $\bar{x}_{i^m}\in A_{i^m}$ for each i^m and $\bar{y}_{j^m}\in B_{j^m}$ for each j^m. As in the previous lemma, we denote Lebesgue measure on [0, 1] by l. Let $\{\Delta_{i^m,j^m}^k|i^m, j^m\in\Lambda^m\}$ be a collection of disjoint subintervals of [0, 1] of length

$$l(\Delta_{i^m,j^m}^k)=\mu\{x|(x, y^k(x))\in A_{i^m}\times B_{j^m}\},$$

and such that $\Delta_{i_1^m,j_1^m}^k$ lies to the left of $\Delta_{i_2^m,j_2^m}^k$ if either $i_1^m=i_2^m$ and $j_1^m<j_2^m$, *or*, $i_1^m<i_2^m$. For each k and m, then, the collection of intervals $\{\Delta_{i^m,j^m}^k|i^m, j^m\in\Lambda^m\}$ subdivides [0, 1]. Define the maps $z_k^m: [0,1]\to X\times Y$ by setting $z_k^m(s)=(\bar{x}_{i^m}, \bar{y}_{j^m})$ if $s\in\Delta_{i^m,j^m}^k$, $k=0, 1, 2, \cdots$. These are defined in analogy to the construction of Skorokhod. Hence it is shown in [15, p. 10] that for each fixed $k=0, 1, 2, \cdots$, z_k^m converges pointwise as $m\to\infty$ to a function $z_k(\cdot)$, and in turn, $\lim_{k\to\infty} z_k(s)=z_0(s)$, (l) a.e. Furthermore, $\bar{\mathscr{P}}_k=lz_k^{-1}$, $k=1, 2, \cdots$, and $\mathscr{P}=lz_0^{-1}$. We write $z_k^m=(h_k^m, \alpha_k^m)$, and $z_k=(h_k, \alpha_k)$, $k=0, 1, 2, \cdots$, where $h_k, h_k^m: [0,1]\to X$ and $\alpha_k, \alpha_k^m: [0,1]\to Y$.

Now for fixed $k>0$, by our specification, if $m\geqq m_k$, y^k is constant on each A_{i^m}, so that

$$l(\Delta_{i^m,j^m}^k)=\begin{cases}\mu(A_{i^m}) & \text{if } y^k: A_{i^m}\to B_{j^m},\\ 0 & \text{otherwise.}\end{cases}$$

We write $\sum_{i^m}^k=\bigcup_{j^m}\Delta_{i^m,j^m}^k$. Since $l(\Delta_{i^m,j^m}^k)\neq 0$ for exactly one value of j^m, we observe that if $m\geqq\max(m_{k_1}, m_{k_2})$, then the intervals $\sum_{i^m}^{k_1}$ and $\sum_{i^m}^{k_2}$ differ by at most a set of l-measure zero. Furthermore, $h_{k_1}^m(s)=h_{k_2}^m(s)=\bar{x}_{i^m}$ a.e. (l), for $s\in\sum_{i^m}^{k_1}\cap\sum_{i^m}^{k_2}$. It follows that $h_{k_1}=h_{k_2}$ a.e. (l). Let $h=h(s)$ denote the common limit: $h=h_k$, $k=0, 1, 2, \cdots$. Then $z_k=(h, \alpha_k)$, $k=0, 1, 2, \cdots$, where $\lim_{k\to\infty}\alpha_k(s)=\alpha_0(s)$ a.e. (l).

We now show that α_0 is measurable with respect to $h^{-1}(\mathscr{B})$, the σ-field on [0, 1] generated by h. Let Δ denote the completed σ-field generated by the sets Δ_{i^m,j^m}^k for $m\geqq m_k$, $k=1, 2, \cdots$. Then in fact, Δ is generated by the sets $\sum_{i^m}^k$, $m\geqq m_k$, $k=1, 2, \cdots$. We claim that $\Delta=h^{-1}(\mathscr{B})$. We observe that $\mathscr{P}=lz_0^{-1}$ implies $\mu=lh^{-1}$. Therefore if E is a subset of X such that $\mu(E)=0$, then $l(h^{-1}(E))=0$. Let $k>0$ be fixed and let $E\in\mathscr{B}$. Then for $m\geqq m_k$, $(h_k^m)^{-1}(E)=\{s|h_k^m(s)\in E\}=\bigcup\Delta_{i^m,j^m}^k$, the union taken over the countable set $\{i^m|\bar{x}_{i^m}\in E\}$. Thus h_k^m is Δ-measurable for each $m\geqq m_k$, and it follows that h, the pointwise limit, is

also Δ-measurable. On the other hand, we observe that $h(t) = \lim_{p\to\infty} h_k^p(t)$ for any fixed k. Let m be given, $m \geqq m_k$, for some fixed k. Then h_k^m maps $\sum_{i^m}^k$ into A_{i^m}. Fix $i^m = (i_1, \cdots, i_m)$. If $p \geqq m$ and if $i^p = (i_1, \cdots, i_m, i_{m+1}, \cdots, i_p) = (i^m, i_{m+1}, \cdots, i_p)$, then $\sum_{i^p}^k \subseteq \sum_{i^m}^k$ and it follows that h_k^p maps $\sum_{i^m}^k$ into A_{i^m}. Denote the interior of A_{i^m} by $\operatorname{int} A_{i^m}$. If $s \in h^{-1}(\operatorname{int} A_{i^m})$, then $s \in (h_k^p)^{-1}(\operatorname{int} A_{i^m})$ for p large enough, so that $s \in \sum_{i^m}^k$. Therefore $h^{-1}(\operatorname{int} A_{i^m}) \subseteq \sum_{i^m}^k$. Furthermore, it is clear that if $s \in \sum_{i^m}^k$, then $h_k^p(s) \in A_{i^m}$ for each $p \geqq m$, and $h(s) \in \operatorname{int} A_{i^m} \cup \partial A_{i^m} = \operatorname{cl} A_{i^m}$. Thus,

$$h^{-1}(\operatorname{int} A_{i^m}) \subseteq \sum_{i^m}^k \subseteq h^{-1}(\operatorname{cl} A_{i^m}).$$

But, by construction, $\operatorname{int} A_{i^m}$ and $\operatorname{cl} A_{i^m}$ differ by a set of μ-measure zero, and hence $h^{-1}(\operatorname{int} A_{i^m})$ and $h^{-1}(\operatorname{cl} A_{i^m})$ differ by a set of l-measure zero. It follows that $h^{-1}(A_{i^m}) = \sum_{i^m}^k$ up to a set of l-measure zero, and we conclude that $\Delta = h^{-1}(\mathcal{B})$.

Furthermore, for $m \geqq m_k$ and F a Borel subset of Y, $(\alpha_k^m)^{-1}(F) = \cup\, \Delta_{i^m, j^m}^k$, the union taken over the countable set $\{(i^m, j^m) | \bar{y}_{j^m} \in F\}$, so that α_k^m is Δ-measurable. It follows that α_k is Δ-measurable for $k = 1, 2, \cdots$, and that α_0, the pointwise limit, is Δ-measurable.

By Lemma 1.6, there is a measurable map $\beta : X \to [0, 1]$ that satisfies $l = \mu\beta^{-1}$. Define $\bar{h} : X \to X$ by $\bar{h}(x) = h(\beta(x))$ and define $\bar{\alpha} : X \to X$ by $\bar{\alpha}(x) = \alpha_0(\beta(x))$. If $\bar{\Delta}$ is the σ-field $\beta^{-1}(\Delta)$, then $\bar{h}$ and $\bar{\alpha}$ are both $\bar{\Delta}$-measurable, and $\bar{\Delta} = \bar{h}^{-1}(\mathcal{B})$ is the σ-field generated by $\bar{h}$. By the well-known theorem on functional dependence ([7, p. 603, Thm. 1.5]), there is a measurable mapping $y : X \to Y$ such that $y(\bar{h}(x)) = \bar{\alpha}(x)$. We note that $\mu\bar{h}^{-1} = \mu\beta^{-1}h^{-1} = lh^{-1} = \mu$, so that $\bar{h}$ is a measure-preserving transformation of X into itself. It is now an exercise to show that $\mu(i_X, y)^{-1} = \mathcal{P}$, and Lemma 1.7 is proved.

We have observed that our underlying probability space Ω is a complete, separable metric space and, in the next section, we would like to use the result in Lemma 1.7 with $X = \Omega$ and $Y = C^n[t_o, T]$. However, μ fails to satisfy the conditions of Lemma 1.7, since r_0 is an atom, having probability $\mu\{r_0\} = \exp[-\lambda_{r_0}(T - t_0)]$. We will therefore require the following corollary to Lemma 1.7.

COROLLARY 1.8. *Let X and μ be as in Lemma* 1.7, *except that we assume that X has an isolated point x_0 which is an atom for μ. Then Lemma* 1.7 *still holds.*

Proof. If $X = \{x_0\}$, then $\mu\{x_0\} = 1$ and it follows that $\{y_k(x_0)\}_{k=1}^\infty$ must be a Cauchy sequence in Y; i.e., there is a $y = y(x_0) \in Y$ such that $\|y_k(x_0) - y(x_0)\| \to 0$ as $k \to \infty$. Hence Corollary 1.8 holds trivially in this case.

We assume, then, that $X^0 = X \backslash \{x_0\}$ is nonempty and has positive μ-measure. Let $\eta = \mu(\{x_0\})$. By our assumption, X^0 is a complete, separable metric space with Borel sets $\mathcal{B}^0$. Define μ^0 on $\mathcal{B}^0$ by setting $\mu^0(E) = \mu(E)/(1 - \eta)$ for $E \in \mathcal{B}^0$. Then $(X^0, \mathcal{B}^0, \mu^0)$ is a nonatomic probability space.

We have assumed that if $\mathcal{P}_k = \mu(i_X, y_k)^{-1}$, then $\mathcal{P}_k \to \mathcal{P}$ weakly as $k \to \infty$, for some probability measure $\mathcal{P}$ on the Borel subsets of $X \times Y$. By [2, p. 37, Thm. 6.2], the sequence $\{\mathcal{P}_k\}_{k=1}^\infty$ is tight. Hence for any $\varepsilon > 0$, there is a compact subset E_ε of $X \times Y$ such that $\mathcal{P}_k(E_\varepsilon) > 1 - \varepsilon$ for all k. Choose $\varepsilon = \eta/2$. Then $\mathcal{P}_k(E_{\eta/2}) > 1 - (\eta/2)$ implies that $(x_0, y_k(x_0)) \in E_{\eta/2}$ for all k. Hence the sequence $\{y_k(x_0)\}_{k=1}^\infty$

has a limit point y_0. It follows that the point (x_0, y_0) is an atom of $\mathcal{P}$. Indeed, if $\delta > 0$ and $N_\delta(x_0, y_0) = \{(x_0, y) | \|y - y_0\| \leqq \delta\}$, then by [2, p. 11, Thm. 2.1],

$$\mathcal{P}(N_\delta(x_0, y_0)) \geqq \limsup_k \mathcal{P}_k(N_\delta(x_0, y_0)).$$

But $\mathcal{P}_k(N_\delta(x_0, y_0)) \geqq \eta$ for infinitely many values of k, and it follows that $\mathcal{P}(N_\delta(x_0, y_0)) \geqq \eta$ for all $\delta > 0$. Hence $\mathcal{P}(\{(x_0, y_0)\}) \geqq \eta$.

Define $\mathcal{P}_k^0 = \mu^0(i_{X^0}, y_k)^{-1}$ on the Borel subsets of $X \times Y$. That is, if E is a Borel subset of $X^0 \times Y$, then

$$\begin{aligned}\mathcal{P}_k^0(E) &= \mu^0\{x | (x, y_k(x)) \in E\} \\ &= [1/(1-\eta)]\mu\{x | (x, y_k(x)) \in E\} \\ &= [1/(1-\eta)]\mathcal{P}_k(E).\end{aligned}$$

Define $\mathcal{P}^0(E) = [1/(1-\eta)]\mathcal{P}(E)$ for all Borel subsets E of $X^0 \times Y$. We observe that $\mathcal{P}_k(X^0 \times Y) = 1 - \eta$ for all k and that $X^0 \times Y$ is closed. Hence $\mathcal{P}(X^0 \times Y) \geqq 1 - \eta$ by [2, p. 11, Thm. 2.1]. But we have seen already that $\mathcal{P}(\{x_0\} \times Y) = \eta$, so we conclude that $\mathcal{P}(X^0 \times Y) = 1 - \eta$ and $\mathcal{P}(\{(x_0, y_0)\}) = \mathcal{P}(\{x_0\} \times Y) = \eta$. It follows that $\mathcal{P}_k^0$, $k = 1, 2, \cdots$, and $\mathcal{P}^0$ are probability measures, and it is clear that $\mathcal{P}_k^0 \to \mathcal{P}^0$ weakly as $k \to \infty$. By Lemma 1.6, there is a measurable mapping $y^0 : X^0 \to Y$ such that $\mathcal{P}^0 = \mu^0(i_{X^0}, y^0)^{-1}$. We define a mapping $y : X \to Y$ by setting $y(x) = y^0(x)$ if $x \in X^0$, and $y(x_0) = y_0$. It is an exercise to show that $\mu(i_X, y)^{-1} = \mathcal{P}$, and Corollary 1.8 is proved.

2. Main results. We are now ready to prove our existence theorems. Our first theorem deals with the minimum expected time problem.

THEOREM 2.1. *Assume that: A and M are closed; U is compact; $|f(t, r, x, u)| \leqq K$ for some constant K and for all (t, r, x, u) in $[t_0, T] \times Z^+ \times A \times U$; $\phi(t, x) = t$ for $(t, x) \in [t_0, T] \times C^n[t_0, T]$ and $Q(t, r, x)$ is closed and convex for each (t, r, x) in $[t_0, T] \times Z^+ \times A$. Then there is an admissible pair $[x, u]$ which minimizes $J(\cdot)$.*

Proof. Since the theorem is trivial if $x_0 \in M$, we assume that $x_0 \notin M$. We note that it is a consequence of our assumptions on f and Q that

$$Q(t, r, x) = \bigcap_{\delta > 0} \operatorname{cl\,co} \bigcup_{y \in N_\delta(x)} Q(t, r, y)$$

for each $(t, r, x) \in [t_0, T] \times Z^+ \times A$, where $N_\delta(x) = \{y | |x - y| < \delta\}$ and $\operatorname{cl\,co} W$ denotes the closure of the convex hull of the set W (see [4, p. 377]). This property of set-valued functions, called property (Q) (here, with respect to the x-variable only), was introduced by Cesari [4] for deterministic control problems with unbounded control spaces.

Let $j = \inf J[x, u]$, the infimum taken over admissible pairs. Let $\{[x_k, u_k]\}_{k=1}^\infty$ be a sequence of admissible pairs such that $J[x_k, u_k] \leqq j + (1/k)$. Let $\tau_k(\omega)$ be the first time t such that $x_k(t, \omega)$ belongs to M if there is such a time; otherwise, $\tau_k(\omega) = T$. Then $J[x_k, u_k] = E\{\tau_k(\omega) | x_k(t_0) = x_0, r(t_0) = r_0\}$. By the definition of admissible pair, however,

$$\dot{x}_k(t, \omega) = f(t, r(t, \omega), x_k(t, \omega), u_k(t, \omega))$$

for almost all t w.p. 1. Thus $|\dot{x}_k(t)| \leqq K$ a.e., w.p. 1, for all k. Since $x_k(t_0, \omega) = x_0$ for all k and all ω, it follows that the set $\{x_k(\omega) | k = 1, 2, \cdots, \omega \in \Omega\}$ is an equibounded, equicontinuous family. Thus there is a compact subset $\mathscr{A}$ of $C^n[t_0, T]$ such that $x_k(\omega) \in \mathscr{A}$ for all k and all ω. Let i_Ω be the identity map on Ω. Let $\mathscr{P}_k$ be the distribution for (i_Ω, x_k), that is, $\mathscr{P}_k(E) = \mu[(i_\Omega, x_k)^{-1}(E)]$ for all Borel sets E of $\Omega \times C^n[t_0, T]$. We observe that the sequence $\{\mathscr{P}_k\}$ is tight (see [2, p. 37]). Indeed, given $\varepsilon > 0$, let F be a compact subset of Ω such that $\mu(F) > 1 - \varepsilon$. Clearly, then $F \times \mathscr{A}$ is a compact subset of $\Omega \times C^n[t_0, T]$ and $\mathscr{P}_k(F \times \mathscr{A}) > 1 - \varepsilon$. By a theorem of Prohorov [2, p. 37, Thm. 6.1], the sequence $\{\mathscr{P}_k\}$ has a weakly convergent subsequence, which we again denote by $\{\mathscr{P}_k\}$. Assume that $\mathscr{P} = \lim_{k \to \infty} \mathscr{P}_k$. Since Ω is a complete separable metric space with the isolated atom r_0, and $C^n[t_0, T]$ is a separable Banach space, we may apply Corollary 1.8 with $X = \Omega$ and $Y = C^n[t_0, T]$. Thus there is a measurable map $x : \Omega \to C^n[t_0, T]$ such that $\mathscr{P} = \mu(i_\Omega, x)^{-1}$. That is, (i_Ω, x_k) converges in distribution to (i_Ω, x) as $k \to \infty$.

We wish to show that $x(t, \omega)$ satisfies the constraints of our control problem. To this end, we define

$$S = \{(\omega, y) \in \Omega \times C^n[t_0, T] | y \text{ is absolutely continuous;}$$
$$y(t_0) = x_0;\ y(t) \in A \text{ for all } t \text{ and } \dot{y}(t) \in Q(t, r(t, \omega), y(t)) \text{ a.e.}\}.$$

We will show that S is closed in $\Omega \times C^n[t_0, T]$. Let (ω_k, y_k) be a sequence in S such that $\omega_k \to \omega$ in Ω and $y_k \to y$ in $C^n[t_0, T]$. Then $\lim_{k \to \infty} \rho(\omega_k, \omega) = 0$. By the definition of ρ, we may assume that ω_k, $k = 1, 2, \cdots$, and ω all have the same number of jumps, say N. We write $\omega = (t_1, \cdots, t_N, r_0, r_1, \cdots, r_N)$, and $\omega_k = (t_1^k, \cdots, t_N^k, r_0, r_1, \cdots, r_N)$. It follows that $\lim_{k \to \infty} t_i^k = t_i$ for $1 \leqq i \leqq N$. Let t be in $[t_0, T]$, $t \neq t_i$, for any $i = 1, \cdots, N$. Then there is an i such that $t_{i-1} < t < t_i$. For k large enough, $k \geqq \bar{k}$ implies $t_{i-1}^k < t < t_i^k$, so that $r(t, \omega_k) = r_{i-1} = r(t, \omega)$. Therefore $r(t, \omega_k) = r(t, \omega)$ for k large enough, for almost all t.

Now, $(\omega_k, y_k) \in S$ implies $\dot{y}_k(t) \in Q(t, r(t, \omega_k), y_k(t))$. But $q \in Q(t, r, y)$ implies $|q| \leqq K$, so the functions y_k are equi-Lipschitzian. It follows from the uniform convergence of y_k to y that y is Lipschitzian, and hence, absolutely continuous. Also, by the uniform convergence, $y(t_0) = x_0$, and since A is assumed closed, $y(t) \in A$ for all t. We would like to conclude that (ω, y) satisfies the orientor field relation. This is the main conclusion in what is called a "closure theorem" in deterministic control theory.

We essentially duplicate an argument in [1] and [6] to show that $(\omega, y) \in S$. The sequence $\{\dot{y}_k\}_{k=1}^\infty$ is clearly uniformly bounded in $(L^\infty[t_0, T])^n$, and so, by the Alaoglu theorem, there exists a subsequence which converges in the weak* topology. We shall rename this subsequence, if necessary, and denote it by $\{\dot{y}_k\}_{k=1}^\infty$. It follows that this same subsequence converges weakly in $(L^1[t_0, T])^n$. Hence by the Mazur theorem [18, p. 120], there is a sequence of convex combinations of the $\{\dot{y}_k\}$ which converges *strongly* in $(L^1[t_0, T])^n$. We let $\{z_k\}_{k=1}^\infty$ be this last sequence, and we note that, without loss of generality, we may assume that, for each k, z_k is expressible as a convex combination of $\{\dot{y}_i\}_{i=k}^\infty$. Furthermore, it is clear that $\dot{y}_k \to \dot{y}$ weakly in $(L^1[t_0, T])^n$, and this implies that $z_k \to \dot{y}$ strongly in the same space. Hence a subsequence, say $\{z_k\}$ again, converges pointwise a.e.

Now, it follows from the assumption that $(\omega_k, y_k) \in S$ for each k, and the construction of z_k, that

$$z_k(t) \in \text{co} \bigcup_{i=k}^{\infty} Q(t, r(t, \omega_i), y_i(t))$$

a.e. Let t be a point such that $z_k(t) \to \dot{y}(t)$ and such that $r(t, \omega_i) = r(t, \omega)$ for i large enough. Then

$$\dot{y}(t) \in \text{cl co} \bigcup_{i=k}^{\infty} Q(t, r(t, \omega_i), y_i(t)),$$

and, since this relation holds for all k,

$$\dot{y}(t) \in \bigcap_{k=N}^{\infty} \text{cl co} \bigcup_{i=k}^{\infty} Q(t, r(t, \omega_i), y_i(t)),$$

for any positive integer N. In particular, for N large enough,

$$\dot{y}(t) \in \bigcap_{k=N}^{\infty} \text{cl co} \bigcup_{i=k}^{\infty} Q(t, r(t, \omega), y_i(t)).$$

Let $\delta_k = \sup_{i \geq k} |y_i(t) - y(t)|$, so that δ_k decreases to 0 as $k \to \infty$, and, for $\delta > 0$, define $Q(t, \delta) = \bigcup_{z \in N_\delta(y(t))} Q(t, r(t, \omega), z)$. Then $\bigcup_{i=k}^{\infty} Q(t, r(t, \omega), y_i(t)) \subseteq Q(t, \delta_k)$ and

$$\dot{y}(t) \in \bigcap_{k=N}^{\infty} \text{cl co } Q(t, \delta_k).$$

Now $0 < \alpha < \beta$ implies $Q(t, \alpha) \subseteq Q(t, \beta)$, so the sets $\{Q(t, \delta)\}_{\delta > 0}$ form a nested family. Thus $\bigcap_{k=N}^{\infty} \text{cl co } Q(t, \delta_k) = \bigcap_{\delta > 0} \text{cl co } Q(t, \delta)$, and by property (Q), the expression on the right equals $Q(t, r(t, \omega), y(t))$. It follows that

$$\dot{y}(t) \in Q(t, r(t, \omega), y(t)).$$

Since this relation holds for almost all t, we have shown that $(\omega, y) \in S$, and that S is closed.

We note that $\mathcal{P}_k(S) = 1$ for all k since $[x_k, u_k]$ is an admissible pair, and hence (i_Ω, x_k) satisfies the constraints given in the definition of S. By [2, p. 11, Thm. 2.1], $\mathcal{P}(S) \geqq \limsup_k \mathcal{P}_k(S) = 1$, so that $\mathcal{P}(S) = 1$. It follows that: x is absolutely continuous w.p. 1; $x(t) \in A$ for all t w.p. 1; and

$$\dot{x}(t, \omega) \in Q(t, r(t, \omega), x(t, \omega))$$

for almost all t w.p.1.

It remains to show that x is nonanticipative, that there is a control $u \in \mathcal{U}$ so that $[x, u]$ is admissible and that $J[x, u] = j$.

To show that x is nonanticipative, by Lemma 1.1, we need show only that x_t is $\bar{\mathcal{B}}_t$-measurable for each t in $[t_0, T]$. Let t be a fixed element of $[t_0, T]$. Define $\Omega_t = P_t(\Omega)$, so that Ω_t is a closed subset of Ω and is therefore a complete, separable metric space in its own right. Define the probability measure $\mu_t = \mu P_t^{-1}$ on the Borel sets of Ω_t. Define $M_t : \Omega \times C^n[t_0, T] \to \Omega_t \times C^n[t_0, t]$ by $M_t(\omega, y) = (P_t\omega, y_t)$. Let $D_t = \{\omega \in \Omega | \omega = (t_1, \cdots, t_N, r_0, \cdots, r_N)$ and $t_i = t$ for some $i\}$. Thus D_t is the

set of paths having a jump *at* time t. It is easily verified that $\mu(D_t) = 0$. If D is the set of discontinuities of M_t, then $D \subseteq D_t \times C^n[t_0, T]$. But $\mathcal{P}(D_t \times C^n[t_0, T]) = \mu(D_t) = 0$, so that $\mathcal{P}(D) = 0$. Hence by [2, p. 31, Cor. 1], $M_t(i_\Omega, x_k)$ converges in distribution to $M_t(i_\Omega, x)$. If E is a Borel subset of $\Omega_t \times C^n[t_0, t]$, define

$$\mathcal{P}^t(E) = \mu\{\omega \in \Omega | (P_t\omega, x_t(\omega)) \in E\},$$

$$\mathcal{P}_k^t(E) = \mu\{\omega \in \Omega | (P_t\omega, x_{k_t}(\omega)) \in E\},$$

so that, by the above, $\mathcal{P}_k^t$ converges weakly to $\mathcal{P}^t$.

We now regard x_{k_t} as a mapping from Ω_t into $C^n[t_0, t]$, and we define the sequence of probability measures $\tilde{\mathcal{P}}_k^t$ as follows:

$$\tilde{\mathcal{P}}_k^t(E) = \mu_t\{\omega \in \Omega_t | (\omega, x_{k_t}(\omega)) \in E\}$$

for all Borel subsets E of $\Omega_t \times C^n[t_0, t]$. Then

$$\begin{aligned}\tilde{\mathcal{P}}_k^t(E) &= \mu(P_t^{-1}\{\omega \in \Omega_t | (\omega, x_{k_t}(\omega)) \in E\}) \\ &= \mu\{\omega \in \Omega | (P_t\omega, x_{k_t}(P_t\omega)) \in E\}.\end{aligned}$$

But x_k is nonanticipative for each k, so that $x_{k_t}(P_t\omega) = x_{k_t}(\omega)$ w.p. 1. Hence $\tilde{\mathcal{P}}_k^t = \mathcal{P}_k^t$ for all k, and it follows that $\tilde{\mathcal{P}}_k^t$ converges weakly to $\mathcal{P}^t$. By Corollary 1.8 (with $X = \Omega_t$, $Y = C^n[t_0, t]$), there is a measurable map $\tilde{x} : \Omega_t \to C^n[t_0, t]$ satisfying $\mathcal{P}^t(E) = \mu_t\{\omega \in \Omega_t | (\omega, \tilde{x}(\omega)) \in E\} = \mu\{\omega \in \Omega | (P_t\omega, \tilde{x}(P_t\omega)) \in E\}$ for all Borel subsets E of $\Omega_t \times C^n[t_0, t]$. Hence, for all such E,

$$\mathcal{P}^t(E) = \mu\{\omega \in \Omega | (P_t\omega, x_t(\omega)) \in E\} = \mu\{\omega \in \Omega | (P_t\omega, \tilde{x}(P_t\omega)) \in E\}.$$

Let E_t be a Borel subset of $C^n[t_0, t]$. We wish to show that $x_t^{-1}(E_t) \in \bar{\mathcal{B}}_t$, and for this to hold, it must happen that $x_t^{-1}(E_t) = P_t^{-1}(F)$, up to a set of measure zero, for some Borel subset F of Ω_t. Define $F_t = \tilde{x}^{-1}(E_t) \subseteq \Omega_t$. We claim that $x_t^{-1}(E_t) = P_t^{-1}(F_t)$, up to a set of measure zero.

Define $I_t(f) = \int f(\omega, y)\, d\mathcal{P}^t(\omega, y)$, the integral taken over $\Omega_t \times C^n[t_0, t]$, for any real-valued, measurable map f on $\Omega_t \times C^n[t_0, t]$. By the change of variable formula, $I_t(f)$ may be expanded as either of two expressions:

(B) $$I_t(f) = \int_\Omega f(P_t\omega, x_t(\omega))\, d\mu(\omega),$$

(C) $$I_t(f) = \int_\Omega f(P_t\omega, \tilde{x}(P_t\omega))\, d\mu(\omega).$$

Define $f_1(\omega, y)$ to be the characteristic function of $F_t \times E_t$ and $f_2(\omega, y)$ to be the characteristic function of $\Omega_t \times E_t$. Then utilizing expression (C), it is clear that $I_t(f_1) = I_t(f_2) = \mu(P_t^{-1}(F_t))$. Utilizing expression (B), we see that

$$I_t(f_1) = \int_{P_t^{-1}(F_t)} \chi_{E_t}(x_t(\omega))\, d\mu(\omega) \leqq \mu(P_t^{-1}(F_t)).$$

However, by the alternate expansion of $I_t(f_1)$, we know that equality holds, and the characteristic function $\chi_{E_t}(x_t(\omega)) = 1$ w.p. 1 for $\omega \in P_t^{-1}(F_t)$. In other words, $x_t^{-1}(E_t) \supseteq P_t^{-1}(F_t)$ (up to a set of μ-measure zero). Finally, from (C), $I_t(f_2) = \mu(x_t^{-1}(E_t))$, so that $\mu(x_t^{-1}(E_t)) = \mu(P_t^{-1}(F_t))$, and we may conclude that $x_t^{-1}(E_t) =$

$P_t^{-1}(F_t)$ up to a set of μ-measure zero. It follows that x_t is $\bar{\mathcal{B}}_t$-measurable, and by Lemma 1.1, that x is nonanticipative.

By Lemma 1.4, $\dot{x}$ is nonanticipative. We have thus constructed a map $x : [t_0, T]\times\Omega\to R^n$ such that: x is absolutely continuous in t w.p. 1; $x(t)\in A$ for all t w.p. 1; $x(t_0, \omega) = x_0$ for all ω; $\dot{x}(t, \omega)\in Q(t, r(t, \omega), x(t, \omega))$ for a.a. t, w.p. 1 and x, $\dot{x}$ are nonanticipative. By Lemma 1.5, there is a nonanticipative function $u(t, \omega)$ such that $u(t, \omega)\in U$ for a.a. t, w.p. 1, and such that

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

for a.a. t, w.p. 1. Thus $[x, u]$ is an admissible pair.

We observe that if $\bar{\mathcal{P}}_k(B) = \mathcal{P}_k(\Omega\times B)$ and $\bar{\mathcal{P}}(B) = \mathcal{P}(\Omega\times B)$ for all Borel subsets B of $C^n[t_0, T]$, then $\bar{\mathcal{P}}_k$ converges weakly to $\bar{\mathcal{P}}$.

We now show that $J[x, u] = j$. For $t_0\leqq t < T$, let $\mathcal{A}_t = \{y\in C^n[t_0, T] | y(s)\in M$ for some $s\in[t_0, t]\}$. Define $\mathcal{A}_T = C^n[t_0, T]$. Then $\mathcal{A}_t$ is closed for each t. In fact, if $y_k\in\mathcal{A}_t$, $y_k\to y$ uniformly as $k\to\infty$, then there are points $s_k\leqq t$ such that $y_k(s_k)\in M$. Let s_0 be the first time such that $y(s_0)\in M$, or, $s_0 = T$ if $y(s)\notin M$ for $t_0\leqq s\leqq T$. Suppose, if possible, that $\{s_k\}$ has a subsequence (say $\{s_k\}$, again) such that $s_k\to\bar{s} < s_0$. Clearly then, $y_k(s_k)\to y(\bar{s})$, and, since M is closed, $y(\bar{s})\in M$. This contradicts the definition of s_0. Therefore $s_0\leqq\liminf_{k\to\infty} s_k\leqq t$, and $y\in\mathcal{A}$. By the Portmanteau theorem [2, p. 11, Thm. 2.1], $\bar{\mathcal{P}}(\mathcal{A}_t)\geqq\limsup_{k\to\infty}\bar{\mathcal{P}}_k(\mathcal{A}_t)$. We define $F(t) = \bar{\mathcal{P}}(\mathcal{A}_t)$, $F_k(t) = \bar{\mathcal{P}}_k(\mathcal{A}_t)$ for all k and $t_0\leqq t\leqq T$. Since $\{\mathcal{A}_t\}$ is an increasing family of sets, F and F_k are nondecreasing functions on $[t_0, T]$. In fact, F_k is the distribution function of the random variable $\tau_k(\omega)$. Thus, $J[x_k, u_k] = E\{\tau_k(\omega)\} = \int_{t_0}^{T} t\, dF_k(t)$. Let $\tau(\omega)$ be the first time t such that $x(t, \omega)$ belongs to M if there is such a time; otherwise, $\tau(\omega) = T$. Then $J[x, u] = E\{\tau(\omega)\} = \int_{t_0}^{T} t\, dF(t)$, since F is clearly the distribution function of $\tau(\omega)$. But we have seen that $F(t)\geqq\limsup_k F_k(t)$ for all t. Therefore $-F(t)\leqq\liminf_{k\to\infty} -F_k(t)$ and, since $0\leqq F(t), F_k(t)\leqq 1$ for all t,

$$-\int_{t_0}^{T} F(t)\, dt\leqq\liminf_{k\to\infty}\left(-\int_{t_0}^{T} F_k(t)\, dt\right).$$

But $F(T) = F_k(T) = 1$ and $F(t_0) = F_k(t_0) = 0$, so by a comparison of the terms of the integration-by-parts formula, we conclude that

$$\int_{t_0}^{T} t\, dF(t)\leqq\liminf_{k\to\infty}\int_{t_0}^{T} t\, dF_k(t).$$

Therefore $J[x, u]\leqq\liminf_{k\to\infty} J[x_k, u_k] = j$. Since $[x, u]$ is admissible, $J[x, u]\geqq j$, and equality follows. This concludes the proof of Theorem 2.1.

Our second existence theorem incorporates a different cost functional.

THEOREM 2.2. *Assume that A is closed; U is compact; $|f(t, r, x, u)|\leqq K$ for some constant K and for all (t, r, x, u) in $[t_0, T]\times Z^+\times A\times U$; $\phi(t, x) = \psi(x)$ for some continuous functional $\psi : C^n[t_0, T]\to R$ and for all (t, x) in $[t_0, t]\times C^n[t_0, T]$; $Q(t, r, x)$ is closed and convex for each (t, r, x) in $[t_0, T]\times Z^+\times A$. Then there is an admissible pair $[x, u]$ which minimizes $J(\cdot)$.*

Proof. The proof of Theorem 2.2 parallels that for Theorem 2.1: we may assume that $[x_k, u_k]$ are admissible pairs so that if $j = \inf J[x, u]$, then $j =$

$\lim_{k\to\infty} J[x_k, u_k]$. We may assume that the distributions $\mathcal{P}_k = \mu(i_\Omega, x_k)^{-1}$ converge weakly to a distribution $\mathcal{P}$, induced by (i_Ω, x) for some trajectory x. As in the proof of Theorem 2.1, it can be shown that there is a control $u \in \mathcal{U}$ such that $[x, u]$ is admissible. Now, $J[x_k, u_k] = E\{\psi(x_k)\}$. But, $(i_\Omega, x_k) \to (i_\Omega, x)$ in distribution implies $x_k \to x$ in distribution. Thus, $\psi(x_k) \to \psi(x)$ in distribution. (See [2, p. 29]). However, as in the proof of Theorem 2.1, there is a compact subset $\mathcal{A}$ of $C^n[t_0, T]$ such that $x_k, x \in \mathcal{A}$ for all k. Hence the random variables $\psi(x_k)$, $\psi(x)$ are uniformly bounded. It follows that $\sup_k E\{|\psi(x_k)|^2\} < +\infty$, so that the $\psi(x_k)$ are uniformly integrable (see [2, p. 32]). By [2, p. 32, Thm. 5.4], $E\{\psi(x_k)\} \to E\{\psi(x)\}$ as $k \to \infty$. We conclude that $J[x, u] = E\{\psi(x)\} = j$, so that $[x, u]$ minimizes $J(\cdot)$, and Theorem 2.2 is proved.

COROLLARY 2.3. *Let ψ in Theorem* 2.2 *be defined so that $\psi(x) = \gamma(x(T))$ where $\gamma : R^n \to R$ is continuous, and suppose that the remaining hypotheses of Theorem* 2.2 *hold. Then there is an admissible pair $[x, u]$ which minimizes $J(\cdot)$.*

Remark 1. Uniform boundedness of the function f may be replaced in Theorems 2.1 and 2.2 by any growth condition which entails that any minimizing sequence be equi-Lipschitzian. This last property was the one used to show that the sequence $\{x_k\}$ is tight, and that S is closed. For deterministic analogues, see [3], for example. Extensions to unbounded control spaces are also possible, under the same proviso as above, but then property (Q), with respect to x, of the sets $Q(t, r, x)$, must be assumed separately.

Remark 2. In case the above problem is actually deterministic, that is, $\Omega = \{r_0\}$, then the proofs reduce to well-known deterministic proofs. For in this case: x_k is an absolutely continuous function on $[t_0, T]$; $\mathcal{P}_k$ is unit point mass at the element (r_0, x_k) of $\Omega \times C^n[t_0, T]$; $\mathcal{P}_k$ converges weakly to $\mathcal{P}$ (induced by x) implies $\mathcal{P}$ is unit point mass at $(r_0, x) \in \Omega \times C^n[t_0, T]$ and this last implies that $x_k \to x$ uniformly in $C^n[t_0, T]$. (See [2, p. 12].) Then showing S is closed in $C^n[t_0, T]$ is tantamount to showing that $x \in S$, which is the content of a deterministic closure theorem.

We conclude by showing that Theorems 2.1 and 2.2 both apply to the linear regulator problem. Let $f(t, r, x, u) = C(t, r)x + D(t, r)u + v(t, r)$, *where* $C(t, r)$ is an $n \times n$ matrix, $D(t, r)$ is an $n \times m$ matrix and $v(t, r)$ is an n-vector for each (t, r). Assume also that the functions $C(\cdot)$, $D(\cdot)$ and $v(\cdot)$ are continuous. If C is a matrix, we will denote by $\|C\|_{\text{op}}$ the usual operator norm: $\|C\|_{\text{op}} = \sup_{|y|=1} |Cy|$. We then have the following corollary of Theorems 2.1 and 2.2.

COROLLARY 2.4. *Assume that A is closed; U is compact and convex; $(\|C(t, r)\|_{\text{op}} + \|D(t, r)\|_{\text{op}} + |v(t, r)|) \leqq K$ for some constant K and for all $(t, r) \in [t_0, T] \times Z^+$; M is closed; $\phi(\cdot)$ is as in Theorem* 2.1 or *in Theorem* 2.2. *Then there is an admissible pair which minimizes $J(\cdot)$.*

Proof. Since the operator norms of $C(\cdot)$ and $D(\cdot)$ and the Euclidean norm of $v(\cdot)$ are uniformly bounded, it is an exercise to show that there is a constant $\bar{K}$ such that if $[x, u]$ is admissible, then $|x(t) - x_0| \leqq \bar{K}$ for $t_0 \leqq t \leqq T$. Let $\bar{A} = A \cap \{x \in R^n \mid |x - x_0| \leqq \bar{K}\}$. Then $\bar{A}$ is closed, and f restricted to $[t_0, T] \times Z^+ \times \bar{A} \times U$ is uniformly bounded. It is no restriction on our problem, then, to replace A by $\bar{A}$. Since U is compact and convex, $Q(t, r, x)$ is closed and convex for each (t, r, x). Theorems 2.1 and 2.2 now apply directly, with the space constraint set $\bar{A}$ instead of A.

Acknowledgment. The author particularly wishes to thank Raymond Rishel for his generous expenditure of time and many valuable suggestions.

REFERENCES

[1] L. D. BERKOVITZ, *Existence theorems in problems of optimal control*, Studia Math., 45 (1972), pp. 275–285.
[2] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
[3] L. CESARI, *Existence theorems for optimal controls of the Mayer type*, this Journal, (1968), pp. 517–552.
[4] ———, *Existence theorems for weak and usual solutions in Lagrange problems with unilateral constraints*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412, 413–429.
[5] ———, *Problems of Optimization*, to appear.
[6] L. CESARI AND M. B. SURYANARAYANA, *Convexity and property (Q) in optimal control theory*, this Journal, (1974), pp. 705–720.
[7] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.
[8] I. I. GIKHMAN AND A. V. SKOROKHOD, *Introduction to the Theory of Random Processes*, W. B. Saunders, Philadelphia, 1969.
[9] S. KARLIN, *A First Course in Stochastic Processes*, Academic Press, New York, 1969.
[10] J. L. KELLEY, *General Topology*, Van Nostrand, Princeton, N.J., 1955.
[11] N. N. KRASSOVSKII AND E. A. LIDSKII, *Analytical design of controllers in stochastic systems with velocity limited controlling action*, J. Appl. Math. and Mech., 25 (1961), pp. 627–643.
[12] E. A. LIDSKII, *Optimal control of systems with random properties*, Ibid., 27 (1963), pp. 33–45.
[13] E. J. MCSHANE AND R. B. WARFIELD, *On Fillippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.
[14] R. RISHEL, *Dynamic programming and minimum principles for systems with jump Markov disturbances*, this Journal, 13 (1975), pp. 338–371.
[15] A. V. SKOROKHOD, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Mass., 1965.
[16] D. D. SWORDER, *Feedback control of a class of linear systems with jump parameters*, IEEE Trans. Automatic control, AC-14 (1969), pp. 9–14.
[17] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, A. T. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131–212.
[18] K. YOSIDA, *Functional Analysis*, 2nd ed., Springer-Verlag, Berlin, 1968.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

TIME OPTIMAL CONTROL OF INFINITE-DIMENSIONAL SYSTEMS*

G. KNOWLES†

Abstract. The problem of reaching a fixed, closed convex body, W, of a locally convex topological vector space X in minimum time, by a control system steered by a sequence of independently operating controls of the bounded amplitude type, is considered. Conditions are derived for the optimal control to exist, be unique and bang-bang, and in the case W is a ball in a Banach space, a necessary and sufficient condition for controllability is obtained. By means of examples, the application of these results to the control of distributed systems is shown, and extensions of results of Ergov [4] and Friedman [5] obtained.

1. Introduction. The problem considered here is that of steering a control system, of the form described in [8] and [9], to reach a fixed closed, convex body W of a locally convex topological vector space X, in minimum time.

Namely, suppose Ω is a set in R^n (possibly empty), and for every $t \in [0, t_0]$, some time interval, we are given σ-algebras $\mathcal{T}_t$ of subsets of $\Omega \times [0, t]$, and a sequence of vector measures $m_i(t) : \mathcal{T}_t \to X$, $i = 1, 2, \cdots$. We consider the control system whose output for a control $f = (f_i)_{i=1}^{\infty}$, ($\{f_i\}$ a sequence of uniformly bounded measurable functions) is the element of X given by $m(t, f) = \sum_{i=1}^{\infty} \int_{\Omega \times [0,t]} f_i \, dm_i(t)$. If we restrict the values of the controls so that $f(\omega, \tau) \in F(\omega, \tau)$, for some given set $F(\omega, \tau)$ contained in the countable product of the real line, $(\omega, \tau) \in \Omega \times [0, t_0]$, then we ask if there is a minimum time, t^*, for which $m(t^*, f) \in W$ for some such control f. The controls with this property are called optimal controls.

In § 3 we give conditions for the existence of optimal controls, and derive a necessary condition which they satisfy. This leads, in § 4, to the introduction of normal control systems, for which the optimal control is uniquely determined by the necessary condition, and as a consequence is bang-bang. These concepts are then applied in § 5 to control systems described by partial differential equations, and examples are given which extend results of [4] and [5] on bang-bang control for parabolic problems. Finally, in § 6, by combining the ideas of [2] and § 3, we derive necessary and sufficient conditions for approximate controllability and obtain a computationally more useful form of the necessary condition.

2. Definitions. In this section, we summarize some of the relevent theory of vector measures which will be needed in this note. For a more detailed and complete study see [8] or [9].

Suppose X is a quasi-complete locally convex topological vector space (l.c.t.v.s.), with continuous dual X'. For a subset $A \subset X$, denote by $\overline{\text{co}}\, A$ the closed, convex hull of A, and by ex A the set of extreme points of A. If a linear functional $x' \in X'$ achieves its maximum value on A at a point $x_0 \in A$, then x' is said to

* Received by the editors May 27, 1975, and in revised form November 8, 1975.

† Institut für Angewandte Mathematik und Informatik, Universität Bonn, 53 Bonn, Germany.
This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 72.

support A at x_0, and x_0 is called a support point of A. Similarly a point $x_0 \in A$ is called an exposed point of A if there exists an $x' \in X'$ such that $\langle x', x\rangle < \langle x', x_0\rangle$ whenever $x \in A$, $x \neq x_0$. The functional x' is said to expose A at x_0, and the set of exposed points of A is denoted by $\exp A$.

Let T be a set, and $\mathcal{T}$ a σ-algebra of subsets of T. If T is a Borel subset of R^n, we denote by $\mathcal{B}(T)$ the Borel σ-algebra on T. $\mathcal{BM}(\mathcal{T})$ stands for the space of bounded, $\mathcal{T}$-measurable functions on T, and for a set $V \subset R$,

$$\mathcal{BM}_V(\mathcal{T}) = \{f : f \in \mathcal{BM}(\mathcal{T}), f(t) \in V, t \in T\}.$$

By a vector measure m on T we mean a countably additive map $m : \mathcal{T} \to X$. For a set $E \in \mathcal{T}$, put $m(\mathcal{T}_E) = \{m(F) : F \in \mathcal{T}, F \subseteq E\}$; $m(\mathcal{T}) = m(\mathcal{T}_T)$. If $x' \in X'$, we define a measure $\langle x', m\rangle : \mathcal{T} \to R$ by $\langle x', m\rangle(E) = \langle x', m(E)\rangle$, $E \in \mathcal{T}$.

A real-valued $\mathcal{T}$-measurable function f on T is said to be m-integrable if it is integrable with respect to every measure $\langle x', m\rangle$, $x' \in X'$, and if for every $E \in \mathcal{T}$, there exists an element $x_E \in X$ such that

$$\langle x', x_E\rangle = \int_E f \, d\langle x', m\rangle, \qquad x' \in X'. \tag{1}$$

We denote

$$x_E = \int_E f \, dm, \qquad x_T = \int f \, dm, \qquad E \in \mathcal{T}.$$

For the properties of this integral, we refer to [9, §§II.2, II.3], and, in particular, Lemma II.3.1, where it is shown that every bounded $\mathcal{T}$-measurable function on T is m-integrable.

If f is an m-integrable function, then the mapping $n : \mathcal{T} \to X$ defined by $n(E) = \int_E f \, dm$, $E \in \mathcal{T}$, is called the indefinite integral of f with respect to m. By the Orlich–Pettis theorem, n is a vector measure.

A function $f \in \mathcal{BM}(\mathcal{T})$ is called m-null, if its indefinite integral is (identically) the zero measure. Two functions f, $g \in \mathcal{BM}(\mathcal{T})$ are called m-equivalent if $f - g$ is m-null, and the class of all functions in $\mathcal{BM}(\mathcal{T})$ m-equivalent to f is denoted by $[f]_m$. A set $E \in \mathcal{T}$ is called m-null if its characteristic function is m-null, and $[E]_m$ is defined similarly. Set $\mathcal{T}(m) = \{[E]_m : E \in \mathcal{T}\}$ and $L_{[0,1]}(m) = \{[f]_m : f \in \mathcal{BM}_{[0,1]}(\mathcal{T})\}$.

If m,n are two set-functions on $\mathcal{T}$ (real or vector-valued) we say $m \ll n$ if $[E]_n = 0$ implies $[E]_m = 0$, $E \in \mathcal{T}$. We call m,n equivalent if $m \ll n$ and $n \ll m$.

Let A be an index set directed by the relation $\leqq$. A net $\{[E_\alpha]_m\}_{\alpha \in A}$ in $\mathcal{T}(m)$ is said to be $\tau(m)$-convergent to $[E]_m$ ($\tau(m)$-Cauchy) if, for every neighborhood U of 0 in X, there exists an $\alpha_U \in A$ such that $m(\mathcal{T}_{E_\alpha \Delta E}) \subset U$, for every $\alpha \in A$ with $\alpha_U \leqq \alpha$ (such that $m(\mathcal{T}_{E_\alpha \Delta E_\beta}) \subset U$, for every $\alpha, \beta \in A$ with $\alpha_U \leqq \alpha$, $\alpha_U \leqq \beta$).

A vector measure $m : \mathcal{T} \to X$ is said to be closed if $\mathcal{T}(m)$ is $\tau(m)$-complete, in other words, if every $\tau(m)$-Cauchy net in $\mathcal{T}(m)$ is $\tau(m)$-convergent.

The properties of closed vector measures are described in detail in [9]. In particular, if X is metrizable, any measure $m : \mathcal{T} \to X$ is closed, the indefinite (Pettis) integral of a vector function with a scalar measure, is closed, and if m is closed, then $\overline{\text{co}}\, m(\mathcal{T}) = \{\int_T f \, dm : f \in \mathcal{BM}_{[0,1]}(\mathcal{T})\}$.

A sequence of closed vector measures $m_i : \mathcal{T} \to X$, $i = 1, 2, \cdots$, will be called a control system if $\sum_{i=1}^{\infty} x_i$ is convergent, for any $x_i \in m_i(\mathcal{T})$, $i = 1, 2, \cdots$. Since $0 \in m_i(\mathcal{T})$, $i = 1, 2, \cdots$, this convergence is unconditional. We write $m = (m_i)$.

Let R^∞ be the countable product of the real line treated as an l.c.t.v.s. under the product topology, and CCR^∞ the family of compact, convex, nonempty subsets of R^∞. If $I_i = [-1, 1]$, $i = 1, 2, \cdots$, set $I^\infty = \prod_{i=1}^{\infty} I_i$. A function $f = (f_i) : T \to R^\infty$ will be called $\mathcal{T}$-measurable, if each component $f_i : T \to R$, is $\mathcal{T}$-measurable, $i = 1, 2, \cdots$. We define $\mathcal{BM}(R^\infty, \mathcal{T})$ to be the set of all $\mathcal{T}$-measurable functions $f = (f_i) : T \to R^\infty$ which are uniformly bounded, i.e., $\sup\{\|f_i\|_\infty : i = 1, 2, \cdots\} < \infty$. It follows from [8, Lemma 3], that if $f \in \mathcal{BM}(R^\infty, \mathcal{T})$ and $m = (m_i)$ is a control system, then $\sum_{i=1}^{\infty} \int f_i \, dm_i \in X$.

A set-valued function F defined on T whose values are subsets of R^∞ will be called bounded if there exists a compact set $V \subset R^\infty$ such that $F(t) \subset V$, $t \in T$. We call a set-valued function $F : T \to CCR^\infty$ measurable, if, for every $x' \in (R^\infty)'$ the mapping $t \to \sup\{\langle x', x\rangle : x \in F(t)\}$, $t \in T$, is $\mathcal{T}$-measurable. For $F : T \to CCR^\infty$, the set-valued function ex F is defined by $(\text{ex } F)(t) = \text{ex } F(t)$, $t \in T$.

In the sequel, for a given bounded measurable set-valued function $F : T \to CCR^\infty$, we will consider the set

$$\mathcal{BM}_F(R^\infty, \mathcal{T}) = \{f : f \in \mathcal{BM}(R^\infty, \mathcal{T}) \text{ and } f(t) \in F(t),\ t \in T\}$$

as the class of admissible controls. The case $F(t) = I^\infty$, $t \in T$, corresponds to the case of bounded-amplitude controls. We reserve the term "bang-bang" control for those controls f with $f(t) \in \text{ex } F(t)$, $t \in T$, and so in the case $F(t) = I^\infty$, $t \in T$, the bang-bang controls are those for which every component $|f_i| \equiv 1$, $i = 1, 2, \cdots$.

Suppose $m = (m_i)$ is a control system. Two functions $f = (f_i)$ and $g = (g_i)$ from $\mathcal{BM}(R^\infty, \mathcal{T})$ are called m-equivalent, if f_i and g_i are m_i-equivalent for every $i = 1, 2, \cdots$. The class of functions in $\mathcal{BM}(R^\infty, \mathcal{T})$ m-equivalent to f is denoted by $[f]_m$, and for a bounded set-valued function F on T, put

$$L_F(R^\infty, m) = \{[f]_m : f \in \mathcal{BM}_F(R^\infty, \mathcal{T})\}.$$

When it is convenient and will not cause confusion, we write f in place of $[f]_m$ to simplify the notation. On this set we can define a locally convex Hausdorff topology, which we will here call $\sigma(m)$, such that the mapping $f \to \sum_{i=1}^{\infty} \int f_i \, dm_i$, $f \in L_F(R^\infty, m)$, is continuous from the $\sigma(m)$ topology on $L_F(R^\infty, m)$ into the weak topology on X. If $F : T \to CCR^\infty$ is bounded and measurable, $L_F(R^\infty, m)$ is also $\sigma(m)$-compact. For the details see [9, Theorem IX.1.1].

3. Necessary condition for optimality. We begin by describing the particular form of the control system to be considered in the rest of this note.

Suppose Ω is a set (possibly empty) and $\mathcal{T}$ a σ-algebra of subsets of Ω. For every $t \in [0, t_0]$, some time interval, define $\mathcal{T}_t = \mathcal{T} \times \mathcal{B}([0, t])$ to be the product σ-algebra on $\Omega \times [0, t]$. Suppose that for each $t \in [0, t_0]$, we are given a control system $m(t) = (m_i(t))$, $m_i(t) : \mathcal{T}_t \to X$, $i = 1, 2, \cdots$. Then for any $f = (f_i) \in \mathcal{BM}(R^\infty, \mathcal{T}_t)$ define

$$m(t, f) = \sum_{i=1}^{\infty} \int_0^t \int_\Omega f_i(\omega, \tau) \, d(m_i(t))(\omega, \tau), \qquad t \in [0, t_0]. \tag{2}$$

It follows from our earlier remarks that $m(t, f) \in X$, and so $m(t, f)$ can be regarded as the position reached by the control system in time t, when steered by the control f.

For the class of admissible controls, we take $\mathcal{BM}_F(R^\infty, \mathcal{T}_{t_0})$ where $F: \Omega \times [0, t_0] \to CCR^\infty$ is a given bounded, measurable set-valued function. This formally requires that admissible controls are functions on $\Omega \times [0, t_0]$; however, from (2) it is clear that at a time $t < t_0$, only the values of the controls on $\Omega \times [0, t]$ affect the output at time t. For this reason, we adopt the convention that when we are only interested in the behavior of the system up to some specified time $t < t_0$, the controls will be considered as functions on only $\Omega \times [0, t]$. Then the attainable set at time t, that is, the set of all points reachable in time t by using all admissible controls, is just

$$A(t) = \{m(t, f) : f \in \mathcal{BM}_F(R^\infty, \mathcal{T}_t)\}.$$

It follows from [8, Thm. 1] that this set is a convex, weakly compact subset of X for each $t \in [0, t_0]$.

We remark here that analogues of Theorems 1, 2 and 3 below can also be proven for systems whose output can be represented as

$$m(t, f) = \sum_{i=1}^{\infty} \int_\Omega f_i(\omega)\, d(m_i(t))(\omega)$$

in place of (2). The proofs require only notational changes. (Such a system occurs at the end of § 5, equations (18)–(21).)

If W is a fixed closed, convex subset of X, we consider the problem of reaching W in minimum time, under the following assumptions.

(A) For some time $t_1 \in [0, t_0]$, $A(t_1) \cap W \neq \varnothing$.

(B) For any $t^* \in (0, t_0]$, and any $x' \in X'$,

$$\sup\{|\langle x', m(t, f) - m(t^*, f)\rangle| : f \in \mathcal{BM}_F(R^\infty, \mathcal{T}_{t_0})\} \to 0 \quad \text{as } t \downarrow t^*.$$

(C) For any control f, the function $t \to m(t, f)$, $t \in [0, t_0]$, is continuous into the given topology on X.

The first assumption is that of controllability and will be discussed in § 6. The second guarantees that the attainable set moves (in some weak sense) continuously in time, and the third that, for fixed f, the trajectory traced out by the system is continuous in time. Then we have the following.

LEMMA 1. *If* (A) *and* (B) *are satisfied, then a time optimal control exists.*

Proof. By assumption (A) $\{t : A(t) \cap W \neq \varnothing\}$ is nonempty, and so set $t^* = \inf\{t : A(t) \cap W \neq \varnothing\}$. We must show $A(t^*) \cap W \neq \varnothing$. In the usual way, select a sequence of times, $t_k \downarrow t^*$, and controls f^k such that $m(t_k, f^k) \in A(t_k) \cap W$. The restriction of each f^k to $\Omega \times [0, t^*]$ belongs to $L_F(R^\infty, m(t^*))$ and as this set is $\sigma(m(t^*))$-compact ([9, Cor. VIII.3.1]), there must exist a subnet f^j of f^k and an $f^* \in L_F(R^\infty, m(t^*))$, such that $f^j \to f^*$ in the $\sigma(m(t^*))$ topology. Then $m(t^*, f^*) \in A(t^*)$, and $m(t^*, f^j) \to m(t^*, f^*)$ weakly in X. However, as

$$\begin{aligned}\langle x', m(t_j, f^j) - m(t^*, f^*)\rangle = {} & \langle x', m(t_j, f^j) - m(t^*, f^j)\rangle \\ & + \langle x', m(t^*, f^j) - m(t^*, f^*)\rangle,\end{aligned}$$

it follows by (B) that $m(t_j, f^j) \to m(t^*, f^*)$ weakly in X, and since $m(t_j, f^j) \in W$ for each j, and W is weakly closed (as it is convex), $m(t^*, f^*) \in W$. In other words, f^* is an optimal control.

LEMMA 2. *If a time optimal control exists, with $t^* > 0$ the minimum time, and if W has nonempty interior in X, and* (C) *holds, then $A(t^*) \cap \operatorname{int} W = \varnothing$, and there exists a nonzero $x' \in X'$ such that, for any optimal control f^*,*

$$\langle x', m(t^*, f)\rangle \leqq \langle x', m(t^*, f^*)\rangle \leqq \langle x', w\rangle \tag{3}$$

for any admissible control f and any $w \in W$.

Proof. Suppose $f \in \mathcal{BM}_F(R^\infty, \mathcal{T}_{t_0})$ and $m(t^*, f) \in \operatorname{int} W$. Then there must exist an $\varepsilon > 0$ such that, for some continuous seminorm, p, on X, $B_p = \{x \in X : p(x - m(t^*, f)) < \varepsilon\} \subset W$. Since the function $t \to m(t, f)$ is continuous for $t \in [0, t_0]$, and $t^* > 0$, there must exist a $t < t^*$ with $m(t, f) \in B_p$, which contradicts the minimality of t^*. Hence $A(t^*) \cap \operatorname{int} W = \varnothing$.

As $\operatorname{int} W \neq \varnothing$, it follows from [3, Cor. to Thm. 21.11], that there exists a nonzero $x' \in X'$, separating the sets $A(t^*)$ and W; that is, $\langle x', m(t^*, f)\rangle \leqq \langle x', w\rangle$ for all admissible controls f and $w \in W$. If f^* is an optimal control, then $m(t^*, f^*) \in A(t^*) \cap W$, and so (3) follows.

Combining Lemmas 1 and 2 we have the following necessary condition.

THEOREM 1. *If W has nonempty interior, and* (A), (B) *and* (C) *are satisfied, then a time optimal control exists in the minimum time t^*, and if $t^* > 0$, then there exists a nonzero $x' \in X'$ such that for any optimal control $f^* = (f_i^*)$,*

$$\begin{aligned} &\sum_{i=1}^{\infty} \int_0^{t^*} \int_\Omega f_i^* \, d\langle x', m_i(t^*)\rangle \\ &\qquad = \max\left\{ \sum_{i=1}^{\infty} \int_0^{t^*} \int_\Omega f_i \, d\langle x', m_i(t^*)\rangle : f \in \mathcal{BM}_F(R^\infty, \mathcal{T}_{t_0}) \right\}. \end{aligned} \tag{4}$$

In particular, if $F(\omega, \tau) = I^\infty$, $(\omega, \tau) \in \Omega \times (0, t_0)$, then $f \in L_{\operatorname{ex} F}(R^\infty, (\langle x', m_i(t^*)\rangle))$, that is, $|f_i| = 1$, $\langle x', m_i(t^*)\rangle$ a.e., $i = 1, 2, \cdots$.

Proof. The relation (4) is an easy consequence of Lemma 2 and (1).

From (4) $\sum_{i=1}^{\infty} \int_0^{t^*} \int_\Omega f_i^* \, d\langle x', m_i(t^*)\rangle \in \operatorname{ex}\{\sum_{i=1}^{\infty} \int_0^{t^*} \int_\Omega f_i \, d\langle x', m_i(t^*)\rangle : f \in \mathcal{BM}_F(R^\infty, \mathcal{T}_{t_0})\}$, and so in the case $F \equiv I^\infty$, we have, by [14, Thm. 4], $f \in L_{\operatorname{ex} F}(R^\infty, (\langle x', m_i(t^*)\rangle))$.

4. Normal systems. In this section, we consider the consequences of the necessary condition (4) on the uniqueness and bang-bangness of the optimal control, when the set-valued function $F = I^\infty$, that is, the set of admissible controls is $\{f : f = (f_i) \in \mathcal{BM}_F(R^\infty, \mathcal{T}_{t_0})$ and $|f_i(\omega, \tau)| \leqq 1$ all $\omega, \tau\}$.

Firstly, as with any necessary condition of the Pontryagin type, (4) may give no information about the optimal control and certainly need not uniquely determine it. In fact, clearly, (4) gives no information about the values of f_i^* on any set $E_i \subset \Omega \times [0, t^*]$ which is $\langle x', m_i(t^*)\rangle$-null. However, if for every $i = 1, 2, \cdots$, $m_i(t^*) \ll \langle x', m_i(t^*)\rangle$, then on this set E_i, f_i^* has no effect on the system, and so there we can give this control any value we choose.

Accordingly we call the control system $m(t^*)$ essentially normal in X, if for any nonzero $x' \in X'$, the measures $m_i(t^*) \ll \langle x', m_i(t^*)\rangle$, $i = 1, 2, \cdots$. We say that

the optimal control is essentially unique, if any two optimal controls are $m(t^*)$-equivalent. Similarly, we say the optimal control, f^*, is essentially bang-bang if there is a bang-bang control $m(t^*)$-equivalent to f^*, and we say the optimal control is essentially determined by (4), if any two solutions of (4) are $m(t^*)$-equivalent. Then from the remarks above, we have

THEOREM 2. *Suppose* (A), (B), (C) *hold,* $\operatorname{int} W \neq \varnothing$, *and* $t^* > 0$ *is the minimum time. Then if* $m(t^*)$ *is essentially normal in* X, *the optimal control is essentially bang-bang, essentially unique and essentially determined by the necessary condition* (4).

Proof. We know from Theorem 1 that $f^* \in L_{\text{ex } F}(R^\infty, (\langle x', m_i(t^*)\rangle))$ for some nonzero $x' \in X'$. However, the essential normality of the system implies that each measure $\langle x', m_i(t^*)\rangle$ is equivalent to $m_i(t^*)$, $i = 1, 2, \cdots$, and hence $L_{\text{ex } F}(R^\infty, (\langle x', m_i(t^*)\rangle)) = L_{\text{ex } F}(R^\infty, m(t^*))$ as sets, and so $f^* \in L_{\text{ex } F}(R^\infty, m(t^*))$.

It is not hard to see that essential normal systems are a natural extension of the same concept discussed in [7] for finite-dimensional control systems. Using the results of [9, Chap. 6], we have, in fact, analogues of Theorem 15.1 and its Corollary from [7].

LEMMA 3. *A control* f^* *is essentially determined by* (4), *for some nonzero* $x' \in X'$, *if and only if* $m(t^*, f^*) \in \exp A(t^*)$.

Proof. If f^* is essentially determined by (4) for some $x' \in X'$, then it follows easily from the definitions that this x' exposes $A(t^*)$ at $m(t^*, f^*)$.

Conversely, suppose x' exposes $A(t^*)$ at $m(t^*, f^*)$. Then for this x', (4) holds, and if g is another solution of (4), we must have $m(t^*, g) = m(t^*, f^*)$, as x' exposes the set $A(t^*)$. However, the point $m(t^*, f^*)$ $(= m(t^*, g))$ is also an extreme point of $A(t^*)$, and so by [14, Thm. 4] g and f^* are $m(t^*)$-equivalent.

As a converse to the uniqueness part of Theorem 2 we have,

LEMMA 4. *Suppose* (A), (B), (C), *hold,* $\operatorname{int} W \neq \varnothing$, *and* $t^* > 0$ *is the minimum time. If for any nonzero* $x' \in X'$, *the solution of* (4) *is essentially unique, then* $m(t^*)$ *is essentially normal in* X.

Proof. Suppose $x' \in X'$ is nonzero, and $E_i \in \mathcal{T}_{t^*}$ are an any sets such that $|\langle x', m_i(t^*)\rangle|(E_i) = 0$, $i = 1, 2, \cdots$. Let F_i^+ (F_i^-) be the positive (respectively, negative) part of $\mathcal{T}_{t^*}$ relative to the measure $\langle x', m_i(t^*)\rangle$. Then it is easily seen that both the functions $(\chi_{F_i^+} - \chi_{F_i^-})$ and $(\chi_{F_i^+ \Delta E_i} - \chi_{F_i^-})$ are solutions of (4) (with respect to x'); hence by our assumption, these functions are $m(t^*)$-equivalent. This can only happen if E_i is $m_i(t^*)$-null for every $i = 1, 2, \cdots$.

Then combining Lemmas 3 and 4, we have

THEOREM 3. *The system* $m(t^*)$ *is essentially normal in* X *if and only if every supporting hyperplane to* $A(t^*)$ *exposes it.*

In the case X is a Banach space, $m : \mathcal{T} \to X$ a vector measure, it was first shown by Rybakov that there exists an $x' \in X'$ such that $m \ll \langle x', m\rangle$. The results given here bear a close resemblance to the work of Anantharaman [1], who showed that $m \ll \langle x', m\rangle$ if and only if x' exposes $\overline{\text{co}}\, m(\mathcal{T})$. In fact, it is known that the set of x' with this property forms a dense G_δ subset of X' (see [9, § VI.4]. Essentially normal systems give examples of vector measures for which every nonzero $x' \in X'$ gives a "Rybakov" measure.

Finally we note that if the measures $m_i(t^*)$, $i = 1, 2, \cdots$, are defined by integration with respect to the same scalar measure μ (for instance, Lebesgue measure), then we can define the stronger concept of normality. Namely, we call

$m(t^*)$ normal in X, if for every $x' \in X'$, $x' \neq 0$, $\mu \ll \langle x', m_i(t^*)\rangle$, $i = 1, 2, \cdots$. It follows that normal implies essentially normal, and that the following theorem holds.

THEOREM 4. *If the system is normal in X, Then the optimal control is uniquely determined by* (4) μ *a.e., and consequently, bang-bang* μ *a.e.*

In the next section we show how these concepts apply to the control of systems governed by partial differential equations.

5. Applications. We suppose Ω and $\mathcal{T}$ are as before, λ is a finite measure on $\mathcal{T}$, l is Lebesgue measure on $\mathcal{B}([0, t])$ and $\lambda \times l$ is the product measure on $\mathcal{T} \times \mathcal{B}([0, t])$.

In many situations the vector measures $m_i(t) : \mathcal{T}_t \to X$, $t \in [0, t_0]$, are the form

$$m_i(t)(x)(E) = \int \int_E K_i(x, \omega, t, \tau)\, d\lambda(\omega)\, d\tau, \qquad E \in \mathcal{T}_t,$$

$i = 1, 2, \cdots$, where x takes values in some domain $D \subset R^n$. Suppose X is a quasi-complete l.c.t.v.s. of real-valued functions defined on D, such that, for each $t \in [0, t_0]$, $i = 1, 2, \cdots$, and each $E \in \mathcal{T}_t$, the mapping $x \to m_i(t)(x)(E)$, $x \in D$, belongs to X and defines a closed vector measure. Suppose also that the measures $m_i(t)$, $i = 1, 2, \cdots$, are summable for each $t \in [0, t_0]$, so that $m(t) = (m_i(t))$ is a control system for each $t \in [0, t_0]$.

In this case, the necessary condition (4) becomes

There exists a non-zero $x' \in X'$, such that for any optimal control f^*

$$(4') \quad \sum_{i=1}^{\infty} \int_0^{t^*} \int_\Omega f_i^*(\omega, \tau)\langle x', K_i(\cdot, \omega, t^*, \tau)\rangle\, d\lambda(\omega)\, d\tau$$

$$= \max\left\{ \sum_{i=1}^{\infty} \int_0^{t^*} \int_\Omega f_i(\omega, \tau)\langle x', K_i(\cdot, \omega, t^*, \tau)\rangle\, d\lambda(\omega)\, d\tau : f \in \mathcal{BM}_F(R^\infty, \mathcal{T}_{t_0}) \right\}.$$

The condition for essential normality in X is

(5) For every nonzero $x' \in X'$, the set of points $(\omega, \tau) \in \Omega \times [0, t^*]$ such that $\langle x', K_i(\cdot, \omega, t^*, \tau)\rangle = 0$, and the function, $x \mapsto K_i(x, \omega, t^*, \tau)$, $x \in D$, is not the zero function in X, is $\lambda \times l$-null, $i = 1, 2, \cdots$.

The condition for normality is

(6) For every nonzero $x' \in X'$, $\{(\omega, \tau) : \langle x', K_i(\cdot, \omega, t^*, \tau)\rangle = 0\}$ is $\lambda \times l$-null, $i = 1, 2, \cdots$.

In particular, for the case $F \equiv I^\infty$, from (4') the optimal control has the form

$$f_i^*(\omega, \tau) = \operatorname{sgn}(\langle x', K_i(\cdot, \omega, t^*, \tau)\rangle), \qquad (\omega, \tau) \in \Omega \times [0, t^*].$$

Now suppose Ω is a bounded domain in R^n, and set $D^\alpha = D_1^{\alpha_1} \cdots D_n^{\alpha_n}$, where $D_j = \partial/\partial x_j$, $\alpha = (\alpha_1, \cdots, \alpha_n)$ and $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$. Consider the differential operator

$$Lu \equiv \frac{\partial u}{\partial t} - A(x, t, D)u \equiv \partial u/\partial t - \sum_{|\alpha| \leqq 2m} a_\alpha(x, t) D^\alpha u,$$

where m is a positive integer. Assume the boundary $\partial\Omega$ of Ω is sufficiently smooth, the coefficients $a_\alpha(x, t)$ are sufficiently smooth in $\bar{\Omega}\times(0, \infty)$ and that L is parabolic in the sense of Petrowski in $\bar{\Omega}\times(0, \infty)$.

Consider the following initial-boundary value problem

$$Lu = g \qquad \text{in } \Omega\times(0, t_0), \tag{7}$$

$$u(x, 0) = \phi(x), \qquad x \in \Omega, \tag{8}$$

$$B_j(x, t, D)u = f_i(x, t), \qquad (x, t) \in \partial\Omega\times(0, t_0), \quad 1 \leqq j \leqq m, \tag{9}$$

where the B_j are boundary operators which are sufficiently "regular" for the problem (7)–(9) to have a unique smooth solution.

Firstly consider the case of boundary control. That is, the functions g and ϕ are fixed and we suppose the control function $f = (f_1, \cdots, f_m)$ is chosen to be measurable and such that $f(\omega, \tau) \in F(\omega, \tau)$, $(\omega, \tau) \in \partial\Omega\times(0, t_0)$, for some fixed bounded measurable set-valued function $F : \partial\Omega\times(0, t_0) \to CCR^m$.

For smooth functions g, ϕ, f, we can represent the solution of (7)–(9) in the form

$$u(x, t) = \beta(x, t) + \sum_{j=1}^{m} \int_0^t \int_{\partial\Omega} G_j(x, \omega, t, \tau) f_j(\omega, \tau)\, d\lambda(\omega)\, d\tau, \tag{10}$$

where β is a fixed smooth function, λ is a surface measure on $\partial\Omega$, and $G_j (1 \leqq j \leqq m)$ are the appropriate Green's functions for the problem. We define (10) to be the solution of (7)–(9) for any bounded measurable function $f = (f_1, \cdots, f_m)$. Then from the remarks at the beginning of this section it can be seen the results of §§ 3 and 4 apply to this problem, where the set-functions $m_i(t) : \mathcal{T}_t \to X$, $t \in [0, t_0]$, are defined by

$$m_i(t)(x)(E) = \int \int_E G_i(x, \omega, t, \tau)\, d\lambda(\omega)\, d\tau, \qquad E \in \mathcal{T}_t, \quad x \in \Omega, \tag{11}$$

$i = 1, 2, \cdots, m$, and so by (10)

$$u(x, t) = \beta(x, t) + \sum_{j=1}^{m} \int_0^t \int_{\partial\Omega} f_j(\omega, \tau)\, d(m_j(t))(\omega, \tau), \qquad x \in \Omega.$$

The space X of real-valued functions defined on Ω must be chosen to satisfy the practical requirements of the system, bearing in mind the topology on X must be such that the set-functions (11) are countably additive.

In particular, consider for simplicity the case $m = 1$, $F(\omega, \tau) = [-1, 1]$, $(\omega, \tau) \in \partial\Omega\times(0, t_0)$, and take the boundary condition (9) to be

$$\frac{\partial u}{\partial \mu} + a(x, t)u = f(x, t) \quad \text{on } \partial\Omega\times(0, t_0), \tag{12}$$

where $\partial/\partial\mu$ is the outward transversal derivative on the lateral boundary and $a(x, t)$ is a smooth function. By using the known estimates for the Green's function

for this problem, it can be shown that if $X = L^p(\Omega)$ $(1 \leqq p < \infty)$ and (A) holds, then a time optimal control exists and condition (C) is true ([5, Lemma 1]). Consequently, by Theorem 1 there exists a nonzero $x' \in (L^p(\Omega))' = L^q(\Omega)$ $((1/p)+(1/q)=1)$, such that for any optimal control f^*,

$$\int_0^{t^*} \int_{\partial\Omega} f^*(\omega, \tau)\left(\int_\Omega x'(x)G(x, \omega, t^*, \tau)\, dx\right) d\lambda(\omega)\, d\tau$$

$$= \max\left\{\int_0^{t^*} \int_{\partial\Omega} f(\omega, \tau)\left(\int_\Omega x'(x)G(x, \omega, t^*, \tau)\, dx\right) d\lambda(\omega)\, d\lambda : |f| \leqq 1\right\};$$

in particular,

$$f^*(\omega, \tau) = \operatorname{sgn}\left(\int_\Omega x'(x)G(x, \omega, t^*, \tau)\, dx\right), \qquad (\omega, \tau) \in \partial\Omega \times (0, t).$$

we now give conditions for the system determined by (7), (8), (12) to be normal.

LEMMA 5. *If $a(x, t)$ and the coefficients of L are analytic functions, and if $\partial\Omega$ is an analytic manifold, then the system* (7), (8), (12) *is normal in* $L^p(\Omega)$ $(1 \leqq p < \infty)$.

Proof. Choose t such that $0 < t < t^*$ and define $K(\omega, \tau) = \int_\Omega x'(x)G(x, \omega, t^*, \tau)\, dx$, $(\omega, \tau) \in \bar{\Omega} \times [0, t)$. Then K satisfies $L^*K = 0$ in $\Omega \times [0, t)$, $\partial K/\partial\mu + aK = 0$ on $\partial\Omega \times (0, t)$, and so by [12] K is analytic in $\bar{\Omega} \times [0, t)$.

If the system is not normal, from our earlier remarks (equation (6)) there must exist a nonnegligible subset E of $\mathcal{T}_t$ such that K is zero on E for any t sufficiently close to t^*.

For $\tau \in [0, t)$, let $E^\tau = \{\omega : (\omega, \tau) \in E\}$. Then by Fubini's theorem there must exist a subset $\Delta \subset [0, t)$ of positive measure, such that for each $\tau \in \Delta$, E^τ is non-λ-null. Since the function $\omega \to K(\omega, \tau)$ for fixed $\tau \in \Delta$, is analytic on $\partial\Omega$, and $\partial\Omega$ is an analytic manifold, it follows that this function is identically zero on $\partial\Omega$. Hence for $(\omega, \tau) \in \partial\Omega \times \Delta$, $K(\omega, \tau) = 0$. It then follows from [5, Lemma 2] that $K(\omega, t) = 0$ for all $\omega \in \Omega$. Now taking a sequence $t_n \uparrow t^*$, we obtain $x'(x) = 0$ a.e. on Ω, which contradicts our initial assumption.

In conclusion we have,

THEOREM 5. *For the problem* (7), (8), (12), *if* (A) *holds,* $X = L^p(\Omega)$ $(1 \leqq p < \infty)$, *then there exists an optimal control and a nonzero* $x' \in L^q(\Omega)$ $((1/p)+(1/q)=1)$ *such that* (4′) *holds, and consequently, any optimal control f^* satisfies*

$$f^*(\omega, \tau) = \operatorname{sgn}\left(\int_\Omega x'(x)G(x, \omega, t^*, \tau)\, dx\right), \quad (\omega, \tau) \in \partial\Omega \times [0, t^*).$$

If the coefficients of L and $a(x, t)$ are analytic functions and if $\partial\Omega$ is an analytic manifold, then the optimal control is uniquely determined by (4′), $\lambda \times l$ *a.e., and is bang-bang* ($\lambda \times l$ *a.e.*). *If* $f^*(\omega, \tau) = f^*(\tau)$, $(\omega, \tau) \in \partial\Omega \times [0, t^*)$, *then the optimal control has at most a countable number of switchings.*

The only extra information follows from the analyticity of the function K defined in Lemma 5.

Suppose now the function g in (7) is the control function, and ϕ and f in (8), (9) are fixed smooth functions. With the aid of the Green's function we write the solution of (7)–(9) in the form

$$u(x, t) = \beta_0(x, t) + \int_0^t \int_\Omega G(x, \omega, t, \tau) g(\omega, \tau)\, d\omega\, d\tau, \qquad (x, t) \in \bar{\Omega} \times [0, t_0],$$

where β_0 is a fixed smooth function. We assume the boundary conditions are sufficiently "regular" so that G satisfies

$$|G(x, \omega, t, \tau)| \leqq \left(\frac{c}{(t-\tau)^{n/2m}}\right) \exp\left\{-c\left[\frac{|x-\omega|^{2m}}{t-\tau}\right]^{1/(2m-1)}\right\}.$$

Then, as before, we have

THEOREM 6. *If* (A) *holds,* $X = L^p(\Omega)$ $(1 \leqq p < \infty)$, *then an optimal control exists and the necessary condition* (4′) *holds at the minimum time* t^*. *If the coefficients of* L *and* B_j $(1 \leqq j \leqq m)$ *are analytic,* Ω *is an open set,* $\partial\Omega$ *an analytic manifold, and* (I′) *of* [12] *holds, then the optimal control is uniquely determined by* (4′) (*l a.e.*) *and is bang-bang* (*l a.e.*).

If the partial differential equation can be solved by separation of variables, then the normality of the system can be deduced from the completeness properties of the eigenfunctions of the equation. This can be of interest when we wish to consider the normality of systems under stronger topologies on X.

For instance, suppose the control system is governed by a partial differential equation on an open domain $\Omega \subset R^m$ whose solution for any control f, can be written in the form

$$u(x, t) = \sum_{n=1}^{\infty} v_n(x) \int_0^t \int_{\Omega_0} g_n(\omega, t, \tau) f(\omega, \tau)\, d\omega\, d\tau, \qquad x \in \bar{\Omega},\ t \in [0, t_0],$$

where Ω_0 is a (possibly empty) subset of Ω. Then if X is a space of real-valued functions on $\bar{\Omega}$, such that $v_n \in X$, $n = 1, 2, \cdots$, and the set-function, for each $t \in [0, t_0]$,

$$m(t)(x)(E) = \sum_{n=1}^{\infty} v_n(x) \int\int_E g_n(\omega, t, \tau)\, d\omega\, d\tau, \qquad x \in \bar{\Omega},$$

$E \in \mathcal{B}(\Omega_0 \times [0, t))$ is a vector measure in X, then (c.f. [4]) we have Lemma 6.

LEMMA 6. *If the linear span of the functions* $\{v_n : n = 1, 2, \cdots\}$ *is not dense in* X, *then the system* $m(t^*)$ *is not essentially normal in* X *for any* $t^* \in [0, t_0]$. *Conversely, if for almost all* $\omega \in \Omega_0$, *the functions* $\tau \to g_n(\omega, t^*, \tau)$, $\tau \in [0, t^*)$, *are linearly independent* (*that is, if* (a_n) *are real numbers such that* $\sum a_n g_n(\omega, t^*, \tau) = 0$, *for almost all* $\tau \in [0, t^*)$, *then* $a_n = 0$, $n = 1, 2, \cdots$), *if for any* $x' \in X'$ *and almost all* $\omega \in \Omega_0$, *the function,* $\tau \to \sum_{n=1}^{\infty} \langle x', v_n \rangle g_n(\omega, t^*, \tau)$, $\tau \in [0, t^*)$, *is analytic, and if the functions* $\{v_n\}$ *span* X, *then* $m(t^*)$ *is normal in* X.

Proof. If the functions $\{v_n\}$ don't span X, by the Hahn–Banach theorem, there exists a nonzero $x' \in X'$ such that $\langle x', v_n \rangle = 0$ for all $n = 1, 2, \cdots$. Then $|\langle x', m(t^*)\rangle|(\Omega_0 \times [0, t^*)) = 0$ for any $0 \leqq t^* \leqq t_0$, and so the system cannot be essentially normal in X.

Conversely, suppose $x' \in X'$, $x' \neq 0$, and $|\langle x', m(t^*)\rangle|(E) = 0$ for some subset $E \in \mathcal{B}(\Omega_0 \times [0, t))$, E having nonzero Lebesgue measure. For $(\omega, \tau) \in E$, $\sum_{n=1}^{\infty} \langle x', v_n\rangle g_n(\omega, t^*, \tau) = 0$. Hence by Fubini's theorem and analyticity we can show, in the same way as in Lemma 5, that there exists a nonnull set $B \subset \Omega_0$ such that $\sum \langle x', v_n\rangle g_n(\omega, t^*, \tau) = 0$ for all $(\omega, \tau) \in B \times (0, t^*)$. Then by linear independence, we must have $\langle x', v_n\rangle = 0$ for all $n = 1, 2, \cdots$, and so $x' = 0$ as the functions $\{v_n\}$ span X. This contradicts the initial assumption; consequently $m(t^*)$ is normal in X.

As an example consider the following parabolic boundary value problem:

$$\frac{\partial u}{\partial t}(x, t) - \frac{\partial^2 u}{\partial x^2}(x, t) + q(x)u(x, t) = 0, \qquad (x, t) \in (0, 1) \times (0, t_0], \tag{13}$$

$$u(1, t) + \alpha \frac{\partial u}{\partial x}(1, t) = f(t), \tag{14}$$

$$\frac{\partial u}{\partial x}(0, t) = 0, \qquad 0 < t \leqq t_0, \tag{15}$$

$$u(x, 0) = 0, \qquad 0 \leqq x \leqq 1, \tag{16}$$

where $t_0 > 0$ and $\alpha > 0$ are fixed parameters, and $q(x) \geqq 0$ is a fixed continuously differentiable function on $[0, 1]$. The normality of this system for $q = 0$ and $X = L_2([0, 1])$ was shown in [4]. We prove that it is normal in $C([0, 1])$.

For sufficiently smooth functions f, the solution can be written

$$u(x, t) = \sum_{n=1}^{\infty} A_n \mu_n v_n(x) \int_0^t f(t) e^{-\mu_n (t-\tau)}\, d\tau, \tag{17}$$

where $\{v_n\}$ and $\{\mu_n\}$ are, respectively, the normed eigenfunctions and eigenvalues of the boundary value problem

$$-v''(x) + q(x)v(x) = \mu v(x), \qquad x \in (0, 1),$$

$$v'(0) = 0,$$

$$v(1) + \alpha v'(1) = 0,$$

and $A_n = \int_0^1 v_n(x)\, dx$ for $n \geqq 1$. For any bounded measurable function f, define (17) to be the solution of (13)–(16). Then the analyticity and linear independence conditions of the Lemma are satisfied (e.g., [13, Lemma 3]); hence the system (13)–(16) is normal in any space X which is spanned by the eigenfunctions $\{v_n\}$. In particular, it is known (see [6, Prop. 3.4] or [11, p. 143]) that $\{v_n\}$ span $C([0, 1])$.

In the case of (13) with initial control, that is,

$$\frac{\partial u}{\partial t}(x, t) - \frac{\partial^2 u}{\partial x^2}(x, t) + q(x)u(x, t) = 0, \qquad (x, t) \in (0, 1) \times (0, t_0], \tag{18}$$

$$u(1, t) + \alpha \frac{\partial u}{\partial x}(1, t) = 0, \tag{19}$$

$$\frac{\partial u}{\partial x}(0, t) = 0, \qquad 0 < t \leqq t_0, \tag{20}$$

$$u(x, 0) = f(x), \qquad 0 \leqq x \leqq 1, \tag{21}$$

the solution is of the form

$$u(x, t) = \sum_{n=1}^{\infty} e^{-\mu_n t} v_n(x) \int_0^1 f(\omega) v_n(\omega)\, d\omega, \qquad 0 \leqq x \leqq 1,$$

and it follows from the properties of $\{v_n\}$ and $\{\mu_n\}$ (e.g., [6, Prop. 3.3]) that for any $x' \in (C([0, 1]))'$, the function

$$\omega \to \sum_{n=1}^{\infty} e^{-\mu_n t} \langle x', v_n \rangle v_n(\omega), \qquad \omega \in [0, 1], \tag{22}$$

has an analytic continuation onto the whole complex plane. Hence the system is normal in $C([0, 1])$, and since the function (22) can have only a finite number of zeros on $[0, 1]$, the optimal control can have only a finite number of switchings.

As a final remark, it may be interesting to consider the problems involved if the restriction, int $W \neq \varnothing$, is dropped. In particular, suppose X is an infinite-dimensional Banach space, and $W = \{x\}$ for some fixed $x \in X$. Then under assumptions (A), (B) and (C), a time optimal control exists, and if t^* is the minimum time, $x \in \partial A(t^*)$. However, the existence of a supporting hyperplane to $A(t^*)$ at x cannot be guaranteed, as the set $A(t^*)$ need not contain interior points. Indeed, $A(t^*)$ is weakly compact ([8, Thm. 1]), and if it has nonempty interior, then X must be reflexive. However, if X is reflexive and the measures $m_i(t^*)$, $i = 1, 2, \cdots$, have σ-finite variation, then $A(t^*)$ will be compact and so will only have interior points if it is finite-dimensional. (See [9, Cor. IX.4.2.] for the case of measures with bounded variation; the σ-finite case follows by reduction.)

This parallels the fact that the bang-bang principle need not hold for such systems. Consider the parabolic problem (18)–(21). For the problem of approximate controllability, the optimal control is unique, bang-bang and has only a finite number of switchings. However, if we wish to hit exactly a distribution of temperature $x \in C([0, 1])$ in minimum time, then the optimal control need not be bang-bang. This follows easily from [10, Thm. 3], as the functions $\{v_n\}$ span $L^1([0, 1])$, and so if f is any bounded measurable function on $[0, 1]$, with $\int_0^1 f(x) v_n(x)\, dx = 0$, for all $n = 1, 2, \cdots$, then $f = 0$. Then by [10, Thm.3], for any time $t > 0$, we have $A_F(t) \neq A_{\text{ex } F}(t)$,or there exist temperature distributions reachable by admissible controls in time t, but not reachable by bang-bang controls.

6. Approximate controllability. Suppose X is a Banach space, and we are given an element $z \in X$ and some $\varepsilon > 0$. In this section, we consider the problem of when $B(z, \varepsilon)$, the closed ball of radius ε about z in X, can be reached by the control system in some time $t > 0$.

Consider once again the situation modeled in § 3. If $(\Omega, \mathcal{T}, \lambda)$ is a finite measure space, set $\mathcal{T}_t = \mathcal{T} \times \mathcal{B}([0, t])$, and let $\lambda \times l$ be the product measure on $\mathcal{T}_{t_0}$. For each $t \in [0, t_0]$, we are given a vector measure $m(t) : \mathcal{T}_t \to X$. For the set of admissible controls, we take the unit ball in $L^\infty(\lambda \times l)$; that is, consider the case $F(\omega, \tau) = [-1, 1]$, all $(\omega, \tau) \in \Omega \times [0, t_0]$. Then if for any bounded measurable function f and time t, $m(t, f)$ is defined as in (2), the mapping $m(t) : L^\infty(\lambda \times l) \to X$ given by

$$m(t)(f) = m(t, f), \qquad f \in L^\infty(\lambda \times l),$$

is a bounded linear operator. (The boundedness is a consequence of the fact that $m(t)(f \in L^\infty(\lambda \times l) : 0 \leqq f \leqq 1) = \overline{\text{co}}\, m(t)(\mathcal{T}_t)$ is weakly compact and hence norm bounded in X [9, Thm. IV.6.1.].) Denote the adjoint of $m(t)$ by $m(t)^*$.

The symbol "$\|\cdot\|$" stands for the norm in X (or in X'), and "$\|\cdot\|_p$" for the norm in $L_p(\lambda \times l)$ $(1 \leqq p \leqq \infty)$. Note that we can regard $m(t^*) : X' \to L^1(\lambda \times l)$.

THEOREM 7. *If $B(z, \varepsilon)$ is reached in time t, then for all $x' \in X'$,*

$$|\langle x', z\rangle| - \varepsilon\|x'\| \leqq \|m(t)^*(x')\|_1. \tag{23}$$

Proof. Since $B(z, \varepsilon)$ is reached in time t, there exists a control f with $\|f\|_\infty \leqq 1$ and $\|z - m(t, f)\| \leqq \varepsilon$. Then for any $x' \in X'$,

$$|\langle x', z - m(t)(f)\rangle| \leqq \|x'\|\|z - m(t)(f)\| \leqq \varepsilon\|x'\|,$$

and so

$$\begin{aligned} |\langle x', z\rangle| - \varepsilon\|x'\| &\leqq |\langle x', m(t)(f)\rangle| = |\langle m(t)^*(x'), f\rangle| \\ &\leqq \|f\|_\infty \|m(t^*)(x')\|_1 \leqq \|m(t)^*(x')\|_1 \end{aligned}$$

by Hölder's inequality.

COROLLARY 1. *If $B(z, \varepsilon)$ is reached in time t, then $\|z\| - \varepsilon \leqq \|m(t)\|$.*

Proof. From (23) we have

$$\sup\{|\langle x', z\rangle| : \|x'\| = 1\} - \varepsilon \leqq \sup\{\|m(t)^*(x')\|_1 : \|x'\| = 1\}. \tag{24}$$

As the range of the operator $m(t^*)$ is a subset of L^1, we have that

$$\sup\{\|m(t^*)(x')\|_1 : \|x'\| = 1\} = \|m(t)^*\|,$$

and the result follows from (24), since an operator and its adjoint have the same norm.

COROLLARY 2. *If a time optimal control exists and $t^* > 0$ is the minimum time, and if condition* (C) *holds, then for any $x' \in X'$ satisfying* (3),

$$|\langle x', z\rangle| - \varepsilon\|x'\| = \|m(t^*)^*(x')\|_1. \tag{25}$$

Proof. Suppose f^* is an optimal control and x' satisfies (3). Then

$$\langle x', m(t^*, f^*)\rangle = \sup\{\langle x', m(t^*, f)\rangle : \|f\|_\infty \leqq 1\},$$

and since the set $A(t^*)$ is symmetric about zero,

$$\langle x', m(t^*, -f^*)\rangle = \inf\{\langle x', m(t^*, f)\rangle : \|f\|_\infty \leqq 1\}.$$

That is,

$$\begin{aligned} |\langle x', m(t^*, f^*)\rangle| &= \sup\{|\langle x', m(t^*, f)\rangle| : \|f\|_\infty \leqq 1\} \\ &= \sup\{|\langle m(t^*)^*(x'), f\rangle| : \|f\|_\infty \leqq 1\} \\ &= \|m(t^*)^*(x')\|_1. \end{aligned}$$

Also x' supports $B(z, \varepsilon)$ at $m(t^*, f^*)$, and so

$$0 \leqq \langle x', m(t^*, f^*)\rangle = \inf\{\langle x', z - w\rangle : \|w\| \leqq \varepsilon\}$$
$$= \langle x', z\rangle - \sup\{\langle x', w\rangle : \|w\| \leqq \varepsilon\}$$
$$= \langle x', z\rangle - \varepsilon\|x'\| = |\langle x', z\rangle| - \varepsilon\|x'\|,$$

and the result follows.

It may be worth noticing that if t_1 and x_1' ($\neq 0$) satisfy (25), and the function $t \to \|m(t)^*(x_1')\|_1$ is monotone increasing in a neighborhood of t_1, then by Theorem 7, $A(t) \cap B(z, \varepsilon) = \varnothing$ for any $t < t_1$, and so the minimum time $t^* \geqq t_1$. Hence if we compute a control f_1 corresponding to t_1 and x_1' in (4), and if $\|z - m(t_1, f_1)\| \leqq \varepsilon$, then t_1 is the minimum time and f_1 an optimal control.

In conclusion we show that the converse of Theorem 7 is also true.

THEOREM 8. *If for every $x' \in X'$ (23) holds, then $B(z, \varepsilon)$ is reached in time t.*

Proof. Suppose $B(z, \varepsilon) \cap A(t) = \varnothing$. Since both these sets are closed and convex, and the attainable set is weakly compact, we can find a nonzero $x' \in X'$ which separates them strictly. In other words, there exist constants $c, \delta > 0$, such that

$$\sup\{\langle x', m(t)(f)\rangle : \|f\|_\infty \leqq 1\} \leqq c - \delta$$

and

$$\inf\{\langle x', x\rangle : x \in B(z, \varepsilon)\} \geqq c.$$

It follows by using the symmetry of the sets $B(z, \varepsilon)$ and $A(t)$, as in Corollary 2 that

$$\|m(t)^*(x')\|_1 \leqq c - \delta \quad \text{and} \quad |\langle x', z\rangle| - \varepsilon\|x'\| \geqq c,$$

which contradicts (23).

Acknowledgment. I would like to record my thanks to the referee for the numerous improvements and clarifications he suggested.

REFERENCES

[1] R. ANANTHARAMAN, *On exposed points of the range of a vector measure*, Vector and Operator Valued Measures and Applications, Proc. Sympos. Snowbird Resort, Alta, Utah, 1972, Academic Press, New York, 1973, pp. 7–22.

[2] H. A. ANTOSIEWICZ, *Linear control systems*, Arch. Rational Mech. Anal., 12 (1963), pp. 313–324.

[3] G. CHOQUET, *Lectures on Analysis*, vol. II, W. A. Benjamin, New York, 1969.

[4] JU. V. EGOROV, *Certain problems in the theory of optimal control*, Dokl. Akad. Nauk SSSR, 145 (1962), pp. 720–723.

[5] A. FRIEDMAN, *Optimal control for parabolic equations*, J. Math. Anal. Appl., 18 (1967), pp. 479–491.

[6] K. GLASHOFF, *Optimal Control of One-dimensional Linear Parabolic Differential Equations*, Springer-Verlag Lecture Notes, vol. 477, New York, 1975, pp. 102–120.

[7] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[8] I. KLUVANEK AND G. P. KNOWLES, *Attainable sets in infinite-dimensional spaces*, Math. Systems Theory, 7 (1974), pp. 344–351.

[9] ——— *Vector Measures and Control Systems*, North-Holland Math. Studies, Amsterdam, 1976.

[10] G. P. KNOWLES, *Lyapunov vector measures*, this Journal, 13 (1974), pp. 294–303.

[11] S. G. MIKHLIN, *The Numerical Performance of Variational Methods*, Wolters-Nordhoff, Groningen, the Netherlands, 1971.

[12] H. TANABE, *On differentiability and analyticity of weighted elliptic boundary value problems*, Osaka J. Math., 1 (1965), pp. 163–190.

[13] K. TSUJIOKA, *Remarks on controllability of second order evolution equations in Hilbert spaces*, this Journal, 8 (1970), pp. 90–99.

[14] I. KLUVANEK, *The range of a vector-valued measure*, Math. Systems Theory, 7 (1973), pp. 44–54.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

INITIAL STATE DETERMINATION FOR DISTRIBUTED PARAMETER SYSTEMS*

TOSHIHIRO KOBAYASHI†

Abstract. The purpose of this paper is to give an approximate initial state which depends continuously on the measurement data.

In a distributed parameter system, the observability of the system is sufficient for an initial state to be uniquely determined from the measurement data, but is not sufficient for the initial state to depend continuously on the measurement data. That is, the problem of the initial state determination is not generally well-posed in the above sense.

In this paper, a well-posed approximate method is given for the initial state determination. The difference between a positive operator and a positive definite one in a Hilbert space plays an important role in this method.

1. Introduction. From the physical viewpoint, the system state functions may not be directly measurable and, instead, only certain restricted ones are actually obtained. In order to construct feedback control, however, complete knowledge of the state functions is required. It is necessary to determine the system state from the restricted measurement data. Therefore the state determination problem is very important from theoretical and practical points of view.

This problem is closely related to the concept of system observability [7]. In a distributed parameter system, observability assures that an initial state can be uniquely determined from the measurement data. As the space of initial states is an infinite-dimensional one, observability does not generally assure that the initial state depends continuously on the measurement data. That is, the problem of initial state determination for a distributed parameter system is not necessarily well-posed; this is different from a lumped parameter system [8], [9].

From the physical point of view, the measurement data have errors which may be very small. Even if the distributed parameter system is observable, the initial state determined from the measurement data is quite different from the desired initial state. From the numerical calculation point of view, if the problem is not well-posed, rounding errors may make the numerical solution meaningless, regardless of the accuracy of the arithmetic.

From the above facts, it is not sufficient to investigate only the observability of the distributed parameter system when we consider the problem of the initial state determination. In this paper, an approximate method is presented which reduces the nonwell-posed problem to a well-posed one. The method of generalized inverses [1] is used to construct a filter of bounded operators that converges pointwise to the inverse of the observability operator (possibly unbounded). That is, the positive observability operator is approximated by a family of positive definite ones.

* Received by the editors July 8, 1975, and in revised form November 19, 1975.

† Department of Control Engineering, Kyushu Institute of Technology, Tobata, Kitakyushu, Japan.

2. System description. In this section, system description is following Lions [5]. So let H and V be two Hilbert spaces with

$$V \subset H, \qquad V \text{ dense in } H; \tag{2.1}$$

the sign $\subset$ denotes both algebraic and topological inclusion. This means that the identity mapping of V in H is continuous. We denote by $(\cdot, \cdot)_V$ (respectively, $(\cdot, \cdot)_H$) and $\|\cdot\|_V$ (respectively, $\|\cdot\|_H$) the scalar product in V (respectively, H) and the norm on V (respectively, H). Let V' be the dual of V; we identify H with its dual so that

$$V \subset H \subset V'. \tag{2.2}$$

If $f \in V'$, $v \in V$, (f, v) denotes their scalar product; if $f \in H$, it coincides with the scalar product in H.

For each $t \in (0, T)$, we are given a continuous bilinear form $a(t; u, v)$ on V, having the following properties:

$\forall u, v \in V$, the function $t \to a(t; u, v)$ is measurable and

$$|a(t; u, v)| \leqq L\|u\|_V \cdot \|v\|_V, \qquad L = \text{constant independent of } t, u, v. \tag{2.3}$$

For fixed u in V, the linear form

$$v \to a(t; u, v)$$

is continuous on V; therefore it can be written

$$a(t; u, v) = (A(t)u, v), \qquad A(t)u \in V'. \tag{2.4}$$

We deduce also from (2.3) that

$$\|A(t)u\|_{V'} \leqq L\|u\|_V, \qquad \forall u \in V, \tag{2.5}$$

where $\|\cdot\|_{V'}$ is the dual norm of $\|\cdot\|_V$.

The family of operators $A(t) \in \mathscr{L}(V; V')$ (the space of continuous linear mappings from V onto V') is coercive; that is,

$$\begin{gathered} \text{there exists } \beta \text{ and } \alpha > 0 \text{ such that} \\ a(t; u, u) + \beta\|u\|_H^2 \geqq \alpha\|u\|_V^2, \qquad u \in V. \end{gathered} \tag{2.6}$$

Consider now the distributed parameter system described by the following evolutional equation;

$$\frac{du(t)}{dt} + A(t)u(t) = 0, \qquad t \in (0, T), \tag{2.7}$$

and

$$u(0) = u_0, \qquad u_0 \text{ given in } H. \tag{2.8}$$

Here $u' = du/dt$ is taken in the sense of *distributions* on $(0, T)$.

For this equation we have the following existence and uniqueness lemma.

LEMMA 1 (Lions [5]). *Under the assumptions* (2.3) *and* (2.6), *the system* (2.7) *and* (2.8) *has a unique solution u such that* $u \in L^2(0, T; V)$ *and* $u' \in L^2(0, T; V')$. *Furthermore, the solution u depends continuously on* u_0.

Remark. $L^2(0, T; F)$ denotes the space (equivalence class) of functions f defined on [0, T] with values in a Hilbert space F such that $\int_0^T \|f(t)\|_F^2 \, dt < \infty$.

From Lemma 1, there exists an operator $U(t)$ such that $U(t) \in \mathcal{L}(H; V)$, and the solution of the system (2.7) and (2.8) is given by

$$u(t) = U(t)u_0, \qquad t \in (0, T).$$

In physical situations, the space of observations K is finite-dimensional. The output of the system is given by

$$z(t) = M(t)u(t), \qquad 0 < t < T, \tag{2.9}$$

where $M(t)$ is a continuous linear operator from V to K for fixed $t \in (0, T)$, and there is a positive constant μ such that

$$\|M(t)u(t)\|_K \leqq \mu \|u(t)\|_V, \qquad u(t) \in V, \quad t \in (0, t). \tag{2.10}$$

The observed output $z(t)$ is written

$$z(t) = M(t)U(t)u_0, \qquad 0 < t < T. \tag{2.11}$$

From Lemma 1 and (2.10), it follows that $z \in L^2(0, T; K)$.

3. Observability. In this section, we investigate observability of the dynamical system described by (2.7) and (2.8) with the observation equation (2.9).

We start with the following definition.

DEFINITION 1. The system described by (2.7) and (2.8) with the observation equation (2.9) is said to be *observable at time* T if an initial state $u(0)$ can be uniquely determined from the observed measurement data $z(t)$ over the time interval $(0, T)$.

Let us define the observability operator $G(T)$ by

$$G(T)u_0 = \int_0^T U^*(t)M^*(t)M(t)U(t)u_0 \, dt, \qquad u_0 \in H. \tag{3.1}$$

Here $(\cdot)^*$ denotes the adjoint operator of an operator $(\cdot)$. From Lemma 1, $G(T) \in \mathcal{L}(H; H)$. Then we have the theorem for observability.

THEOREM 1. *The system described by* (2.7) *and* (2.8) *with the observation equation* (2.9) *is observable at time* T *if and only if the self-adjoint operator* $G(T)$ *is positive* [2], [6]; *that is,*

$$(G(T)h, h)_H \geqq 0, \qquad \forall h \in H,$$

and $(G(T)h, h)_H = 0$ *implies* $h = 0$.

The proof for this theorem will be given in the Appendix. The explicit conditions of observability were given by Kobayashi [3] for various types of systems with averaged, pointwise, scanning and boundary outputs.

We can give another important theorem as follows.

THEOREM 2. *Suppose that an operator* G *defined on a Hilbert space is self-adjoint and positive. Then its inverse* G^{-1} *is continuous if and only if* G *is positive definite* [2], [6]; *that is, there is a positive constant* γ *such that*

$$(Gh, h)_H \geqq \gamma \|h\|_H^2, \qquad \forall h \in H.$$

The proof for this theorem will be also given in the Appendix.

Remark. If H is finite-dimensional, the positive operator is always positive definite. This is shown as follows. In this case, the operator G is an $n \times n$ matrix for an n-dimensional Euclidean space H. Let eigenvalues of the matrix G be $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ and the corresponding eigenvectors be $\{\phi_1, \phi_2, \cdots, \phi_n\}$. Any n-dimensional vector $h \in H$ is described by

$$h = \sum_{i=1}^{n} h_i \phi_i.$$

Since

$$Gh = \sum_{i=1}^{n} h_i G\phi_i = \sum_{i=1}^{n} h_i \lambda_i \phi_i,$$

$$\begin{aligned}(Gh, h) &= \left(\sum_{i=1}^{n} h_i \lambda_i \phi_i, \sum_{i=1}^{n} h_i \phi_i \right) \\ &= \sum_{i,j=1}^{n} (h_i \lambda_i \phi_i, h_j \phi_j) \\ &= \sum_{i=1}^{n} \lambda_i h_i^2\end{aligned} \tag{3.2}$$

Now replacing h by ϕ_i $(i = 1, 2, \cdots, n)$, we obtain

$$(G\phi_i, \phi_i) = \lambda_i (\phi_i, \phi_i) > 0, \qquad i = 1, 2, \cdots, n,$$

which shows $\lambda_i > 0$ $(i = 1, 2, \cdots, n)$. Let $\gamma = \min \lambda_i$ (>0), and then

$$(Gh, h) \geqq \gamma \sum_{i=1}^{n} h_i^2 = \gamma \|h\|^2, \qquad \forall h \in H,$$

from (3.2). If the space H is infinite-dimensional, however, $\lambda_i > 0$ $(i = 1, 2, \cdots)$ does not necessarily imply $\inf \lambda_i > 0$. Thus there is no positive constant γ in general.

Next let us seek a unique initial state from the measurement data. Premultiply both sides of (2.10) by $U^*(t)M^*(t)$ and integrate over an interval $(0, T)$. We obtain

$$G(T)u_0 = \int_0^T U^*(t)M^*(t)z(t)\, dt. \tag{3.3}$$

Now define an operator $P(T)$ by

$$P(T)z = \int_0^T U^*(t)M^*(t)z(t)\, dt \tag{3.4}$$

and then $P(T) \in \mathscr{L}(L^2(0, T; K); H)$. Equation (3.3) becomes

$$G(T)u_0 = P(T)z. \tag{3.5}$$

If the system (2.7), (2.8) and (2.9) is observable at time T, the operator $G(T)$ is positive and has its inverse $G(T)^{-1}$. Thus the initial state is uniquely determined by

$$u_0 = G(T)^{-1}P(T)z. \tag{3.6}$$

For the distributed parameter system, the initial state space H is infinite-dimensional. Then Theorem 2 shows that the inverse operator G^{-1} is not necessarily continuous, even if the operator G is positive, that is, the system is observable at time T. If G^{-1} is not continuous, a small observation error induces a quite large error to the initial state determined by (3.6).

It is not practical to use (3.6) to seek an initial state from the measurement data. From numerical points of view, numerical algorithms which seek the initial state from (3.6) do not converge if G^{-1} is not continuous. When we treated only lumped parameter systems, these difficulties did not occur.

Therefore we should consider a new approximate method of determining the initial state which depends continuously on the measurement data.

4. Well-posed approximate method. In this section, we consider the approximate method of determining the initial state, which is well-posed. First consider the following estimation problem.

Problem I. Seek an optimal initial state which minimizes

$$J(u_0)=\int_0^T \|z_0(t)-M(t)u(t)\|_K^2\,dt \tag{4.1}$$

for the system (2.7), (2.8) and (2.9), where $z_0(t)$ denotes the measurement data and $M(t)u(t)$ denotes the output of the model system.

This problem is ordinarily used to determine numerically an initial state for lumped parameter systems. Next we introduce a regularized estimation problem corresponding to Problem I.

Problem II. Seek an optimal initial state which minimizes

$$J_\varepsilon(u_0)=\int_0^T \|z_0(t)-M(t)u(t)\|_K^2\,dt+\varepsilon(Nu_0, u_0)_H, \qquad \varepsilon>0, \tag{4.2}$$

for the system (2.7), (2.8) and (2.9), where

$$N\in\mathscr{L}(H, H), \quad (Nu_0, u_0)_H \geqq c\|u_0\|_H^2, \quad c>0.$$

If the system (2.7), (2.8) and (2.9) is observable at time T, both Problem I and Problem II have unique solutions, respectively. For Problem I specifically, there is a unique initial state u_0 such that $J(u_0)=0$, without measurement errors.

Now we shall seek the optimal solution for Problem I and Problem II. The performance index of Problem II, $J_\varepsilon(u_0)$, is transformed as follows:

$$J_\varepsilon(u_0)=\int_0^T \|z_0(t)-M(t)U(t)u_0\|_K^2\,dt+\varepsilon(Nu_0, u_0)_H \tag{4.3}$$

$$=\int_0^T (z_0(t)-M(t)U(t)u_0, z_0(t)-M(t)U(t)u_0)_K\,dt+\varepsilon(Nu_0, u_0)_H. \tag{4.4}$$

Since the operators $U(t)$ and $M(t)$ are continuous and $J_\varepsilon(u_0)$ is differentiable, the necessary optimality condition is

$$J_\varepsilon'(u_{0\varepsilon})\cdot h=0, \qquad \forall h\in H. \tag{4.5}$$

Now $J'_\varepsilon(u_{0\varepsilon})$ is explicitly calculated, and then (4.5) becomes

$$\int_0^T (M(t)U(t)u_{0\varepsilon} - z_0(t), M(t)U(t)h)_K\, dt + \varepsilon(Nu_{0\varepsilon}, h)_H = 0, \qquad \forall h \in H. \tag{4.6}$$

On the other hand, by using adjoint operators $U^*(t)$ and $M^*(t)$, (4.6) is transformed into

$$\int_0^T (U^*(t)M^*(t)M(t)U(t)u_{0\varepsilon} - U^*(t)M^*(t)z_0(t), h)_H\, dt + \varepsilon(Nu_{0\varepsilon}, h)_H = 0, \qquad \forall h \in H. \tag{4.7}$$

Moreover from (3.1) and (3.4), (4.7) becomes

$$(\varepsilon Nu_{0\varepsilon} + G(T)u_{0\varepsilon} - P(T)z_0, h)_H = 0, \qquad h \in H. \tag{4.8}$$

Since this equation holds for any $h \in H$, we obtain

$$(\varepsilon N + G(T))u_{0\varepsilon} = P(T)z_0 \tag{4.9}$$

as the equation by which the optimal solution $u_{0\varepsilon}$ is defined.

Similarly we obtain for Problem I

$$G(T)u_0 = P(T)z_0 \tag{4.10}$$

as the equation by which the optimal solution u_0 is determined. Now the optimal solution u_0 for Problem I is uniquely determined by

$$u_0 = G(T)^{-1}P(T)z_0, \tag{4.11}$$

if the system (2.7), (2.8) and (2.9) is observable at time T, that is, $G(T)$ is positive. However, the inverse $G(T)^{-1}$ is not continuous in general from Theorem 2; then the solution u_0 does not necessarily depend continuously on the measurement data z_0.

For Problem II, if the system (2.7), (2.8) and (2.9) is observable at time T, the operator $G_\varepsilon(T) = \varepsilon N + G(T)$ is positive definite, because

$$\begin{aligned}(G_\varepsilon(T)h, h)_H &= \varepsilon(Nh, h)_H + (G(T)h, h)_H \\ &\geqq \varepsilon c\|h\|_H^2, \qquad h \in H.\end{aligned} \tag{4.12}$$

From Theorem 2, there is a continuous inverse operator $G_\varepsilon(T)^{-1}$. Therefore the optimal solution $u_{0\varepsilon}$ for Problem II is uniquely determined by

$$u_{0\varepsilon} = G_\varepsilon(T)^{-1}P(T)z_0. \tag{4.13}$$

Moreover $u_{0\varepsilon}$ depends continuously on the measurement data z_0, as P(T) is a continuous operator.

Next we shall be able to show

$$\lim_{\varepsilon \to 0} \|u_{0\varepsilon} - u_0\|_H^2 = 0. \tag{4.14}$$

Putting $h = u_{0\varepsilon} - u_0$ in (4.6), $h = u_0 - u_{0\varepsilon}$ in the optimality condition for Problem I,

$$\int_0^T (M(t)u(t; u_0) - z_0(t), M(t)u(t; h))_K\, dt = 0, \qquad h \in H, \tag{4.15}$$

and adding, respectively, both sides of two equations, we obtain

$$\int_0^T (M(t)u(t; u_{0\varepsilon}) - M(t)u(t, u_0), M(t)u(t; u_{0\varepsilon}) - M(t)u(t; u_0))_K \, dt + \varepsilon(Nu_{0\varepsilon}, u_{0\varepsilon} - u_0)_H = 0. \tag{4.16}$$

Here $u(t; h) = U(t)h$. From this equation we have

$$(Nu_{0\varepsilon}, u_{0\varepsilon} - u_0)_H \leqq 0, \tag{4.17}$$

since $\varepsilon > 0$. But (4.17) implies

$$(Nu_{0\varepsilon}, u_{0\varepsilon}) \leqq (Nu_{0\varepsilon}, u_0) \leqq \|Nu_{0\varepsilon}\| \cdot \|u_0\| \leqq C\|u_{0\varepsilon}\| \cdot \|u_0\|$$

C being positive constant as $N \in \mathscr{L}(H; H)$. Since $(Nu_{0\varepsilon}, u_{0\varepsilon}) \geqq c\|u_{0\varepsilon}\|^2$, we obtain, consequently,

$$\|u_{0\varepsilon}\| \leqq \frac{C}{c}\|u_0\|. \tag{4.18}$$

Thus from every sequence of $\varepsilon \to 0$, we can extract a subsequence η such that $u_{0\eta} \to w$ weakly in H. From (4.9) we have

$$(\varepsilon Nu_{0\eta}, h) + (G(T)u_{0\eta}, h) = (P(T)z_0, h), \qquad \forall h \in H. \tag{4.19}$$

As $\eta \to 0$, (4.19) becomes

$$(G(T)w, h) = (P(T)z_0, h), \qquad \forall h \in H. \tag{4.20}$$

From (4.10) and (4.20), we have

$$(G(T)w, h) = (G(T)u_0, h), \qquad \forall h \in H. \tag{4.21}$$

Here putting $h = w - u_0$, we obtain

$$(G(T)(w - u_0), w - u_0) = 0. \tag{4.22}$$

From positiveness of $G(T)$ (the hypothesis of the system being observable at time T), we have

$$w = u_0.$$

Here $\{u_{0\eta}\}$ is an arbitrary, weakly convergent subsequence and its weak limit u_0 does not depend on how we choose subsequences. Therefore the extraction of a subsequence is unnecessary and $u_{0\varepsilon} \to u_0$ weakly in H (see [4]). Moreover from (4.17),

$$(N(u_{0\varepsilon} - u_0), u_{0\varepsilon} - u_0) \leqq -(Nu_0, u_{0\varepsilon} - u_0).$$

From this,

$$c\|u_{0\varepsilon} - u_0\|_H^2 \leqq -(Nu_0, u_{0\varepsilon} - u_0)_H,$$

which implies

$$\lim_{\varepsilon \to 0} \|u_{0\varepsilon} - u_0\|_H^2 = 0.$$

Now we notice that u_0 and $u_{0\varepsilon}$ are the solutions minimizing $J(u_0)$ and $J_\varepsilon(u_0)$, respectively, when the measurement data is z_0. Therefore u_0 is not the actual

initial state (denote it by u_0^*). We should evaluate $\|u_{0\varepsilon}-u_0^*\|_H$. Define u_ε^0 by

$$G_\varepsilon(T)u_\varepsilon^0=P(T)z. \tag{4.23}$$

Then

$$\|u_{0\varepsilon}-u_0^*\|\leqq\|u_{0\varepsilon}-u_\varepsilon^0\|+\|u_\varepsilon^0-u_0^*\|. \tag{4.24}$$

For the second term of the right-hand side, we can apply (4.14) to the case of $z_0=z$. As a result we obtain

$$\lim_{\varepsilon\to 0}\|u_\varepsilon^0-u_0^*\|_H=0. \tag{4.25}$$

Next, as for the first term, we obtain

$$u_{0\varepsilon}-u_\varepsilon^0=G_\varepsilon^{-1}(T)P(T)(z_0-z)$$

from (4.9) and (4.23). Since $\|G_\varepsilon(T)h\|\geqq\varepsilon c\|h\|$ for any $h\in H$,

$$\|G_\varepsilon^{-1}(T)\|\leqq\frac{1}{\varepsilon c}.$$

Thus we have

$$\|u_{0\varepsilon}-u_\varepsilon^0\|_H\leqq\frac{\|z_0-z\|}{\varepsilon c}\|P(T)\|.$$

If we can evaluate the measurement error by

$$\|z_0-z\|\leqq\delta,$$

we have

$$\|u_{0\varepsilon}-u_\varepsilon^0\|\leqq\frac{\delta}{\varepsilon c}\|P(T)\|. \tag{4.26}$$

The right-hand side of (4.26) tends to 0 as ε and δ tend to 0 where a relation $\delta=o(\varepsilon)$ holds (δ has a higher order than ε). Thus if ε and δ are chosen with $\delta=o(\varepsilon)$,

$$\lim_{\varepsilon,\delta\to 0}\|u_{0\varepsilon}-u_0^*\|_H=0. \tag{4.27}$$

On the other hand, we can give a posteriori estimate for $\|u_{0\varepsilon}-u_0^*\|_H$, if $\|P(T)^{-1}\|$ can be estimated. We put $u_0^*=u_{0\varepsilon}+y_\varepsilon$ in the identity $G(T)u_0^*=P(T)z$. Then we have

$$G(T)(u_{0\varepsilon}+y_\varepsilon)=P(T)z.$$

From this, it follows that

$$G_\varepsilon(T)u_{0\varepsilon}-P(T)z_0+G_\varepsilon(T)y_\varepsilon-\varepsilon N(u_{0\varepsilon}+y_\varepsilon)+P(T)(z_0-z)=0.$$

Since $G_\varepsilon(T)u_{0\varepsilon}-P(T)z_0=0$, we get

$$G_\varepsilon(T)y_\varepsilon-P(T)(z-z_0)-\varepsilon N(u_{0\varepsilon}+y_\varepsilon)=0. \tag{4.28}$$

This means that the element y_ε realizes the lower bound of the functional

$$I(u_0)=\int_0^T \|z(t)-z_0(t)+\varepsilon P(T)^{-1}N(u_{0\varepsilon}+y_\varepsilon)-M(t)u(t;u_0)\|_K^2\,dt + \varepsilon(Nu_0, u_0)_H. \tag{4.29}$$

Therefore

$$I(y_\varepsilon)\leqq I(0)=\int_0^T \|z(t)-z_0(t)+\varepsilon P(T)^{-1}N(u_{0\varepsilon}+y_\varepsilon)\|_K^2\,dt. \tag{4.30}$$

From this,

$$\varepsilon c\|y_\varepsilon\|_H^2\leqq\|z-z_0+\varepsilon P(T)^{-1}N(u_{0\varepsilon}+y_\varepsilon)\|_{L^2(0,T;K)}^2.$$

That is,

$$\begin{aligned}\sqrt{\varepsilon c}\|y_\varepsilon\|_H &\leqq\|z-z_0\|_{L^2(0,T;K)}+\varepsilon\|P(T)^{-1}\|\cdot\|N\|(\|u_{0\varepsilon}\|_H+\|y_\varepsilon\|_H)\\ &\leqq\delta+\varepsilon C\|P(T)^{-1}\|(\|u_{0\varepsilon}\|_H+\|y_\varepsilon\|_H).\end{aligned}$$

If we choose ε such that $\sqrt{c}-C\sqrt{\varepsilon}\|P(T)^{-1}\|>0$, we obtain

$$\|u_{0\varepsilon}-u_0^*\|_H\leqq\frac{\delta/\sqrt{\varepsilon}+C\sqrt{\varepsilon}\|P(T)^{-1}\|\cdot\|u_{0\varepsilon}\|_H}{\sqrt{c}-C\sqrt{\varepsilon}\|P(T)^{-1}\|}. \tag{4.31}$$

Summarizing, we get the following.

THEOREM 3. *If the system* (2.7), (2.8) *and* (2.9) *is observable at time T,*

(i) *for Problem* I, *there is a unique optimal solution* u_0 *which does not necessarily depend continuously on the measurement data* z_0;

(ii) *for Problem* II, *there is a unique solution* $u_{0\varepsilon}$ *which depends continuously on the measurement data* z_0 *and satisfies with the convergence property* (4.27).

This theorem shows that the optimal solution for Problem II, $u_{0\varepsilon}$, is an approximate initial state which depends continuously on the measurement data. Various methods for optimal control problems are applied to solve Problem II.

5. Conclusions. In this paper, we have investigated the initial state determination problem for a distributed parameter system. The problem relates to the concept of observability.

The initial state space is infinite-dimensional for the distributed parameter system. Even if the distributed system is observable, we cannot necessarily determine the initial state continuously dependent on the measurement data by the same method as that for a lumped parameter system. The initial state determined from the measurement data is quite different from the actual initial state, since the measurement data always have errors which may be very small. Therefore, we gave an approximate method which determines the approximate initial state continuously dependent on the measurement data. We analyzed this method from the difference between positive operators and positive definite ones in a Hilbert space. However, how to choose ε for this method was not sufficiently investigated in this paper. From the numerical calculation point of view, we gave a

well-posed approximate method, but we did not investigate the property of well-conditioned one. How to choose ε depends on each problem.

It is interesting to apply this method to a system which is not observable, by using pseudoinverse operators. It is also possible to apply this method to an identification problem for the distributed parameter system.

Appendix.

Proof of Theorem 1. As $z \in L^2(0, T; K)$,

$$\begin{aligned}\|z\|^2_{L^2(0,T;K)} &= \int_0^T \|z(t)\|^2_K \, dt \\ &= \int_0^T (M(t)U(t)u_0, M(t)U(t)u_0)_K \, dt \\ &= \int_0^T (U^*(t)M^*(t)M(t)U(t)u_0, u_0)_H \, dt.\end{aligned}$$

Since (3.1) and $G(T) \in \mathscr{L}(H; H)$, the integrand can be put into the inner product,

$$\|z\|^2_{L^2(0,T;K)} = (G(T)u_0, u_0)_H.$$

This shows that the operator $G(T)$ is nonnegative. Thus the initial state $u(0)$ is uniquely determined if and only if the operator $G(T)$ is positive. Therefore we obtain Theorem 2.

Proof of Theorem 2 (Mikhlin [6]). *Sufficiency.* Let the operator G be positive definite; that is, there exists a constant $\gamma > 0$ such that

$$(Gh, h) \geqq \gamma \|h\|^2, \qquad \forall h \in H. \tag{A.1}$$

Then, since

$$(Gh, h) \leqq \|Gh\| \cdot \|h\|, \tag{A.2}$$

$$\|Gh\| \geqq \gamma \|h\|. \tag{A.3}$$

This last inequality shows that the inverse operator G^{-1} exists and is bounded.

Necessity. Let G be a positive, but not a positive definite operator. That is, for any nonzero $h \in H$,

$$(Gh, h) > 0, \tag{A.4}$$

but there exists no positive constant γ such that

$$(Gh, h) \geqq \gamma \|h\|^2, \qquad \forall h \in H. \tag{A.5}$$

Equation (4) means that $\lambda = 0$ is not an eigenvalue of G, from which G^{-1} exists. Then the range set R_G is dense in H. This is shown as follows. Consider $\xi \in H$ such that

$$(Gh, \xi) = 0, \qquad \forall h \in H. \tag{A.6}$$

Equation (6) is rewritten as

$$(h, G^*\xi) = 0, \qquad \forall h \in H, \tag{A.7}$$

from which it follows that ξ belongs to the domain of definition of the operator G^* and that $G^*\xi = 0$. But $G^* = G$, and therefore $G\xi = 0$, from which $\xi = 0$, because 0 is not an eigenvalue of G. Thus R_G is dense in H.

Lastly, consider

$$m = \inf_h \frac{(Gh, h)}{(h, h)}. \tag{A.8}$$

Since G is positive but not positive definite, $m = 0$, from which it follows that 0 is a point of the continuous spectrum of G. Therefore G^{-1} is unbounded in H.

REFERENCES

[1] BEN-ISRAEL AND T. N. GREVILLE, *Generalized Inverses, Theory and Applications*, John Wiley, New York, 1974.

[2] S. G. KLEIN, *Linear Equations in Banach Spaces*, Moscow, 1967.

[3] T. KOBAYASHI, *Controllability and observability of distributed paramater systems*, Mem. Kyushu Inst. Tech. Engrg., (1975), no. 5, pp. 11–29.

[4] E. S. LEVINTIN AND B. T. POLJAK, *Convergence of minimizing sequences in conditional extremum problems*, Soviet Math. Dokl., Vol. 7 (1966), pp. 764–767.

[5] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[6] S. G. MIKHLIN, *The Problem of the Minimum of a Quadratic Functional*, Holden-Day, San Francisco, 1965.

[7] S. ROLEWICZ, *On optimal observability of linear systems with infinite-dimensional states*, Studia Math., 48 (1972), pp. 411–416.

[8] Y. SAKAWA, *Observability and related problems for partial differential equations of parabolic type*, this Journal, 13 (1975), pp. 15–27.

[9] P. K. C. WANG, *Mathematical modelling of systems with distributed parameters*, Control of Distributed Parameter Systems, American Society of Mechanical Engineers, New York, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

SPECTRAL MINIMALITY FOR INFINITE-DIMENSIONAL LINEAR SYSTEMS*

AVRAHAM FEINTUCH†

Abstract. Let $\{A, b, c\}$ be an infinite-dimensional system on a Hilbert space H. Suppose $f(\lambda) = ((\lambda I - A)^{-1}b, c)$ is the transfer function of the system. It is shown that if A is a compact operator or spectral operator, then the set of singularities of f is the spectrum of A.

1. Introduction. In recent years, attempts have been made to generalize the central results in finite-dimensional linear systems to infinite-dimensional Hilbert space.

It was seen that properties such as the state space isomorphism theorem [3, p. 113] and the spectral minimality property [1] do not hold for arbitrary infinite-dimensional systems (see, for example, [1], [7], [2]). Thus a number of authors decided to consider specific classes of linear operators such as restricted shift operators [1], [5], [6] or normal operators [2] and obtained interesting results. (See also [9], [10], [11]).

In this paper, we attempt to continue the work in this direction. We consider compact operators and spectral operators. Both of these classes are natural generalizations of finite-dimensional operators. We prove the spectral minimality property for such systems.

2. Preliminaries and notation. We begin with two linear spaces U and Y which we assume to be finite-dimensional. U will be called the control space and Y the output space. Let $\{A_i\}_{i=0}^{\infty}$ be a sequence of linear transformations from U to Y, and consider the linear transformation

$$y_n = \sum_{j=0}^{n-1} A_j u_{n-j-1}$$

sending sequences of elements of U into sequences of elements of Y. This transformation is called a (discrete) linear time-invariant input/output map, and the sequence $\{A_i\}$ is called its impulse response function.

Given a discrete constant linear system described by the equations

$$\begin{aligned} x_{n+1} &= Ax_n + Bu_n, \\ y_n &= Cx_n, \end{aligned} \tag{1}$$

where x_n belongs to a linear space X called the state space, $A \in B(X)$, $B \in L(U, X)$, $C \in L(X, Y)$, we will say that $\{A, B, C\}$ is a bounded realization of the above input output relation if $A_i = CA^{i-1}B$ for all i.

* Received by the editors February 13, 1975, and in revised form October 14, 1975.

† Department of Mathematics, Ben Gurion University of the Negev, Beersheva 84120, Israel.

DEFINITION. The system (1) is *controllable* if $\bigcap_i \ker B^*A^{*i} = \{0\}$ and *observable* if $\bigcap_i \ker CA^i = \{0\}$. The system will be called *canonical* if it is both controllable and observable.

If U, X, Y are finite-dimensional, then by the state space isomorphism theorem, two canonical systems $\{A, B, C\}$ and $\{A_1, B_1, C_1\}$ realize the same impulse response function, if and only if they are similar.

Here we will take X to be an infinite-dimensional Hilbert space, and to simplify certain proofs, we take our system to be single input, single output; i.e., we assume $U = Y = \mathbb{C}$. It is not hard to extend our resuls to the case $U = \mathbb{C}^p$, $Y = \mathbb{C}^r$. We also mention that the same technique holds if X is a Banach space.

Thus $B\alpha = \alpha b$ for some $b \in X$, and $Cx = (x, c)$ for some $c \in X$. We will denote the system by the triple $\{A, b, c\}$ with $A \in B(X)$. Thus $\{A, b, c\}$ is controllable if b is a cyclic vector for A and observable if c is a cyclic vector for A^*, and the weighting pattern $\{a_i\}$ is realized by $\{A, b, c\}$ if $a_i = (A^{i-1}b, c)$. The function

$$f(z) = ((z - A)^{-1}b, c)$$

is called the transfer function of the system. We will make use of the following characterization of all realizable transfer functions proved by Baras and Brockett in [1] and independently by Fuhrmann [5].

THEOREM 1. *The weighting pattern $\{a_i\}$ has a bounded realization, if and only if its transfer function is analytic at infinity and vanishes there.*

3. Spectral minimality. Let $f(z) = ((z - A)^{-1}b, c)$ be the transfer function of the system $\{A, b, c\}$. Let $\rho_0(A)$ denote the unbounded component of the resolvent $\rho(A)$. By $\sigma_0(A)$ we denote the compliment of $\rho_0(A)$. It was pointed out by Baras and Brockett that $\sigma(f) = \{z \in C | f \text{ is not analytic at } z\} \subseteq \sigma_0(A)$. They called this the spectral inclusion property and gave examples to show that often strict inclusion holds.

DEFINITION. A canonical realization $\{A, b, c\}$ of $\{a_i\}$ has the *spectral minimality property* if $\sigma(A) = \sigma(f)$.

The term spectral minimality was coined and introduced by Baras and Brockett in [1] and this property was also studied in [9], [10] and [11]. Here we show this holds when A is compact or spectral.

4. Compact operators. We assume the basic facts about the description of the spectrum of a compact operator, and consider various cases. In the case where A is quasi-nilpotent, we do not even need compactness.

DEFINITION. A is *quasi-nilpotent* if $\sigma(A) = \{0\}$.

THEOREM 2. *If A is quasi-nilpotent and f is the transfer function of the controllable system $\{A, b, c\}$, $c \neq 0$, then $\sigma(f) = \sigma(A) = \{0\}$.*

Proof. Consider $f(z) = ((z - A)^{-1}b, c)$. Then f is analytic everywhere except possibly 0. If f is analytic at 0, then f is entire, and by Theorem 1, f is bounded and $f(\infty) = 0$. By Liouiville's theorem, f is identically zero. Using the inverse Laplace transform, this implies that

$$(A^i b, c) = 0 \quad \text{for all} \quad i.$$

Since b is a cyclic vector for A, this implies that $c = 0$, which is impossible. This completes the proof.

Before proceeding, we need the following technical lemma which is a consequence of a theorem of Rosenthal [8].

LEMMA 3. *Let H and K be Hilbert spaces, $A \in B(H)$, $B \in B(K)$ and $C \in B(K, H)$. If $\sigma(A) \cap \sigma(B) = \varnothing$, then the operator*

$$\begin{pmatrix} A & C \\ 0 & B \end{pmatrix}$$

is similar to $A \oplus B$ on $H \oplus K$.

Proof. See [8, Chap. 0].

THEOREM 4. *Suppose A is a compact linear operator and $\lambda \neq 0$ is an eigenvalue of A. Then*

$$f(z) = ((z - A)^{-1} b, c)$$

is not continuable analytically to λ.

Proof. The subspace

$$N_\lambda = \{x | (A - \lambda I)^n x = 0\}$$

is a finite-dimensional invariant subspace of A. If we represent A as an operator matrix with respect to the decomposition $A = N_\lambda \oplus N_\lambda^\perp$, then

$$A = \begin{pmatrix} A_1 & C \\ 0 & A_2 \end{pmatrix}.$$

Now $\sigma(A_1) = \{\lambda\}$, $\sigma(A_2) = \sigma(A) \setminus \{\lambda\}$. Thus $\sigma(A_1) \cap \sigma(A_2) = \varnothing$. So A is similar to $A_1 \oplus A_2$. Since canonical realizations are invariant under similarity, we may as well assume $A = A_1 \oplus A_2$.

If P is the orthogonal projection on N_λ, then the system $\{A_1, Pb, Pc\}$ is canonical and $\sigma(A_1) = \{\lambda\}$. Thus by Theorem 2,

$$f_\lambda(z) = ((z - A_1)^{-1} Pb, Pc)$$

cannot be continued analytically to λ.

Now $f(z) = f_\lambda(z) + ((z - A_2)^{-1}(1 - P)b, (1 - P)c)$, where the second term is analytic at λ. Thus $f(z)$ is not analytically continuable to λ. This completes the proof.

We are now ready to prove the main result of this section.

THEOREM 5. *Let $\{A, b, c\}$ be a compact linear system and $f(z) = ((z - A)^{-1} b, c)$ its transfer function. Then $\sigma(f) = \sigma(A)$.*

Proof. We consider two cases.

Case (i). $\sigma(A)$ is finite. Let $\{\lambda_1, \cdots, \lambda_k, 0\}$ be $\sigma(A)$. By using Lemma 3 k times, we obtain that A is similar to an operator

$$\begin{pmatrix} A_1 & & & \\ & \ddots & & \\ & & A_k & 0 \\ 0 & & & A_{k+1} \end{pmatrix},$$

where $\sigma(A_i) = \{\lambda_i\}$, $1 \leqq i \leqq k$, and $\sigma(A_{k+1}) = \{0\}$. The result now follows by the argument in the previous theorem.

Case (ii). $\sigma(A)$ is infinite. Then 0 is the limit point of $\sigma(A)$. It was seen in Theorem 4 that $f(z)$ cannot be continued analytically to a nonzero point of the spectrum. If f is analytic at 0, then it is analytic in some neighborhood of 0. But every neighborhood of 0 contains nonzero points of $\sigma(A)$. This completes the proof.

A natural question to ask is whether the state space isomorphism theorem is true for compact linear systems. The following example, pointed out to the author by P. Fuhrmann, shows that this is not the case even if A is taken to be self-adjoint. A similar example has been described in [11, p. 697].

Example 6. Let $\{\lambda_i\}$ be a real sequence to 0, and let

$$A = \begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & 0 \\ 0 & & \ddots \end{pmatrix}$$

be an operator on l^2. Let $b = \{\beta_i\}$, $c = \{\gamma_i\}$ be sequences in l^2 such that $\beta_i \neq 0$, $\gamma_i \neq 0$ for all i.

Consider the systems $\{A, b, c\}, \{A, c, b\}$, and suppose they are similar; i.e., there exists a bounded invertible linear transformation R such that $AR = RA$, $Rb = c$.

Since R commutes with A,

$$R = \begin{pmatrix} \rho_1 & & \\ & \rho_1 & 0 \\ 0 & & \ddots \end{pmatrix},$$

and $Rb = c$ implies $\rho_i\beta_i = \gamma_i$; i.e., $\rho_i = \gamma_i/\beta_i$. If we choose β_i and γ_i such that $\{\beta_i\}$ converges to 0 much faster than $\{\gamma_i\}$, then R will not be bounded.

5. Spectral operators. We begin by presenting the necessary facts about spectral operators.

DEFINITION 7. Let X be a Banach space. A spectral measure in X is a homomorphic map of a Boolean algebra of sets in the plane into a Boolean algebra of projection operators in X such that it maps the unit in its domain into the identity operator. A spectral measure is bounded if the norms of the projections in its range are bounded.

DEFINITION 8. If Σ is a Boolean algebra of subsets of $\mathbb{C}$ which contains $\varnothing$ and $\mathbb{C}$; then a spectral measure E on Σ is called a resolution of the identity for the operator A if, for all $\delta \in \Sigma$,

(i) $E(\delta)A = AE(\delta)$,

(ii) $\sigma(T|E(\delta)X) \subseteq \bar{\sigma}$.

The classic example of a spectral resolution of the identity is given by the spectral theorem for normal operators on Hilbert space. In this case, the projections are self-adjoint, which is not true in general.

DEFINITION 9. A *spectral operator* is an operator with a countably additive resolution of the identity defined on the Borel sets of the plane.

DEFINITION 10. S is a *spectral operator of scalar type* if

$$S = \int \lambda E(d\lambda),$$

where E is the resolution of the identity for S.

The following canonical reduction of spectral operators shows how this class is a natural generalization of finite-dimensional operators.

THEOREM 11. *T is spectral, if and only if $T = S + N$, where S is spectral of scalar type, N is quasi-nilpotent and $SN = NS$. This decomposition is unique.*

For the proof of this, as well as most other facts about spectral operators, see [4].

We now come to the property of spectral operators that is important for this study.

Let $R(z; A)$ denote $(zI - A)^{-1}$ for $z \in \rho(A)$. If $x \in X$, then an analytic extension of $R(z; A)x$ will be an X-valued function f defined and analytic on an open set $D(f) \supseteq \rho(A)$ such that

$$(zI - A)f(z) = x \quad \text{for} \quad z \in D(f).$$

Then clearly $f(z) = R(z; A)x$ for $z \in \rho(A)$. It should be pointed out that the notion of "analytic extension" is different from that of "analytic continuation", for the domain $D(f)$ may contain points which cannot be connected with any point in $\rho(A)$ by a curve in $D(f)$.

To simplify matters, we will assume $\rho(A)$ is connected, and thus the two notions will be identical.

DEFINITION 12. The function $R(z; A)x$ has the *single-valued extension property* if for every pair f, g of analytic extensions of $R(z; A)x$ we have $f(z) = g(z)$ for $z \in D(f) \cap D(g)$.

The union of the sets $D(f)$ as f varies over all analytic extensions of $R(Z; A)x$ is called the resolvent set of x, denoted by $\rho(x)$. Its compliment will be called the spectrum of x, $\sigma(x)$.

If $R(z, A)x$ has the single-valued extension property, then there is a maximal extension $x(z)$ of $R(z; A)$ with domain $\rho(x)$. Then $x(z)$ is a single-valued analytic function with domain $\rho(x)$ and

$$x(z) = R(z; A)x \quad \text{for} \quad z \in \rho(A).$$

THEOREM 13. *If A is a bounded spectral operator in X, then for every $x \in X$, the function $R(z; A)x$ has the single-valued extension property.*

Proof. See [4, p. 1933].

It is interesting to note that the backward shift on H^2 does not have this property.

LEMMA 14. *If $A \in B(X)$ has the single-valued extension property, then for all $x \in X$,*

$$\sigma(x) \subseteq \sigma(A).$$

Also $\sigma(A) = \cup\{\sigma(x) | x \in X\}$.

Proof. See [4, p. 2093].

LEMMA 15. *For $x \in X$, denote by $\mathcal{M}_x$ the closed linear manifold determined by all the vectors $R(z; A)x$ with $z \in \rho(A)$. If A is spectral, then $\mathcal{M}_x$ is an invariant subspace of A and $\sigma(A|\mathcal{M}_x) = \sigma(x)$.*

Proof. See [4, p. 2171].

THEOREM 16. *Let A be a spectral operator and b a cyclic vector for A. Then $\sigma(b) = \sigma(A)$.*

Proof. Suppose b is a cyclic vector for A and consider $\mathcal{M}_b$. If $\mathcal{M}_b \neq X$, let c be a nonzero vector in $\mathcal{M}_b^{\perp}$. Then for all $z \in \rho(A)$, $((zI - A)^{-1}b, c) = 0$. If we take the inverse Laplace transform, we obtain that $(A^n b, c) = 0$ for all n. But since b is cyclic, this implies $c = 0$, which is impossible. Thus $\mathcal{M}_x = X$. By the previous lemma, it follows that $\sigma(A) = \sigma(b)$. This completes the proof.

THEOREM 17. *Let $\{A, b, c\}$ be a canonical spectral system, with transfer function $f(z) = ((zI - A)^{-1}b, c)$. Then $\{A, b, c\}$ has the spectral minimality property.*

Proof. Consider $f(z) = ((z - A)^{-1}b, c)$, and suppose $\lambda \in \sigma(A)$. Suppose $f(z)$ can be analytically continued to λ.

It follows easily that for every polynomial p, the function $((z - A)^{-1}p(A)b, c)$ is analytic at λ. Thus since $p(A)$ commutes with $(z - A)^{-1}$, by taking adjoints we obtain that for any polynomial p, the function $((z - A)^{-1}b, p(A^*)c)$ is analytic at λ. Since the system is canonical, c is a cyclic vector for A^*. Thus the function $b(z) = (z - A)^{-1}b$ has an analytic extension to λ. This contradicts the fact that $\sigma(b) = \sigma(A)$. The proof is complete.

Remark 18. This theorem is a significant generalization of Theorem 2.1 of [2], in the scalar case, since every normal operator is spectral (in fact, of scalar type).

Acknowledgment. I would like to thank Professor P. Fuhrmann for many valuable discussions on the problems considered here.

REFERENCES

[1] J. S. BARAS AND R. W. BROCKETT, *H^2-functions and infinite-dimensional realization theory*, this Journal, 13 (1975), pp. 221–241.

[2] R. W. BROCKETT AND P. A. FUHRMANN, *Normal symmetric dynamical systems*, this Journal, 14 (1976), pp. 107–119.

[3] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, London, 1970.

[4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Part III, Wiley-Interscience, New York, London, 1971.

[5] P. A. FUHRMANN, *On realization of linear systems and applications to some questions of stability*, Math. Systems Theory, 8 (1974), pp. 132–141.

[6] ———, *Exact controllability and observability and realization theory in Hilbert space*, Math. Anal. Appl., to appear.

[7] ———, *On spectral minimality of the shift realizations*, preprint.

[8] H. RADJAVI AND P. ROSENTHAL, *Invariant Subspaces*, Springer-Verlag, Berlin, 1973.

[9] J. S. BARAS, *On canonical realizations with unbounded infinitesimal generators*, Proc. of 11th Annual Allerton Conf. on Circuit and Systems Theory, 1973, pp. 1–10.

[10] ———, *Natural models for infinite-dimensional systems and their structural properties*, Proc. of 8th Princeton Conf. on Information Sciences and Systems, Princeton, N.J., 1974, pp. 195–199.

[11] J. S. BARAS, R. W. BROCKETT AND P. A. FUHRMANN, *State-space models for infinite-dimensional systems*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 693–700.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 5, August 1976

THE INFINITE-DIMENSIONAL RICCATI EQUATION FOR SYSTEMS DEFINED BY EVOLUTION OPERATORS*

RUTH CURTAIN AND A. J. PRITCHARD‡

Abstract. In the paper we consider the linear, quadratic control and filtering problems for systems defined by integral equations given in terms of evolution operators. We impose very weak conditions on the evolution operators and prove that the solution to both problems leads to an integral Riccati equation which possesses a unique solution. By imposing more structure on the evolution operator we show that the integral Riccati equation can be differentiated, and finally by considering an even smaller class of evolution operators we are able to prove that the differentiated version has a unique solution. The motivation for the study of such systems is that they enable us to consider wide classes of differential delay equations and partial differential equations in the same formulation. We derive new results for such a system and show how all of the existing results can be obtained directly by our methods.

Introduction. A number of recent reports and papers, [1]–[8], [13], [14], [16], [21] have considered the optimal control and filtering problems for evolution equations in Hilbert space. These problems lead to an infinite-dimensional Riccati equation which can take different forms depending on the structure assumed on the original dynamics. We feel that the most natural formulation of the problem is in terms of integral equations, or input-output relations, and these lead to an integral version of the Riccati equation. However, in order to make comparisons with finite-dimensional theory and for computational applications, we also ask what extra conditions must be imposed so that the integral Riccati equation may be differentiated, and so that the differentiated form has a unique solution.

The main applications of our results are to problems in which the system is modeled by either partial differential equations or differential-delay equations. There have, of course, been many papers devoted specifically to either one of these areas. For example, Lions [13], and Bensoussan [1] have derived a differential Riccati equation for the optimal control and filtering problems, where the differential operator $\mathcal{A}(t)$ is associated with a continuous bilinear form. Their results have applications to partial differential equations. Vinter [21] has considered the control problem for a class of hyperbolic partial differential equations, allowing for both boundary and distributed control. Mitter and Delfour [9] using essentially the ideas of Lions have achieved similar results for the control problem for differential delay equations. Another approach was adopted by Datko [8] and Pritchard [16] who have considered systems determined by an abstract operator $\mathcal{A}$ which generates a strongly continuous semigroup, thus enabling both differential-delay equations and certain partial differential equations to be considered in the same formulation. Curtain [7] and Mitter and Vinter [14] have examined the filtering problem for autonomous differential delay systems by using the special structure of the semigroup relevant to these systems. The general theory was extended to abstract operators $\mathcal{A}(t)$ which generate evolution operators $\mathcal{U}(t, s)$ by Curtain and Pritchard [5] for the particular application to partial differential

* Received by the editors May 6, 1975.
† Control Theory Centre, University of Warwick, Coventry CV4 7AL, England.

equations of the type studied by Kato and Tanabe [12]. In [4] Curtain adapted this evolution operator approach to obtain similar results for systems where $\mathscr{A}$ is associated with differential-delay equations.

Finally, in the monograph by Bensoussan, Delfour and Mitter [2] the integral Riccati equation is derived using the Lions approach for a more general class of evolution operator $\mathscr{U}(t, s)$ strongly continuous on $0 \leqq s < t \leqq T$.

In this paper we consider an even more general class of evolution operators than those of [2] in that $\mathscr{U}(t, s)$ is only assumed to be weakly continuous on $0 \leqq s < t \leqq T$, and which we call "mild evolution operator". We show that if $\mathscr{U}(t, s)$ is a mild evolution operator, then the optimal control problem leads to an integral Riccati equation. Moreover our approach does not follow Lions but is constructive in the sense that we derive a sequence of weakly continuous operators which converge to the unique solution of the integral Riccati equation. We then introduce the concept of a "quasi-evolution operator" in order to obtain a differential version of the Riccati equation. To ensure uniqueness it is necessary to suppose that $\mathscr{U}(t, s)$ is a strong evolution operator. However, this condition is satisfied by many of the applications to partial differential equations and differential delay equations in the literature. For the filtering problem, one is led to study a "dual" Riccati equation, and we prove that if $\mathscr{U}(t, s)$ is a mild evolution operator, the integral version of the dual Riccati equation has a unique solution. We obtain a differential version if $\mathscr{U}(t, s)$ is a strong evolution operator, but for uniqueness we need to suppose that $\mathscr{U}^*(T-s, T-t)$ is a strong evolution operator. These results include the previous results on the dual equation in [3], [7], [14], and Vinter and Curtain use them to obtain a general filtering theory for evolution operators in [6].

After developing the main results in §§ 1, 2, 3, we illustrate the results with a number of applications in § 4.

1. Perturbation theory for evolution operators. Evolution operators were studied by Kato and Tanabe in [12] for a class of parabolic partial differential equations with time-dependent coefficients. If the coefficients are time-invariant, the evolution operator is just the semigroup generated by the differential operator. Just as semigroups are not restricted to partial differential equations, evolution operators can be used to describe more general systems, including delay equations as in [20]. Here we define three types of evolution operators; mild, quasi and strong evolution operators. Strong evolution operators correspond to strong solutions of the homogeneous abstract evolution equation $\dot{z}(t) = \mathscr{A}(t)z(t)$, as was discussed by Kato and Tanabe in [12]. For control problems, the concept of a strong solution is too restrictive and so the term mild evolution operator is used in [2], [5], [7], [16], [18]. From a deeper study of the quadratic cost control problem and the dual filtering problem, we feel that a still weaker version is appropriate and, in addition, an intermediate evolution operator which we call a quasi-evolution operator.

DEFINITION 1.1. *Mild evolution operator.* Let H be a real Hilbert space and $T = [0, T]$ an interval of the real line and

$$\Delta(T) = \{(t, s) : 0 \leqq s < t \leqq T\}.$$

$\mathcal{U}(\cdot,\cdot):\Delta(T)\to\mathcal{L}(H)$ is a mild evolution operator if

$$\mathcal{U}(t,r)\mathcal{U}(r,s)=\mathcal{U}(t,s) \quad \text{for} \quad 0\leqq s\leqq r\leqq t\leqq T, \tag{1.1}$$

$$\mathcal{U}(t,s) \quad \text{is weakly continuous in } s \text{ on } [0,t] \text{ and in } t \text{ on } [s,T]. \tag{1.2}$$

We note that in [2], [5], [7], [16], [18], a mild evolution operator is used when $\mathcal{U}(\cdot,\cdot)$ satisfies (1.1) and is jointly strongly continuous. An interesting observation is that if we define

$$\mathcal{Y}(t,s)=\mathcal{U}^*(T-s,T-t),$$

then $\mathcal{Y}(\cdot,\cdot)$ defines a mild evolution operator on $\Delta(T)$ and $\mathcal{Y}(\cdot,\cdot)$ may be considered the "dual" to $\mathcal{U}(\cdot,\cdot)$. (See § 3 on the dual Riccati equation.)

Our main result is the following perturbation theorem.

THEOREM 1.1. *If $\mathcal{U}(\cdot,\cdot)$ is a mild evolution operator on $\Delta(T)$ and $D\in\mathcal{B}_\infty(T;\mathcal{L}(H))$ where*

$$\mathcal{B}_\infty(T;\mathcal{L}(H))=\left\{\begin{array}{c} D:T\to\mathcal{L}(H) \text{ such that } D(\cdot)x \text{ is strongly} \\ \text{measurable for each } x\in H \text{ and } \operatorname{ess\,sup}_{t\in T}\|D(t)\|<\infty \end{array}\right\},$$

then the following operator integral equation has a unique solution $\mathcal{U}_D(\cdot,\cdot)$,

$$\mathcal{U}_D(t,s)x=\mathcal{U}(t,s)x+\int_s^t \mathcal{U}(t,r)D(r)\mathcal{U}_D(r,s)x\,dr \tag{1.3}$$

in the class of weakly continuous bounded linear operators on H. $\mathcal{U}_D(\cdot,\cdot)$ is a mild evolution operator and we call it the perturbed mild evolution operator corresponding to the perturbation D. *Furthermore, if*

$$\operatorname*{ess\,sup}_{t\in T}\|D(t)\|\leqq M_1, \qquad \operatorname*{ess\,sup}_{\Delta(T)}\|\mathcal{U}(t,s)\|\leqq M_2,$$

we have

$$\|\mathcal{U}_D(t,s)\|\leqq M_1\exp M_1M_2(t-s).$$

Proof.

(a) *Existence.* We note that $\sup_{\Delta(T)}\|\mathcal{U}(t,s)\|\leqq M_2$, since $\mathcal{U}(\cdot,\cdot)$ is weakly continuous (see Appendix, Property A.3).

Let

$$\mathcal{U}_0(t,s)x=\mathcal{U}(t,s)x \quad \text{for all } x\in H,$$

$$\mathcal{U}_1(t,s)x=\int_s^t \mathcal{U}(t,r)D(r)\mathcal{U}(r,s)x\,dr,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\mathcal{U}_n(t,s)x=\int_s^t \mathcal{U}(t,r)D(r)\mathcal{U}_{n-1}(r,s)x\,dr,$$

where the integrals are well-defined Bochner integrals from the Appendix, Lemma A.1 and Property A.3.

We have

$$\|\mathcal{U}_n(t, s)x\| \leqq M_2(M_1M_2)^{n-1}\frac{(t-s)}{n!}\|x\|$$

by induction. Hence

$$\|\mathcal{U}_n(t, s)\| \leqq M_2(M_1M_2)^{n-1}\frac{(t-s)^n}{n!}$$

and

$$\left\| \sum_{n=1}^{N} \mathcal{U}_n(t, s)\right\| \leqq M_2 \exp\left(M_1M_2(t-s)\right) \quad \text{for all } N.$$

Therefore $\sum_{n=1}^{\infty} \mathcal{U}_n(t, s)$ is convergent on $\Delta(T)$ in the uniform topology and

$$\sup_{\Delta(T)} \left\| \sum_{n=1}^{\infty} \mathcal{U}_n(t, s)\right\| \leqq M^2 \exp\left(M_1M_2T\right).$$

But

$$\begin{aligned}\sum_{n=0}^{\infty} \mathcal{U}_n(t, s)x &= \mathcal{U}(t, s)x + \sum_{n=1}^{\infty} \mathcal{U}_n(t, s)x \\ &= \mathcal{U}(t, s)x + \sum_{n=1}^{\infty} \int_s^t \mathcal{U}(t, r)D(r)\mathcal{U}_{n-1}(r, s)x\, dr \\ &= \mathcal{U}(t, s)x + \int_s^t \sum_{n=1}^{\infty} \mathcal{U}(t, r)D(r)\mathcal{U}_{n-1}(r, s)x\, dr.\end{aligned}$$

Therefore $\mathcal{U}_D(t, s) = \sum_{n=1}^{\infty} \mathcal{U}_n(t, s)$ satisfies (1.4) and

$$\mathcal{U}_D(t, t) = I; \qquad \sup_{\Delta(T)} \|\mathcal{U}_D(t, s)\| \leqq M_2 \exp\left(M_1M_2T\right).$$

(b) *Uniqueness.* Suppose there is another solution $\mathcal{U}_1(t, s)$ and let

$$R(t, s) = \mathcal{U}_1(t, s) - \mathcal{U}_D(t, s).$$

Then

$$R(t, s)x = \int_0^t \mathcal{U}(t, r)D(r)R(r, s)x\, ds,$$

$$\therefore \quad \|R(t, s)x\| \leqq M_1M_2 \int_s^t \|R(r, s)x\|\, ds.$$

So $R(t, s)x = 0$ for all $x \in H$, by Gronwall's inequality.

(c) *Semigroup property.*

$$\mathcal{U}_D(t, r)\mathcal{U}_D(r, s)x = \mathcal{U}(t, r)\mathcal{U}(r, s)x + \int_s^r \mathcal{U}(t, r)\mathcal{U}(r, \beta)D(\beta)\mathcal{U}_D(\beta, s)x\, d\beta$$

$$+ \int_r^t \mathcal{U}(t, \beta)D(\beta)\mathcal{U}_D(\beta, r)\mathcal{U}_D(r, s)x\, d\beta,$$

$$\therefore \quad \mathcal{U}_D(t, r)\mathcal{U}_D(r, s)x - \mathcal{U}_D(t, s)x = \int_r^t \mathcal{U}(t, \beta)D(\beta)(\mathcal{U}_D(\beta, r)\mathcal{U}_D(r, s) - \mathcal{U}_D(\beta, s))x\, d\beta.$$

Let $R(\beta, r, s) = \mathcal{U}_D(\beta, r)\mathcal{U}_D(r, s) - \mathcal{U}_D(\beta, s)$.

$$\therefore \quad \|R(t, r, s)x\| \leqq M_1 M_2 \int_r^t \|R(\beta, r, s)x\| \, d\beta.$$

Since $\mathcal{U}_D(t, t) = I$, Gronwall's inequality implies $R(t, r, s)x = 0 \; \forall x \in H$ and $s \leqq r \leqq t$.

Hence $\mathcal{U}_D(t, r)\mathcal{U}_D(r, s) = \mathcal{U}_D(t, s)$ for $s \leqq r \leqq t$.

(d) *Continuity.* We show that $\mathcal{U}_D(t, \cdot)$ is weakly continuous on $[0, t]$ and $\mathcal{U}_D(\cdot, s)$ is weakly continuous on $[s, T]$. Consider

$$\phi(t, s)x = \int_s^t \mathcal{U}(t, r)D(r)\mathcal{U}_D(r, s)x \, dr.$$

Take $h > 0$, $t_1 \in [s, T)$, $t_2 \in (s, T]$. Then we have

$$\phi(t_1 + h, s)x - \phi(t_1, s)x = \int_s^{t_1} (\mathcal{U}(t_1 + h, r) - \mathcal{U}(t_1, r))D(r)\mathcal{U}_D(r, s)x \, dr + \int_{t_1}^{t_1+h} \mathcal{U}(t_1 + h, r)D(r)\mathcal{U}_D(r, s)x \, dr$$

and

$$\phi(t_2, s)x - \phi(t_2 - h, s)x = \int_s^{t_2-h} (\mathcal{U}(t_2, r) - \mathcal{U}(t_2 - h, r))D(r)\mathcal{U}_D(r, s)x \, dr + \int_{t_2-h}^{t_2} \mathcal{U}(t_2, r)D(r)\mathcal{U}_D(r, s)x \, dr,$$

$$\therefore \quad |\langle y, \phi(t_1 + h, s)x - \phi(t_1, s)x\rangle| \leqq \int_s^{t_1} |\langle y, \mathcal{U}(t_1 + h, r) - \mathcal{U}(t_1, r)D(r)\mathcal{U}_0(r, s)x\rangle| \, dr + \int_{t_1}^{t_1+h} M_1 M_2^2 \exp(M_1 M_2(r - s))\|x\|\|y\| \, dr.$$

Using the Lebesgue dominated convergence theorem and the weak continuity of $\mathcal{U}(\cdot, r)$ on $[s, T)$, we have

$$|\langle y, \phi(t_1 + h, s)x - \phi(t_1, s)x\rangle| \to 0 \quad \text{as } h \to 0.$$

Similarly,

$$|\langle y, \phi(t_2, s)x - \phi(t_2 - h, s)x\rangle| \leqq \int_s^{t_2-h} |\langle y, (\mathcal{U}(t_2, r) - \mathcal{U}(t_2 - h, r))D(r)\mathcal{U}_D(r, s)x\rangle| \, dr + \int_{t_2-h}^{t_2} M_1 M_2^2 \exp(M_1 M_2(r - s))\|x\| \, \|y\| \, dr$$

so $\phi(\cdot, s)$ is weakly continuous on $[s, T]$. To prove continuity with respect to the second variable, we use the fact that $\mathcal{U}_D(\cdot, \cdot)$ is also the unique solution of (1.4):

$$\mathcal{U}_D(t, s)x = \mathcal{U}(t, s)x + \int_s^t \mathcal{U}_D(t, r)D(r)\mathcal{U}(r, s)x \, dr \tag{1.4}$$

(this is proved in Corollary 1.2).

Consider

$$\psi(t, s)x = \int_s^t \mathcal{U}_D(t, r)D(r)\mathcal{U}(r, s)x\, dr.$$

We take $h>0$, $s_1 \in [0, t)$ and $s_2 \in (0, t]$. Then

$$\psi(t, s_1+h)x - \psi(t, s_1)x = \int_{s_1+h}^t \mathcal{U}_D(t, r)D(r)(\mathcal{U}(r, s_1+h) - \mathcal{U}(r, s_1))x\, dr$$
$$- \int_{s_1}^{s_1+h} \mathcal{U}_D(t, r)D(r)\mathcal{U}(r, s_1)x\, dr$$

and

$$\psi(t, s_2-h)x - \psi(t, s_2)x = \int_{s_2}^t \mathcal{U}_D(t, r)D(r)(\mathcal{U}(r, s_2-h) - \mathcal{U}(r, s_2))x\, dr$$
$$+ \int_{s_2-h}^{s_2} \mathcal{U}_D(t, r)D(r)\mathcal{U}(r, s_2-h)x\, dr,$$

$$\therefore \quad |\langle y, \psi(t, s_1+h)x - \psi(t, s_1)x\rangle|$$
$$\leqq \int_{s_1+h}^t |\langle D^*(r)\mathcal{U}_D^*(t, r)y, \mathcal{U}(r, s_1+h) - \mathcal{U}(r, s_1)x\rangle|\, dr$$
$$+ \int_{s_1}^{s_1+h} M_1M_2^2 \exp(M_1M_2\, T)\|x\|\, \|y\|\, dr$$
$$\to 0 \quad \text{as } h \to 0$$

by the Lebesgue dominated convergence theorem, since $\mathcal{U}$ is weakly continuous and $\|D^*(r)\| \leqq M_1$, $\|\mathcal{U}^*(t, r)\| \leqq M_2$.

Similarly,

$$|\langle y, \psi(t, s_2-h)x - \psi(t, s_2)x\rangle| \leqq \int_{s_2}^t |\langle D^*(r)\mathcal{U}_D^*(t, r)y, \mathcal{U}(r, s_2-h) - \mathcal{U}(r, s_2)x\rangle|\, dr$$
$$+ \int_{s_2-h}^{s_2} M_1M_2^2 \exp(M_1M_2\, T)\, dr\|x\|\, \|y\|$$
$$\to 0 \quad \text{as } h \to 0.$$

COROLLARY 1.1. *If $\mathcal{A}$ is the infinitesimal generator of a strongly continuous semigroup $\{\mathcal{T}(t), t \geqq 0\}$, then for $D \in \mathcal{B}_\infty(T; \mathcal{L}(H))$, the perturbed evolution operator $\mathcal{U}_D(t, s)$ is a mild evolution operator. Furthermore, since $\mathcal{T}(t)$ and $\mathcal{T}^*(t)$ are strongly continuous, the strong continuity of $\mathcal{U}_D(t, s)$ and $\mathcal{U}_D^*(t, s)$ follows from* (1.3).

COROLLARY 1.2. *The unique solution of* (1.4), $\mathcal{U}_D(t, s)$, *is also the unique solution of*

$$\mathcal{U}_D(t, s)x = \mathcal{U}(t, s)x + \int_s^t \mathcal{U}_D(t, r)D(r)\mathcal{U}(r, s)x\, dr, \qquad x \in H, \tag{1.4}$$

in the class of weakly continuous bounded linear operators on H.

Proof. As in the proof of Theorem 1.1(a), define

$$\mathcal{U}'_n(t, s)x = \int_s^t \mathcal{U}'_{n-1}(t, r)D(r)\mathcal{U}(r, s)x\, dr; \qquad \mathcal{U}'_0(t, s) = \mathcal{U}(t, s).$$

Then $\mathcal{U}'_D(t, s)x = \sum_0^\infty \mathcal{U}'_n(t, s)x$ is the unique solution of (1.4). We show that $\mathcal{U}'_n(t, s)x = \mathcal{U}_n(t, s)x$ for all n. Suppose the assertion is true for $n = k-1$, $k-2$; then

$$\begin{aligned}
\mathcal{U}'_k(t, s)x &= \int_s^t \mathcal{U}'_{k-1}(t, r)D(r)\mathcal{U}(r, s)x\, dr \\
&= \int_s^t \mathcal{U}_{k-1}(t, r)D(r)\mathcal{U}(r, s)x\, dr \\
&= \int_s^t \left(\int_r^t \mathcal{U}(t, \alpha)D(\alpha)\mathcal{U}_{k-2}(\alpha, r)D(r)\mathcal{U}(r, s)x\, d\alpha \right) dr \\
&= \int_s^t \left(\int_s^\alpha \mathcal{U}(t, \alpha)D(\alpha)\mathcal{U}_{k-2}(\alpha, r)D(r)\mathcal{U}(r, s)x\, dr \right) d\alpha \\
&\qquad \text{(changing the order of integration)} \\
&= \int_s^t \mathcal{U}(t, \alpha)D(\alpha)\mathcal{U}'_{k-1}(\alpha, s)x\, d\alpha \\
&= \int_s^t \mathcal{U}(t, \alpha)D(\alpha)\mathcal{U}_{k-1}(\alpha, s)x\, d\alpha \\
&= \mathcal{U}_k(t, s)x
\end{aligned}$$

and

$$\mathcal{U}'_0(t, s) = \mathcal{U}_0(t, s); \qquad \mathcal{U}'_1(t, s) = \mathcal{U}_1(t, s),$$

so

$$\mathcal{U}'_D(t, s) = \mathcal{U}_D(t, s).$$

The following class of evolution operators is motivated by requiring a differential form of the Riccati equation in § 2.

DEFINITION 1.2. *Quasi-evolution operator.* A *quasi-evolution operator* is a mild evolution operator $\mathcal{U} : \Delta(T) \to H$ such that there exists a nonzero $x \in H$ and a closed linear operator $\mathcal{A}(s)$ on H for almost all $s \in [0, T]$ satisfying

$$\langle y, \mathcal{U}(t, s)x - x\rangle = \int_s^t \langle y, \mathcal{U}(t, \rho)\mathcal{A}(\rho)x\rangle\, d\rho \quad \forall y \in H. \tag{1.5}$$

We denote the set of $x \in H$ for which (1.5) is valid as $\mathcal{D}_A$, and we call $\mathcal{A}(\cdot)$ the generator of $\mathcal{U}(\cdot, \cdot)$.

An immediate consequence of the definition is

$$\frac{\partial}{\partial s}\langle y, \mathcal{U}(t, s)x\rangle = -\langle y, \mathcal{U}(t, s)\mathcal{A}(s)x\rangle \quad \text{a.e. for } x \in \mathcal{D}_A, \quad y \in H, \quad t > s. \tag{1.6}$$

THEOREM 1.2. *If $\mathcal{U}$ is a quasi-evolution operator and $D \in \mathcal{B}_\infty(T; \mathcal{L}(H))$, then the perturbed mild evolution operator corresponding to D is also a quasi-evolution operator.*

Proof.

(a) From Theorem 1.1 and Corollary 1.2, $\mathcal{U}_D$ is uniquely defined by

$$\mathcal{U}_D(t, \rho)x = \mathcal{U}(t, \rho)x + \int_\rho^t \mathcal{U}_D(t, r)D(r)\mathcal{U}(r, \rho)x \, d\rho,$$

$$\therefore \quad \langle y, \mathcal{U}_D(t, \rho)\mathcal{A}(\rho)x\rangle = \langle y, \mathcal{U}(t, \rho)\mathcal{A}(\rho)x\rangle + \left\langle y, \int_\rho^t \mathcal{U}_D(t, r)D(r)\mathcal{U}(r, \rho)\mathcal{A}(\rho)x \, dr\right\rangle$$

for $x \in \mathcal{D}_A$, $y \in H$

$$= \langle y, \mathcal{U}(t, \rho)\mathcal{A}(\rho)x\rangle + \int_\rho^t \langle D^*(r)\mathcal{U}_D^*(t, r)y, \mathcal{U}(r, \rho)\mathcal{A}(\rho)x\rangle \, dr.$$

Both terms on the right side are integrable with respect to ρ on $(0, t)$ by (1.5) and so

$$\int_s^t \langle y, \mathcal{U}_D(t, \rho)\mathcal{A}(\rho)x\rangle \, d\rho = \int_s^t \langle y, \mathcal{U}(t, \rho)\mathcal{A}(\rho)x\rangle \, d\rho + \int_s^t \int_\rho^t \langle D^*(r)\mathcal{U}_D^*(t, r)y, \mathcal{U}(r, \rho)\mathcal{A}(\rho)x\rangle \, dr \, d\rho$$

for $x \in \mathcal{D}_A$, $y \in H$,

$$= \int_s^t \langle y, \mathcal{U}(t, \rho)\mathcal{A}(\rho)x\rangle \, d\rho + \int_s^t \int_s^r \langle D^*(r)\mathcal{U}_D^*(t, r)y, \mathcal{U}(r, \rho)\mathcal{A}(\rho)x\rangle \, d\rho \, dr$$

by (1.5) and changing the order of integration,

$$= \langle y, \mathcal{U}(t, s)x - x\rangle + \int_s^t \langle D^*(r)\mathcal{U}_D^*(t, r)y, \mathcal{U}(r, s)x - x\rangle \, dr \quad \text{by (1.5)},$$

$$\therefore \quad \int_s^t \langle y, \mathcal{U}_D(t, \rho)(\mathcal{A}(\rho) + D(\rho))x\rangle \, d\rho = \langle y, \mathcal{U}_D(t, s)x - x\rangle \quad \text{for } x \in \mathcal{D}_A, \ y \in H,$$

by Corollary 1.2.

So $\mathcal{U}_D(t, s)$ is a quasi-evolution operator with generator $\mathcal{A} + D$.

COROLLARY 1.3. *If $\mathcal{A}$ is the infinitesimal generator of a strongly continuous semigroup $\mathcal{T}(t)$, then for $D \in \mathcal{B}_\infty(T; \mathcal{L}(H))$, the perturbed evolution operator $\mathcal{U}_D(t, s)$ is a quasi-evolution operator. Furthermore, $\mathcal{U}_D(\cdot, \cdot)$ and $\mathcal{U}_D^*(\cdot, \cdot)$ are jointly strongly continuous on $\Delta(T)$.*

Proof. Consider $\mathcal{U}(t, s) = \mathcal{T}(t-s)$. Then from properties of strongly continuous semigroups [15], we have for $x \in \mathcal{D}(\mathcal{A})$,

$$\mathcal{A}\mathcal{T}(\cdot)x = \mathcal{T}(\cdot)\mathcal{A}x$$

is Bochner integrable, and

$$\mathcal{T}(t-s)x - x = \int_s^t \mathcal{T}(t-\alpha)\mathcal{A}x \, d\alpha.$$

Hence $\mathcal{T}(t)$ is a quasi-evolution operator for $x \in \mathcal{D}(\mathcal{A})$. Thus by the above theorem, the perturbed evolution operator is a quasi-evolution operator, and in fact,

$$(\mathcal{U}_D(t, s) - I)x = \int_s^t \mathcal{U}_D(t, r)(\mathcal{A} + D(r))x \, dr \quad \text{for } x \in \mathcal{D}(\mathcal{A})$$

and

$$\frac{\partial}{\partial s}(\mathcal{U}_D(t, s)x) = -\mathcal{U}_D(t, s)(\mathcal{A} + D(s))x \quad \text{a.e. for } x \in \mathcal{D}(\mathcal{A}).$$

The strong continuity comes from the strong continuity of $\mathcal{T}(t)$ and $\mathcal{T}^*(t)$ and the integral equation definition (1.3).

It is natural to ask whether the quasi-evolution operator has any connection with weak solutions of partial differential equations and we have the following result.

LEMMA 1.1. *If $\mathcal{U}(t, s)$ is a quasi-evolution operator on $\Delta(T)$, consider the dual equation*

$$\begin{cases} \dot{z}(t) = \mathcal{A}^*(T-t)z(t), \\ z(s) = z_0, \end{cases} \qquad s \leqq t \leqq T. \tag{1.7}$$

Then $z(t) = \mathcal{U}^*(T-s, T-t)z_0$ is a weak solution of (1.7) *in the sense that*

(a) *$z(t)$ is weakly continuous on $[s, T]$,*

(b) *$z(t)$ satisfies*

$$\int_s^T \langle z(t), \psi'(t) + \mathcal{A}(T-t)\psi(t)\rangle \, dt = \langle z(T), \psi(T)\rangle - \langle z(s), \psi(s)\rangle$$

for all $\mathcal{D}_A$-valued $\psi(t)$ functions such that ψ, ψ' and $\mathcal{A}(T-t)\psi(t)$ are weakly continuous on (s, T).

Proof.

$$\begin{aligned} &\int_s^T \langle \mathcal{U}^*(T-s, T-t)z_0, \psi'(t) + \mathcal{A}(T-t)\psi(t)\rangle \, dt \\ &\quad = \int_s^T \langle z(s), \mathcal{U}(T-s, T-t)\psi'(t) \\ &\qquad + \mathcal{U}(T-s, T-t)\mathcal{A}(T-t)\psi(t)\rangle \, dt \\ &\quad = \int_s^T \frac{\partial}{\partial t}\langle z_0, \mathcal{U}(T-s, T-t)\psi(t)\rangle \, dt \end{aligned}$$

since $\mathcal{U}$ is a quasi-evolution operator and $\langle z_0, \mathcal{U}(T-s, T-t)\psi(t)\rangle$ is absolutely continuous and equals

$$\langle z(T), \psi(T)\rangle - \langle z_0, \psi(s)\rangle.$$

So $z(t)$ is a weak solution.

We remark that in order to obtain a differential form of the Riccati equation, Bensoussan, Delfour and Mitter in [2] have imposed extra conditions on $\mathcal{U}^*(t, s)$ and perturbations on $\mathcal{U}^*(t, s)$, which amount to requiring that $\mathcal{U}(t, s)$ and its perturbations satisfy a stronger version of (1.6). It seems more natural to express these conditions in terms of properties of the original evolution operators $\mathcal{U}(t, s)$, i.e., by requiring $\mathcal{U}(t, s)$ to be a quasi-evolution operator.

The following definition of a strong evolution operator has been used in [4], [5], [12] and [18].

DEFINITION 1.3 (Strong evolution operator). A *strong evolution operator* is a mild evolution operator with an associated generator $\mathcal{A}(t)$, a closed, densely-defined linear operator $\mathcal{A}(t)$ on H for each $t \in T$, such that

$$\mathcal{U}(t, s) : \mathcal{D}(\mathcal{A}(s)) \to \mathcal{D}(\mathcal{A}(t)) \quad \text{for } t > s, \tag{1.8}$$

$$\frac{\partial}{\partial t}(\mathcal{U}(t, s)x) = \mathcal{A}(t)\mathcal{U}(t, s)x \quad \text{for } x \in \mathcal{D}(\mathcal{A}(s)), \quad t > s. \tag{1.9}$$

We remark that some authors have not assumed $\mathcal{A}(t)$ to be closed or even densely defined, but in all the applications considered so far, this is always the case. Besides these are necessary assumptions to ensure that a strong evolution operator is also a quasi-evolution operator.

LEMMA 1.2. *$\mathcal{U}(t, s)$ is a strong evolution operator on $\Delta(T)$. Then $\mathcal{U}(t, s)$ is also a quasi-evolution operator with the same generator and $\mathcal{D}_A = \bigcap_{r\in[s,t]}\mathcal{D}(\mathcal{A}(r))$ provided $\langle \mathcal{U}(t, r)\mathcal{A}(r)x, y\rangle$ is integrable with respect to r on (s, t) for all $y \in H$ and $x \in \mathcal{D}_A$.*

Proof. Now for $h > 0$,

$$\langle \mathcal{U}(t, r+h)x - \mathcal{U}(t, r)x, y\rangle = \langle \mathcal{U}(t, r+h)(I - \mathcal{U}(r+h, r))x, y\rangle \quad \text{(by (1.1))}$$

$$\to -\langle \mathcal{A}(t, r)\mathcal{A}(r)x, y\rangle$$

as $h \to 0+$ for $x \in \mathcal{D}(\mathcal{A}(r))$ by (1.9) and (1.2). So

$$\frac{\partial^+}{\partial r}\langle \mathcal{U}(t, r)x, y\rangle = -\langle \mathcal{U}(t, r)\mathcal{A}(r)x, y\rangle \quad \text{for } y \in H, \quad x \in \mathcal{D}(\mathcal{A}(r)).$$

Since $\langle y, \mathcal{U}(t, r)x\rangle$ is right differentiable and its derivative is integrable, it is absolutely continuous and

$$\int_s^t \langle y, \mathcal{U}(t, r)\mathcal{A}(r)x\rangle\, dr = \langle y, \mathcal{U}(t, r)x - x\rangle \quad \text{for } x \in \bigcap_{s\le r\le t} \mathcal{D}(\mathcal{A}(r));$$

i.e., $\mathcal{U}(t, s)$ is a quasi-evolution operator with generator $\mathcal{A}(t)$ and $\mathcal{D}_A = \bigcap_{s\le r\le t}\mathcal{D}(\mathcal{A}(r))$.

We remark that a sufficient condition for $\langle \mathcal{U}(t, r)\mathcal{A}(r)x, y\rangle$ to be integrable is that it be measurable and $\sup_{r\in T}\|\mathcal{A}(r)x\| < \infty$ for each $x \in \bigcap_{t\in T}\mathcal{D}(\mathcal{A}(t))$ and this is always satisfied in the applications (see § 4).

The following result is useful when we wish to consider a differentiated version of the dual Riccati equation in § 3.

THEOREM 1.3. *If $\mathcal{U}(t, s)$ is a strong evolution operator on $\Delta(T)$, then the "dual" evolution operator $\mathcal{Y}(t,s)=\mathcal{U}^*(T-s, T-t)$ is a quasi-evolution operator on $\Delta(T)$ with generator $\mathcal{A}^*(T-s)$, provided $\|\mathcal{U}^*(t, r)\mathcal{A}^*(T-r)x\|$ is integrable in r on (s, t) for $x \in \bigcap_{s\leqq r\leqq t}\mathcal{D}(\mathcal{A}^*(T-r))=\mathcal{D}_{A^*}\neq\varnothing$.*

Proof.

$$\frac{\partial}{\partial r}\langle\mathcal{Y}^*(T-t, T-r)y, x\rangle=\langle\mathcal{A}(r)\mathcal{Y}^*(T-t, T-r)y, x\rangle,$$

$$r>t, \quad x\in H, \quad y\in\mathcal{D}(\mathcal{A}(t))$$

(from (1.8)),

$$\therefore \quad \frac{\partial}{\partial r}\langle y, \mathcal{Y}(T-t, T-r)x\rangle=\langle y, \mathcal{Y}(T-t, T-r)\mathcal{A}^*(r)x\rangle \quad \text{for } x\in\mathcal{D}(\mathcal{A}^*(r)),$$

$$\therefore \quad \frac{\partial}{\partial r}\langle y, \mathcal{Y}(t, r)x\rangle=-\langle y, \mathcal{Y}(t, r)\mathcal{A}^*(T-r)x\rangle \quad \text{for } t>r, \quad x\in\mathcal{D}(\mathcal{A}^*(T-r))$$

and by our integrability assumptions, we have

$$\langle y, \mathcal{Y}(t, s)x-x\rangle=\int_s^t\langle y, \mathcal{Y}(t, r)\mathcal{A}^*(T-r)x\rangle\, dr$$

for $x\in\mathcal{D}_{A^*}$ and $y\in\mathcal{D}(\mathcal{A}(t))$. Since $\mathcal{D}(\mathcal{A}(t))$ is dense in H, it holds for all $y\in H$.

Unfortunately, strong evolution operators are not stable under perturbations $D\in\mathcal{B}_\infty(T; \mathcal{L}(H))$ as is shown by the counter example in Phillips [15]. In fact, the best general perturbation result from [15] is for $D\in C^1(T; \mathcal{L}(H))$—the space of strongly continuous differentiable $\mathcal{L}(H)$-valued operators, where it is proved that $\mathcal{A}+D(t)$ generates a strong evolution operator.

However, from Corollary 1.3, $\mathcal{A}+D(t)$ does generate a quasi-evolution operator for any $D\in\mathcal{B}_\infty(T; \mathcal{L}(H))$. We have similar perturbation results for strong evolution operators from Lemma 1.2 and a converse result.

COROLLARY 1.4. *If any perturbation $D\in\mathcal{B}_\infty(T; \mathcal{L}(H))$ of a mild evolution operator $\mathcal{U}(t, s)$ generates a strong evolution operator on $\Delta(T)$, such that $\langle\mathcal{U}_D(t, r)\mathcal{A}(r)x, y\rangle$ is integrable with respect to r on (s, t), for all $y\in H$, $x\in\bigcap_{s\leqq r\leqq t}\mathcal{D}(\mathcal{A}(r))$, then $\mathcal{U}(t, s)$ is a quasi-evolution operator.*

COROLLARY 1.5. *If any perturbation $D\in\mathcal{B}_\infty(T; \mathcal{L}(H))$ of a mild evolution operator $\mathcal{U}(t, s)$ generates a strong evolution operator, such that $\|\mathcal{U}_D^*(t, r)\mathcal{A}^*(T-r)x\|$ is integrable in r on (s, t) for all $x\in\mathcal{D}_{A^*}\neq\varnothing$, then $\mathcal{Y}(t, s)=\mathcal{U}^*(T-s, T-t)$ is also a quasi-evolution operator.*

Proof. See Theorem 1.3.

2. Infinite-dimensional quadratic cost control problem and the Riccati equation. The infinite-dimensional control system we consider is

$$(2.1) \qquad z(t)=\mathcal{U}(t, t_0)z_0+\int_{t_0}^t\mathcal{U}(t, s)B(s)u(s)\, ds,^1 \qquad 0\leqq t_0<t<T<\infty,$$

where $\mathcal{U}(\cdot, \cdot)$ is a mild evolution operator on a real Hilbert space H, $u\in$

[1] See footnote 3.

$L_2(T; K)$, where K is a real Hilbert space and $T=[0, T]$, $z_0 \in H$, and $B \in \mathcal{B}_\infty(T; \mathcal{L}(H))$.[2]

We consider the cost functional:

$$\mathcal{C}(u; t_0, z_0) = \langle z(T), Gz(T)\rangle + \int_{t_0}^{T} (\langle z(s), W(s)z(s)\rangle + \langle u(s), R(s)u(s)\rangle)\, ds, \tag{2.2}$$

where $z(t)$ is given by (2.1), $G \in \mathcal{L}(H)$ is self-adjoint and nonnegative, $W \in \mathcal{B}_\infty(T; \mathcal{L}(H))$, $R \in \mathcal{B}_\infty(T; \mathcal{L}(H))$ and for each t, $W(t)$, $R(t)$ are nonnegative and self-adjoint and $R(t)$ satisfies

$$\langle y, R(t)y\rangle \geqq \mu \|y\|^2 \quad \text{a.e. for some } \mu > 0.$$

The quadratic cost control problem is then to find the optimal control $u_0 \in L_2(T; K)$ which minimizes $\mathcal{C}(u; t_0, z_0)$.

Consider a sequence of admissible controls $\{u_k\}$ given by

$$u_k(t) = -F_k(t)z(t), \tag{2.3}$$

i.e.,

$$z(t) = \mathcal{U}_k(t, t_0)z_0,$$

where $\mathcal{U}_k(\cdot, \cdot)$ is the perturbed mild evolution operator corresponding to the perturbation of $\mathcal{U}(\cdot, \cdot)$ by $-B(t)F_k(t)$, and $F_k(t)$ is defined recursively by

$$\begin{aligned} &F_k(t) = R^{-1}(t)B^*(t)Q_{k-1}(t); \quad F_0(t) = 0, \\ &W_k(t) = W(t) + F_k^*(t)R(t)F_k(t), \\ &Q_k(t)x = \mathcal{U}_k^*(T, t)G\mathcal{U}_k(T, t)x + \int_{t_0}^{T} \mathcal{U}_k^*(s, t)W_k(s)\mathcal{U}_k(s, t)x\, ds.^{3} \end{aligned} \tag{2.4}$$

F_k, W_k, Q_k are all bounded linear operators and are uniformly bounded in norm on T, $Q_k \in \mathcal{B}_\infty(T; \mathcal{L}(H))$. Consider the sequence of control problems

$$z_k(t) = \mathcal{U}_k(t, t_0)z_0 + \int_{t_0}^{t} \mathcal{U}_k(t, s)B(s)\bar{u}(s)\, ds, \tag{2.5}$$

where $\bar{u} \in L_2(T; K)$.

A basic technical lemma, which is easily verified by substitution, is

LEMMA 2.1.

$$\begin{aligned} &\langle z_k(t), Q_k(t)z_k(t)\rangle \\ &\quad = \langle z_k(T), Gz_k(T)\rangle + \int_t^T (\langle z_k(s), W_k(s)z_k(s)\rangle - \langle z_k(s), R(s)B(s)\bar{u}(s)\rangle \\ &\qquad\qquad - \langle Q_k(s)B(s)\bar{u}(s), z_k(s)\rangle)\, ds. \end{aligned}$$

[2] $\mathcal{B}_\infty(T; \mathcal{L}(H))$ is the space of $\mathcal{L}(H)$-valued functions which are strongly measurable and uniformly bounded in norm on T.

[3] A well-defined Bochner integral (see Property A.3 and Lemma A.1).

LEMMA 2.2. $\langle z_0, Q_k(t_0)z_0\rangle$ *is monotonically decreasing in* k *for each* $t_0 \in T$ *and all* $z_0 \in H$, *with*

$$\langle z_0, Q_k(t_0)z_0\rangle \leqq \langle z_0, Q_0(t_0)z_0\rangle.$$

Proof. Letting $\bar{u} = 0$ in Lemma 2.1, we have

$$\begin{aligned}\langle z_k(t_0), Q_k(t_0)z_k(t_0)\rangle &= \langle z_k(T), Gz_k(T)\rangle + \int_{t_0}^{T} \langle z_k(s), W_k(s)z_k(s)\rangle\, ds \\ &= \langle z_k(T), Gz_k(T)\rangle \\ &\quad + \int_{t_0}^{T} (\langle z_k(s), W(s)z_k(s)\rangle \\ &\qquad + \langle F_k(s)z_k(s), R(s)F_k(s)z_k(s)\rangle)\, ds \\ &= \langle z_k(T), Gz_k(T)\rangle \\ &\quad + \int_{t_0}^{T} (\langle z_k(s), W(s)z_k(s)\rangle + \langle u_k(s), R(s)u_k(s)\rangle)\, ds,\end{aligned}$$

and this is just the cost for (2.1) with feedback control $u_k(t) = -F_k(t)z(t)$,

$$\therefore \quad \mathscr{C}(u_k; t_0, z_0) = \langle z_0, Q_k(t_0)z_0\rangle.$$

For simplicity, let $t_0 = 0$ and show that $\mathscr{C}(u_k; 0, z_0) \geqq \mathscr{C}(u_{k+1}; 0, z_0)$. Consider (2.5) where $\bar{u}(t) = u_0(t) + F_k(t)z_k(t)$, i.e., (2.5) is equivalent to

$$z_k(t) = \mathscr{U}_k(t, 0)z_o + \int_0^t \mathscr{U}_k(t, s)B(s)u_0(s)\, ds.$$

If $u_0(t) = -F_{k+1}(t)z_k(t)$, (2.5) is also equivalent to

$$z_k(t) = \mathscr{U}_{k+1}(t, 0)z_0,$$

i.e., for $\bar{u}(t) = u_0(t) + F_k(t)z_k(t)$, $z_k(t)$ and $z_{k+1}(t)$ are identical and so we shall dispense with the suffix henceforth. Substituting for this $\bar{u}(t)$ in Lemma 2.1 with $t = 0$, we obtain

$$\begin{aligned}\langle z_0, Q_k(0)z_0\rangle = \langle z(T), Gz(T)\rangle &+ \int_0^T (\langle z(s), W(s)z(s)\rangle \\ &\quad + \langle F_k(s)z(s), R(s)F_k(s)z(s)\rangle)\, ds \\ &- \int_0^T (\langle z(s), Q_k(s)B(s)u_{k+1}(s)\rangle + \langle z(s), Q_k(s)B(s)F_k(s)z(s)\rangle)\, ds \\ &- \int_0^T (\langle Q_k(s)B(s)F_k(s)z(s), z(s)\rangle + \langle Q_k(s)B(s)u_{k+1}(s), z(s)\rangle)\, ds\end{aligned}$$

and

$$B^*(t)Q_k(t) = R(t)F_{k+1}(t); \qquad u_{k+1}(t) = -F_{k+1}(t)z(t),$$

$$\begin{aligned}\therefore \quad \langle z_0, Q_k(0)z_0\rangle &= \langle z(T), Gz(T)\rangle \\ &\quad + \int_0^T [\langle z(s), W(s)z(s)\rangle + 2\langle u_{k+1}(s), R(s)u_{k+1}(s)\rangle]\, ds \\ &\quad + \int_0^T [\langle F_{k+1}(s)z(s), R(s)F_k(s)z(s)\rangle \\ &\quad - \langle B^*(s)Q_k(s)z(s), F_k(s)z(s)\rangle \\ &\quad - \langle F_k(s)z(s), B^*(s)Q_k(s)z(s)\rangle]\, ds \\ &= \langle z(T), Gz(T)\rangle + \int_0^T \langle B^*(s)Q_k(s)z(s), (R^{-1}(s)B^*(s)Q_k(s) \\ &\quad - F_k(s))z(s)\rangle\, ds + \int_0^T \langle F_k(s)z(s), (R(s)F_k(s) \\ &\quad - B^*(s)Q_k(s))z(s)\rangle\, ds \\ &= \langle z(T), Gz(T)\rangle - \int_0^T \langle (R(s)B^*(s)Q_k(s) - F_k(s))z(s), (B^*(s) \\ &\quad \cdot Q_k(s) - R(s)F_k(s))z(s)\rangle\, ds,\end{aligned}$$

i.e.,

$$\langle z_0, Q_k(0)z_0\rangle = \langle z_0, Q_{k+1}(0)z_0\rangle - \int_0^T \langle R(s)y(s), y(s)\rangle\, ds,$$

where

$$y(s) = B^*(s)Q_k(s) - R(s)F_k(s)$$

and R is strictly positive.

$$\therefore \quad \langle z_0, Q_k(0)z_0\rangle \geqq \langle z_0, Q_{k+1}(0)z_0\rangle.$$

The proof for $t_0 \neq 0$ is similar so

$$\langle z_0, Q_k(t_0)z_0\rangle \geqq \langle z_0, Q_{k+1}(t_0)z_0\rangle \quad \text{for each } t_0 \in T, \quad z_0 \in H.$$

THEOREM 2.1. *$Q_k(t)$ of (2.4) converges strongly to a self-adjoint operator $Q \in \mathcal{B}_\infty(T; \mathcal{L}(H))$ which satisfies the integral equation*

$$\begin{aligned}Q(t)x &= \mathcal{U}_\infty^*(T, t)G\mathcal{U}_\infty(T, t)x \\ &\quad + \int_t^T \mathcal{U}_\infty^*(s, t)[W(s) + Q(s)B(s)R^{-1}(s)B^*(s)Q(s)]\mathcal{U}_\infty(s, t)x\, ds,\end{aligned} \tag{2.6}$$

where $\mathcal{U}_\infty(t, s)$ is the perturbed mild evolution operator corresponding to the perturbation of $\mathcal{U}(t, s)$ by $-B(t)R^{-1}(t)B^*(t)Q(t)$.

Proof. $Q_k(t)$ is a sequence of positive self-adjoint operators on H and so

$$\|Q_k(t)\| = \sup_{\|z_0\|=1} \langle z_0, Q_k(t)z_0\rangle$$

$$\leqq \sup_{\|z_0\|=1} \langle z_0, Q_0(t)z_0\rangle \quad \text{(by Lemma 2.2)}$$

$$= C.$$

So for each t, $Q_k(t)$ converges strongly to a self-adjoint positive operator $Q(t)$ and $\|Q(t)\| \leqq C$ on T. Therefore $W_k(t), F_k(t)$ also converge strongly to bounded operators $W_\infty(t)$ and $F_\infty(t)$, respectively, and W_∞, F_∞ are uniformly bounded in norm on T, and by Theorem 1.1, $\operatorname{ess\,sup}_{\Delta(T)}\|\mathcal{U}_k(t, s)\| \leqq M_2 \exp M_2\mu\beta^2 CT$, where

$$\operatorname*{ess\,sup}_{\Delta(T)}\|\mathcal{U}(t, s)\| = M_2, \qquad \operatorname*{ess\,sup}_{t\in T}\|B(t)\| = \beta.$$

$\mathcal{U}_k(t, s)$ is the perturbed mild evolution operator corresponding to $-B(t)F_k(t)$ and $\mathcal{U}_\infty(t, s)$ is the perturbed mild evolution operator corresponding to $-B(t)F_\infty(t)$.

That $\mathcal{U}_k(t, s) \to \mathcal{U}_\infty(t, s)$ strongly as $k \to \infty$ follows from the definition of $\mathcal{U}_k(t, s)$ as the unique solution of

$$\mathcal{U}_k(t, s)x = \mathcal{U}(t, s)x + \int_s^t \mathcal{U}(t, \alpha)B(\alpha)F_k(\alpha)\mathcal{U}_k(\alpha, s)x\, d\alpha$$

for all $x \in H$ and, by applying the Lebesgue dominated convergence theorem. (All operators are uniformly bounded in norm.)

Now $Q_k(t)$ satisfied (2.4) for each k and W_k, Q_k, F_k are all uniformly bounded in norm on T and all converge strongly as $k \to \infty$ to W_∞, F_∞ and Q respectively. So applying the Lebesque dominated convergence theorem, we have that

$$Q(t)x = \mathcal{U}_\infty^*(T, t)G\mathcal{U}_\infty(T, t)x + \int_t^T \mathcal{U}_\infty^*(s, t)(W(s) + Q(s)B(s)R^{-1}(s)B^*(s)Q(s))\mathcal{U}_\infty(s, t)x\, ds.$$

Therefore $Q \in \mathcal{B}_\infty(T; \mathcal{L}(H))$, and is self-adjoint.

THEOREM 2.2. *The optimal control which minimizes $\mathcal{C}(u; t_0, z_0)$ is the feedback control*

$$u_0(t) = -R^{-1}(t)B^*(t)Q(t)z(t).$$

Proof. Consider any admissible control u, so that

$$z(t) = \mathcal{U}(t, t_0)z_0 + \int_{t_0}^t \mathcal{U}(t, s)B(s)u(s)\, ds.$$

Then since $\mathcal{U}_\infty(t, t_0)$ is a perturbation of $\mathcal{U}(t, t_0)$, by $-R^{-1}(t)B^*(t)Q(t)$ we have

$$z(t) = \mathcal{U}_\infty(t, t_0)z_0 + \int_{t_0}^t \mathcal{U}_\infty(t, s)B(s)\bar{u}(s)\, ds,$$

where

$$\bar{u}(t) = u(t) + R^{-1}(t)B^*(t)Q(t)z(t),$$

$$\begin{aligned}\langle z_0, Q(t_0)z_0\rangle &= \langle z(T), Gz(T)\rangle \\ &\quad + \int_{t_0}^{T} (\langle z(s), W(s)z(s)\rangle \\ &\qquad + \langle R^{-1}(s)B^*(s)Q(s)z(s), B^*(s)Q(s)z(s)\rangle)\, ds \\ &\quad - \int_{t_0}^{T} (\langle z(s), Q(s)B(s)\bar{u}(s)\rangle + \langle Q(s)B(s)\bar{u}(s), z(s)\rangle)\, ds \\ &\qquad\qquad \text{(by Lemma 2.1)} \\ &= \langle z(T), Gz(T)\rangle + \int_{t_0}^{T} (\langle z(s), W(s)z(s)\rangle + \langle u(s), R(s)u(s)\rangle)\, ds \\ &\qquad - \int_{t_0}^{T} \langle \bar{u}(s), R(s)\bar{u}(s)\rangle\, ds,\end{aligned}$$

$$\begin{aligned}\therefore \quad \mathscr{C}(u; t_0, z_0) &= \mathscr{C}(u_0; t_0, z_0) + \int_{t_0}^{T} \langle \bar{u}(s), R(s)\bar{u}(s)\rangle\, ds \\ &\geqq \mathscr{C}(u_0; t_0, z_0)\end{aligned}$$

and so u_0 is optimal.

THEOREM 2.3. *$Q(t)$ is the unique solution of* (2.6) *in the class of self-adjoint bounded linear operators in $\mathscr{B}_\infty(T; \mathscr{L}(H))$.*

Proof. Let $Q(t)$ be a solution of the integral equation (2.6), and suppose u is any admissible control in $L_2(T; U)$. Then from the proof of Theorem 2.2, we have

$$\begin{aligned}\mathscr{C}(u; t_0, z_0) = \langle Q(t)z_0, z_0\rangle + \int_{t_0}^{T} \langle u(t) + R^{-1}(t)B^*(t)Q(t)z(t), \\ R(t)(u(t) + R^{-1}(t)B^*(t)Q(t)z(t))\rangle\, ds.\end{aligned}$$

Now by a standard result in [1], $\mathscr{C}(u; t_0, z_0)$ has a unique minimizing control $u_0(\cdot)$, and so $\langle Q(t)z_0, z_0\rangle$ is uniquely defined. Since $Q(t)$ is symmetric, it is unique.

So far we have only assumed that $\mathscr{U}(t, s)$ in (2.1) is a mild evolution operator. In order to obtain a differentiated version of the Riccati equation, we must assume that $\mathscr{U}(t, s)$ is a quasi-evolution operator. We use the following standard technical lemmas.

LEMMA 2.3. *Let $f : (0, T) \times (0, T) \to \mathbb{R}$ be integrable and suppose*

(i) *for almost all s, $f(\cdot, s)$ is absolutely continuous,*

(ii) $f(s, s) \in L_1(0, T)$,

(iii) $\displaystyle\int_0^T ds \int_s^T \left|\frac{\partial f}{\partial \tau}(s, \tau)\right| d\tau < \infty.$

Then $g(t) = \int_0^t f(s, t)\, ds$ *is absolutely continuous with*

$$g'(t) = f(t, t) + \int_0^t \frac{\partial f}{\partial t}(s, t)\, ds \quad a.e. \tag{2.7}$$

Proof (Vinter [18]).

$$\int_0^t f(s, t)\, ds = \int_0^t f(s, s)\, ds + \int_0^t \left(\int_0^\sigma \frac{\partial f}{\partial \sigma}(s, \sigma)\, ds \right) d\sigma$$

by applying Fubini's theorem to

$$\int_0^t \int_0^t \frac{\partial f}{\partial \sigma}(s, \sigma) \chi(\sigma, s)\, d\sigma\, ds,$$

where

$$\chi(\sigma, s) = \begin{cases} 1, & \sigma \geqq s, \\ 0 & \text{otherwise.} \end{cases}$$

LEMMA 2.4. *Let* H *be a real Hilbert space and* $W \in \mathcal{L}(H)$. *Suppose* $g_1(\cdot)$ *and* $g_2(\cdot)$ *are weakly absolutely continuous* H*-valued functions on* $[0, T]$ *such that*

$$\langle g_i(t), x \rangle = \langle g_i(s), x \rangle + \int_0^t \frac{\partial}{\partial s} \langle g_i(s), x \rangle\, ds \quad \text{for all } x \in H; \quad i = 1, 2.$$

Then $f(t) = \langle W g_1(t), g_2(t) \rangle$ *is an absolutely continuous function with*

$$\langle W g_1(t), g_2(t) \rangle = \langle W g_1(0), g_2(0) \rangle + \int_0^t \frac{\partial}{\partial s} \langle W g_1(s), g_2(s) \rangle\, ds.$$

Proof (Vinter [18]). Apply Fubini's theorem to

$$\int_0^t \int_0^t \chi(\sigma, s) \frac{\partial^2}{\partial \sigma \partial s} \langle W g_1(\sigma), g_2(s) \rangle\, ds\, d\sigma,$$

where

$$\chi(\sigma, s) = \begin{cases} 1, & \sigma \geqq s, \\ 0 & \text{otherwise,} \end{cases}$$

i.e.,

$$\int_0^t \int_s^t \frac{\partial^2}{\partial \sigma \partial s} \langle W g_1(\sigma), g_2(s) \rangle\, d\sigma\, ds = \int_0^t \int_0^\sigma \frac{\partial^2}{\partial s \partial \sigma} \langle W g_1(\sigma), g_2(s) \rangle\, ds\, d\sigma$$

whence the result follows.

LEMMA 2.5. *Let* H *be a real Hilbert space and suppose* $P(\cdot)$ *is a weakly absolutely continuous* $\mathcal{L}(H)$*-valued function and* $g(\cdot)$ *is strongly differentiable with the representation*

$$g(t) = g'(0) + \int_0^t g'(s)\, ds.$$

Then $P(\cdot)g(\cdot)$ is weakly absolutely continuous with

$$\frac{d}{dt}\langle P(t)g(t), x\rangle = \langle P(t)g'(t), x\rangle + \frac{\partial}{\partial t}\langle P(t)g(s), x\rangle\Big|_{s=t} \quad \text{a.e. on } T.$$

Proof. As in the proof of Lemma 2.4, we apply Fubini's theorem to

$$\int_0^t \int_0^t \langle \chi(\sigma, s)\frac{\partial}{\partial s}(P(s)g'(\sigma)), x\rangle \, d\sigma \, ds,$$

whence

$$\int_0^t \left\langle P(s)g'(s) + \frac{\partial}{\partial s}(P(s)g(\sigma))\Big|_{\sigma=s}, x\right\rangle ds = \langle P(t)g(t) - P(0)g(0), x\rangle$$

and differentiation yields the required result.

THEOREM 2.4. *Let $\mathcal{U}(t, s)$ be a quasi-evolution operator on H. Then the solution of the integral equation* (2.6) *satisfies the following inner product differentiated Riccati equation*:

$$\frac{d}{dt}\langle Q(t)x, y\rangle + \langle Q(t)x, \mathcal{A}(t)y\rangle + \langle \mathcal{A}(t)x, Q(t)y\rangle$$

$$-\langle Q(t)B(t)R^{-1}(t)B^*(t)Q(t)x, y\rangle + \langle W(t)x, y\rangle = 0 \quad \text{a.e. on } [t_0, T], \tag{2.8}$$

$$Q(T) = G \quad \text{for } x, y \in \mathcal{D}_A.$$

If B, W and R are strongly continuous on T, then (2.8) *is satisfied everywhere on* $[t_0, T]$.

Proof. We differentiate the inner product $\langle (2.6)x, y\rangle$ term by term for $x, y \in \mathcal{D}_A$ using property (1.5) for quasi-evolution operators and Lemma 2.4. The formal term by term differentiation is clear and so we shall consider the justification of the "differentiation under the integral" procedure. Consider

$$g(t) = \left\langle \int_t^T \mathcal{U}_\infty^*(s, t)W(s)\mathcal{U}_\infty(s, t)x \, ds, y\right\rangle \quad (\text{for } x, y \in \mathcal{D}_A)$$

$$= \int_t^T \left\langle W(s)\mathcal{U}_\infty(s, t)x, \mathcal{U}_\infty(s, t)y\right\rangle ds,$$

$$\therefore \quad \frac{dg(t)}{dt} = -\langle W(t)x, y\rangle + \int_t^T \frac{\partial}{\partial t}\langle W(s)\mathcal{U}_\infty(s, t)x, \mathcal{U}_\infty(s, t)y\rangle \, ds,$$

and assuming for the moment that we can differentiate under the integral sign,

$$= -\langle W(t)x, y\rangle - \int_t^T \langle W(s)\mathcal{U}_\infty(s, t)(\mathcal{A}(t) - B(t)R^{-1}(t)B^*(t)Q(t))x, \mathcal{U}_\infty(s, t)y\rangle \, ds$$

$$- \int_t^T \langle W(s)\mathcal{U}_\infty(s, t)x, \mathcal{U}_\infty(s, t)$$

$$\cdot (\mathcal{A}(t) - B(t)R^{-1}(t)B^*(t)Q(t))y\rangle \, ds \quad \text{a.e.,}$$

and using property (1.5) and Theorem 1.2,

$$= -\langle W(t)x, y\rangle - \Big\langle (\mathcal{A}(t) - B(t)R^{-1}(t)B^*(t)Q(t))x, \int_t^T \mathcal{U}_\infty^*(s, t)W(s)\mathcal{U}_\infty(s, t)y\, ds\Big\rangle$$

$$- \Big\langle \int_t^T \mathcal{U}_\infty^*(s, t)W(s)\mathcal{U}_\infty(s, t)x\, ds, (\mathcal{A}(t) - B(t)R^{-1}(t)B^*(t)Q(t))y\Big\rangle.$$

Taking $\mathcal{A}(t)$ outside the integral is valid since it is closed.

The differentiation under the integrand is justified since

$$f(s, t) = \langle W(s)\mathcal{U}_\infty(s, t)x, \mathcal{U}_\infty(s, t)y\rangle$$

satisfies the conditions of Lemma 2.3, i.e.,

$$\int^T \left|\frac{\partial}{\partial t} f(s; t)\right| dt \leqq \int_s^T (|\langle \mathcal{U}_\infty(s, t)\mathcal{A}_2(t)x, W(s)\mathcal{U}_\infty(s, t)y\rangle| + |\langle W(s)\mathcal{U}_\infty(s, t)x, \mathcal{U}_\infty(s, t)\mathcal{A}_2(t)y\rangle|)\, dt,$$

where $\mathcal{A}_2(t) = \mathcal{A}(t) - B(t)R^{-1}(t)B^*(t)Q(t)$, and the integral is finite since $\mathcal{U}_\infty(t, s)$ is a quasi-evolution operator with generator $\mathcal{A}_2(t)$ satisfying (1.5) and by Lemma 2.4, $f(s, t)$ is absolutely continuous in t.

COROLLARY 2.1. *Consider* (2.6) *where* $\mathcal{U}(t, s)$ *is the quasi-perturbed evolution operator generated by* $\mathcal{A} + D(t)$ *where* $D \in \mathcal{B}_\infty(T; \mathcal{L}(H))$ *and* $\mathcal{A}$ *is the infinitesimal generator of a strongly continuous semigroup. Then* (2.6) *has a unique strongly continuous solution which also satisfies the differentiated Riccati equation* (2.8).

Remark 1. It is possible to obtain another form for the differentiated version of the Riccati equation, using only the assumption that $\mathcal{U}(t, s)$ is a mild evolution operator. First we observe:

$$\mathcal{U}_\infty^*(t, s)Q(t)\mathcal{U}_\infty(t, s)x = \mathcal{U}_\infty^*(T, s)G\mathcal{U}_\infty(T, s)x + \int_t^T \mathcal{U}_\infty^*(r, s) \cdot [W(r) + Q(r)B(r)R^{-1}(r)B^*(r)Q(r)]\mathcal{U}_\infty(r, s)x\, dr$$

$$= Q(s)x - \int_s^t \mathcal{U}_\infty^*(r, s) \cdot [W(r) + Q(r)B(r)R^{-1}(r)B^*(r)Q(r)]\mathcal{U}_\infty(r, s)x\, dr, \quad t > s.$$

Now the RHS is a well-defined Bochner integral. Hence

$$\frac{\partial}{\partial t}(\mathcal{U}_\infty^*(t, s)Q(t)\mathcal{U}_\infty(t, s)x) = -\mathcal{U}_\infty^*(t, s) \cdot [W(t) + Q(t)B(t)R^{-1}(t)B^*(t)Q(t)]\mathcal{U}_\infty(t, s)x \quad \text{a.e.}$$

Remark 2. There are several other sets of sufficient conditions to allow for the differentiation of (2.6), which were set up either for the partial differential equation in [2], [5], [8], [13], [16] or for the delay case in [7], [9]. The advantage of Theorem 2.4 is that it is directly applicable to both the delay case and the partial differential equation case as we shall prove in § 4. It is also directly applicable to the dual Riccati equation for the partial differential equation case, which has also been studied in [1] and [3].

The question of uniqueness of solutions of (2.8) is not really important for control applications, since we have given a means of constructing a solution to the

integral Riccati equation, and we know that any other solution to the differential Riccati equation will result in a larger value for the cost. Nevertheless, we feel it is worth including a proof of uniqueness; however it requires much stronger conditions on the evolution operator.

THEOREM 2.5. *Let $\mathcal{U}(t, s)$ be a strong evolution operator with generator $\mathcal{A}(t)$ such that $\langle \mathcal{U}(t, r)\mathcal{A}(r)x, y\rangle$ is integrable with respect to r on (s, t) for all $y \in H$ and $x \in \mathcal{D}_A$. Then if $\bar{\mathcal{D}}_A = H$, (2.8) has a unique solution in the class of self-adjoint weakly continuous operators $P(\cdot)$, such that $\langle x, P(\cdot)y\rangle$ is absolutely continuous for all $x, y \in \mathcal{D}_A$.*

Proof.

(a) Let P_1, P_2 be solutions of (2.8) and write $Q(t) = P_1(t) - P_2(t)$. Then it is readily verified that

$$\begin{aligned}\frac{d}{dt}\langle Q(t)x, y\rangle = &-\langle (\mathcal{A}(t) - C(t)P_1(t))x, Q(t)y\rangle \\ &-\langle Q(t)x, (\mathcal{A}(t) - C(t)P_1(t))y\rangle \\ &-\langle Q(t)C(t)Q(t)x, y\rangle \quad \text{a.e.}\end{aligned} \tag{2.9}$$

and

$$\begin{aligned}\frac{d}{dt}\langle Q(t)x, y\rangle = &-\langle (\mathcal{A}(t) - C(t)P_2(t))x, Q(t)y\rangle - \langle Q(t)x, (\mathcal{A}(t) - C(t)P_2(t))y\rangle \\ &+\langle Q(t)C(t)Q(t)x, y\rangle \quad \text{a.e.},\end{aligned} \tag{2.10}$$

where

$$C(t) = B(t)R^{-1}(t)B^*(t).$$

Let

$$F(t)x = \int_t^T \mathcal{U}_1^*(s, t)Q(s)C(s)Q(s)\mathcal{U}_1(s, t)x\, ds,$$

where $\mathcal{U}_1(t, s)$ is the quasi-perturbed operator generated by $\mathcal{A}(t) - C(t)P_1(t)$. Then for $x, y \in \mathcal{D}_A$, by Lemmas 2.3, 2.4, we may differentiate to obtain

$$\begin{aligned}\frac{d}{dt}\langle F(t)x, y\rangle = &-\langle Q(t)C(t)Q(t)x, y\rangle - \langle F(t)x, (\mathcal{A}(t) - C(t)P_1(t))y\rangle \\ &-\langle \mathcal{A}(t) - C(t)P_1(t)x, F(t)y\rangle \quad \text{a.e.}\end{aligned}$$

and subtracting from (2.9), we have

$$\begin{aligned}\frac{d}{dt}\langle (Q(t) - F(t))x, y\rangle = &-\langle (Q(t) - F(t))x, (\mathcal{A}(t) - C(t)P_1(t))y\rangle \\ &-\langle (\mathcal{A}(t) - C(t)P_1(t))x, (Q(t) - F(t))x\rangle \quad \text{a.e.}, \\ Q(T) = F(T) = 0.\end{aligned} \tag{2.11}$$

Assuming that (2.11) has a unique solution, we have

$$Q(t) = F(t)$$

and

$$\langle Q(t)x, x\rangle = \int_t^T \langle \mathcal{U}_1^*(s, t)Q(s)C(s)Q^*(s)\mathcal{U}_1(s, t)x, x\rangle\, ds \geqq 0 \quad \text{for } x \in H.$$

Similarly, using (2.10) with P_2 perturbations, we find

$$\langle Q(t)x, x\rangle \leqq 0 \quad \text{for } x \in H,$$

$$\therefore \quad Q(t) = 0.$$

(b) Consider the linear equation on H:

$$\frac{d}{dt}\langle P(t)x, y\rangle = -\langle P(t)x, (\mathcal{A}(t) - D(t))y\rangle - \langle (\mathcal{A}(t) - D(t))x, P(t)y\rangle \quad \text{a.e.},$$

$$P(T) = 0, \tag{2.12}$$

where

$$D \in \mathcal{B}_\infty(T; \mathcal{L}(H)).$$

Let $Q(t) = \mathcal{U}^*(t, s)P(t)\mathcal{U}(t, s)$, where $\mathcal{U}(t, s)$ is a strong evolution operator and so $\mathcal{U}(t, s)x$ is strongly differentiable in t for $x \in \mathcal{D}_A$. By Lemmas 2.4 and 2.5, $\langle P(t)\mathcal{U}(t, s)x, \mathcal{U}(t, s)y\rangle$ is absolutely continuous and

$$\begin{aligned}\frac{d}{dt}\langle x, Q(t)y\rangle &= \langle P(t)\mathcal{A}(t)\mathcal{U}(t, s)x, \mathcal{U}(t, s)y\rangle + \langle P(t)\mathcal{U}(t, s)x, \mathcal{A}(t)\mathcal{U}(t, s)y\rangle \\ &\quad - \langle P(t)\mathcal{U}(t, s)x, (\mathcal{A}(t) - D(t))\mathcal{U}(t, s)y\rangle \\ &\quad - \langle (\mathcal{A}(t) - D(t))\mathcal{U}(t, s)x, P(t)\mathcal{U}(t, s)y\rangle \quad \text{a.e.} \\ &= \langle P(t)\mathcal{U}(t, s)x, D(t)\mathcal{U}(t, s)y\rangle + \langle D(t)\mathcal{U}(t, s)x, P(t)\mathcal{U}(t, s)y\rangle \quad \text{a.e.}\end{aligned}$$

and

$$\langle \mathcal{U}^*(t, s)P(t)\mathcal{U}(t, s)x, x\rangle = -\int_t^T \langle (D^*(r)P(r) + P(r)D(r))\mathcal{U}(r, s)x, \mathcal{U}(r, s)x\rangle\, dr$$

for all $x \in H$, since $\bar{\mathcal{D}}_A = H$.

Let $s \to t$. Then

$$\langle P(t)x, x\rangle = -\int_t^T \langle (D^*(r)P(r)P(r)D(r))\mathcal{U}(r, t)x, \mathcal{U}(r, t)x\rangle\, dr \quad \text{for } x \in H.$$

Now $P(t)$ is self-adjoint and so

$$\|P(t)\| = \sup_{\|x\|=1} |\langle P(t)x, x\rangle| \leqq \sup_{\|x\|=1} \int_t^T C\|P(r)\|\,\|x\|^2\, dr,$$

$$\therefore \quad \|P(t)\| \leqq C\int_t^T \|P(r)\|\, dr.$$

Then by Gronwall's inequality, $\|P(t)\| = 0$ on H, i.e., (2.12) has a unique solution on H.

COROLLARY 2.2. *Let $\mathcal{U}(t, s)$ be a quasi-evolution operator and for some $F \in \mathcal{B}_\infty(T; \mathcal{L}(H))$ the perturbed evolution operator $\mathcal{U}_F(t, s)$ is a strong evolution operator. Then the conclusions of Theorem* 2.5 *hold.*

Proof.

(a) Assume only that $\mathcal{U}(t, s)$ is a quasi-evolution operator.

(b) Consider (2.12) on $\mathcal{D}_A$ a quasi-evolution operator with generator $\mathcal{A}(t)$ and $\mathcal{U}_F(t, s)$ a strong evolution operator with generator $\mathcal{A}(t)+F(t)$.

Let $Q(t) = \mathcal{U}_F^*(t, s)P(t)\mathcal{U}_F(t, s)$. Then $\langle Q(t)x, y\rangle$ is absolutely continuous and

$$\begin{aligned}\frac{d}{dt}\langle x, Q(t)y\rangle &= \langle P(t)(\mathcal{A}(t)+F(t))\mathcal{U}_F(t, s)x, \mathcal{U}_F(t, s)y\rangle \\ &\quad +\langle P(t)\mathcal{U}_F(t, s)x, (\mathcal{A}(t)+F(t))\mathcal{U}_F(t, s)y\rangle \\ &\quad -\langle P(t)\mathcal{U}_F(t, s)x, (\mathcal{A}(t)-D(t))\mathcal{U}_F(t, s)y\rangle \\ &\quad +\langle (\mathcal{A}(t)-D(t))\mathcal{U}_F(t, s)x, P(t)\mathcal{U}_F(t, s)y\rangle \quad \text{a.e.} \\ &= \langle P(t)\mathcal{U}_F(t, s)x, (D(t)+F(t))\mathcal{U}_F(t, s)y\rangle + \langle (D(t)+F(t))\mathcal{U}_F(t, s)x, \\ &\quad \cdot P(t)\mathcal{U}_F(t, s)y\rangle \quad \text{a.e.},\end{aligned}$$

but this implies that $Q(t) = 0$ as before, since $D+F \in \mathcal{B}_\infty(T; \mathcal{L}(H))$. So (2.12) again has a unique solution on H.

COROLLARY 2.3. *If $\mathcal{U}(t, s)$ is the quasi-evolution operator generated by $\mathcal{A}+D(t)$, where $D \in \mathcal{B}_\infty(T; \mathcal{L}(H))$ and $\mathcal{A}$ is the infinitesimal generator of a strongly continuous semigroup, then* (2.8) *has a unique solution in the class of strongly continuous operators on H, such that $\langle x, P(t)y\rangle$ is absolutely continuous for $x, y \in \mathcal{D}(\mathcal{A})$.*

3. Dual Riccati equation. Just as in finite dimensions there is a duality between the quadratic cost control problem and the linear filtering problem, this occurs in infinite dimensions via the dual Riccati equation. This means considering the existence and uniqueness of solutions for the following Riccati equations:

$$\begin{aligned}P(t)x = \mathcal{U}_p(t, t_0)P_0\mathcal{U}_p^*(t, t_0)x + \int_{t_0}^{t} \mathcal{U}_p(t, s)[W(s) \\ +P(s)C(s)P(s)]\mathcal{U}_p(t, s)x\, ds,\end{aligned} \tag{3.1}$$

where $\mathcal{U}_p(t, s)$ is a perturbed mild evolution operator corresponding to the perturbation $-P(t)C(t)$ of the mild evolution operator $\mathcal{U}(t, s)$. P_0, $C(t)$ and $W(t)$ are bounded linear operators on H, self-adjoint and positive and $W, C \in \mathcal{B}_\infty(T; \mathcal{L}(H))$.

$$\begin{aligned}\frac{d}{dt}\langle x, P(t)y\rangle - \langle P(t)x, \mathcal{A}^*(t)y\rangle - \langle \mathcal{A}^*(t)x, P(t)y\rangle \\ -\langle W(t)x, y\rangle + \langle P(t)C(t)P(t)x, y\rangle = 0 \quad \text{a.e. on } T\end{aligned} \tag{3.2}$$

for $x, y \in \bigcap_{t\in T}\mathcal{D}(\mathcal{A}^*(t))$, where W, R, C, P_0, are shown as above and this time $\mathcal{U}(t, s)$ is a strong evolution operator with generator $\mathcal{A}(t)$.

THEOREM 3.1. *Under the stated assumptions for* (3.1), (3.1) *has a unique self-adjoint positive operator solution in $\mathcal{B}_\infty(T; \mathcal{L}(H))$.*

Proof. Let $\mathcal{Y}_p(t, s) = \mathcal{U}_p^*(T-s, T-t)$ (that is, $\mathcal{Y}_p(t, s)$ is the *dual* to $\mathcal{U}_p(t, s)$). Then $\mathcal{Y}_p(t, s)$ is a mild evolution operator on $\Delta(T)$ and Theorem 2.3 holds.

From Theorem 2.4 on the differentiated Riccati equation we see that in order to differentiate the inner product version of (3.1), we require $\mathcal{Y}_p(t, s)$ to be a quasi-evolution operator. The following is an easily verifiable sufficient condition for this to hold.

THEOREM 3.2. *Consider* (3.1) *under the stated assumption. Then the unique solution of* (3.1) *satisfies* (3.2) *if* $\mathcal{U}(t, s)$ *is a strong evolution operator and* $\|\mathcal{U}^*(t, r)\mathcal{A}^*(T-r)x\|$ *is integrable in r on* (r, t) *for* $x \in \mathcal{D}_{A^*}$.

Proof. Under the above assumption, by Theorem 1.3, $\mathcal{Y}(t, s) = \mathcal{U}^*(T-s, T-t)$ is a quasi-evolution operator. By Theorem 1.2, the perturbation of $\mathcal{Y}(t, s)$ by $-C^*(T-s)P^*(T-s)$ is also a quasi-evolution operator and is uniquely defined by

$$\mathcal{Y}_p(t, s)x = \mathcal{Y}(t, s)x - \int_s^t \mathcal{Y}(t, \alpha)C^*(T-\alpha)P^*(T-\alpha)\mathcal{Y}_p(\alpha, s)x\, d\alpha \tag{3.3}$$

or equivalently by

$$\mathcal{Y}_p(t, s)x = \mathcal{Y}(t, s)x - \int_s^t \mathcal{Y}_p(t, \alpha)C^*(T-\alpha)P^*(T-\alpha)\mathcal{Y}(\alpha, s)x\, d\alpha. \tag{3.4}$$

By Corollary 1.2, $\mathcal{U}_p(t, s)$ is defined by

$$\mathcal{U}_p(t, s)x = \mathcal{U}(t, s)x - \int_s^t \mathcal{U}(t, \alpha)P(\alpha)C(\alpha)\mathcal{U}_p(\alpha, s)x\, d\alpha \tag{3.5}$$

and it is readily seen that $\mathcal{U}_p^*(T-s, T-t) = \mathcal{Y}_p(t, s)$, as they both satisfy (3.3), which has a unique solution.

Now we rewrite (3.1) in the form

$$P(T-t)x = \mathcal{Y}_p(T-t_0, T-t)P_0\mathcal{Y}_p(T-t_0, T-t)x + \int_{T-t}^{T-t_0} \mathcal{Y}_p^*(T-s, T-t) \cdot [W(T-s) + P(T-s)C(T-s)P(T-s)]\mathcal{Y}_p(T-s, T-t)x\, ds$$

and since $\mathcal{Y}_p(\cdot, \cdot)$ is a quasi-evolution operator we may apply Theorem 2.4 to differentiate (3.1).

COROLLARY 3.1. *Consider* (3.1) *under the stated assumptions. Then the unique solution of* (3.1) *satisfies* (3.2) *if any perturbation* $\mathcal{U}_F(t, s)$ *of* $\mathcal{U}(t, s)$ *is a strong evolution operator and* $\|\mathcal{U}_F^*(t, r)A^*(T-r)x\|$ *is integrable in r on* (s, t) *for* $x \in \mathcal{D}_{A^*}$.

Proof. By Corollary 1.5, $\mathcal{Y}(t, s) = \mathcal{U}^*(T-s, T-t)$ is a quasi-evolution operator and the proof of Theorem 3.2 may be used.

In § 4 we apply Theorem 3.2 to the dual Riccati equation obtained for differential-delay equations to obtain an improvement over previous results in [6] and [7]. However we caution that the differential Riccati equation (3.2) may not mean much as $\bigcap_{t \in T}(\mathcal{A}^*(t))$ is not dense in H in general (see [18]).

The conditions for (3.2) to have a unique solution are even stronger, namely that $\mathcal{Y}(t, s) = \mathcal{U}^*(T-s, T-t)$ or any perturbation be a strong evolution operator and this is impossible to formulate naturally in terms of conditions on $\mathcal{U}(t, s)$. So

we shall restrict ourselves to considering the special case when $\mathcal{U}(t, s)$ is generated by $\mathcal{A}+D(t)$.

THEOREM 3.3. *Consider* (3.1) *when* $\mathcal{U}(t, s)$ *is generated by* $\mathcal{A}+D(t)$, *where* $\mathcal{A}$ *is the infinitesimal generator of a strongly continuous semigroup* $\{\mathcal{T}(t); t \geqq 0\}$ *and* $D \in \mathcal{B}_\infty(T; \mathcal{L}(H))$. *Then* (3.1) *has a unique strongly continuous solution which is also the unique solution of the differentiated Riccati equation* (3.2) *in the class of strongly continuous operators with* $\langle P(t)x, y\rangle$ *absolutely continuous for* $x, y \in \mathcal{D}(\mathcal{A})$.

Proof. (3.1) may be written in the form (3.6), where $\mathcal{Y}_p(t, s)$ is the quasi-evolution operator with generator $\mathcal{A}^*-C^*(T-s)P^*(T-s)$. Now $\mathcal{A}^*$ is the infinitesimal generator of the strongly continuous semigroup $\{\mathcal{T}^*(t); t \geqq 0\}$, so by Corollary 2.1, (3.1) has a unique strongly continuous solution which satisfies the differentiated Riccati equation (3.2). By Corollary 2.3, (3.2) has a unique solution.

This is useful in obtaining a partial uniqueness result for differential delay equations (see § 4).

4. Applications. In this section we illustrate our results by applications to a variety of different problems. We have already seen (Corollaries 1.3, 2.2 and Theorems 2.3, 2.4, 2.5) that the integral and differential forms of the Riccati equation have unique solutions when the evolution operator is generated by $\mathcal{A}+D(t)$, with $\mathcal{A}$ generating a strongly continuous semigroup, and $D \in \mathcal{B}_\infty(T; \mathcal{L}(H))$. We also have (Theorem 3.3) that the integral and differential forms of the dual Riccati equation have unique solutions under these conditions.

4.1. Kato–Tanabe type evolution operators. In [12] and other papers Kato and Tanabe consider a class of abstract evolution equations, where the main requirement is that for each t, $\mathcal{A}(t)$ is the infinitesimal generator of an analytic semigroup. To be more precise, let Σ be the fixed closed angular domain

$$\Sigma = \{\lambda : |\arg \lambda| \leqq \pi/2+\theta,\ 0<\theta<\pi/2\}.$$

Assume that for each $t \in T$, $\mathcal{A}(t)$ is a densely defined closed linear operator such that:

(4.1) The resolvent set $\rho(\mathcal{A}(t))$ of $\mathcal{A}(t)$ contains Σ, and

$$\|(\lambda I-\mathcal{A}(t))^{-1}\| \leqq M/|\lambda| \quad \text{for } \lambda \in \Sigma, \quad t \in T,$$

(4.2) $\mathcal{A}(t)^{-1} \in \mathcal{L}(H)$ and is continuously differentiable in t on T in the uniform operator topology.

For any $\lambda \in \Sigma$, $t \in T$,

$$\left\|\frac{\partial}{\partial t}(\lambda I-\mathcal{A}(t))^{-1}\right\| \leqq \frac{N}{|\lambda|^{1-\rho}}, \qquad 0 \leqq \rho < 1, \tag{4.3}$$

$$\left\|\frac{d}{dt}(\mathcal{A}(t)^{-1})-\frac{d}{ds}(\mathcal{A}(s)^{-1})\right\| \leqq k|t-s|^\alpha; \qquad k>0, \quad \alpha>0. \tag{4.4}$$

Then it is shown in [12], that $\mathcal{A}(t)$ generates a strong evolution operator $\mathcal{U}(t, s)$ on $\Delta(T)$, which is strongly continuous and has the properties

$$\frac{\partial}{\partial s}\mathcal{U}(t, s)x = -\mathcal{U}(t, s)\mathcal{A}(s)x \quad \text{for all } x \in H, \quad t>s, \tag{4.5}$$

and

$$\left\|\frac{\partial}{\partial s}\mathcal{U}(t, s)\right\| \leqq \frac{c}{|t-s|},$$

$$\frac{\partial}{\partial t}\mathcal{U}(t, s)x = \mathcal{A}(t)\mathcal{U}(t, s)x \quad \text{for all } x \in H, \quad t > s, \tag{4.6}$$

and

$$\left\|\frac{\partial}{\partial t}\mathcal{U}(t, s)\right\| \leqq \frac{c}{|t-s|}.$$

In order to apply our results we need one further property of the evolution operator $\mathcal{U}(t, s)$—we require it to be a quasi-evolution operator, and we will prove this under the extra assumption

$$\sup_{t \in T}\|\mathcal{A}(t)x\| < C_x < \infty \quad \text{for } x \in \mathcal{D}_A = \bigcap_{t \in T}\mathcal{D}(\mathcal{A}(t)). \tag{4.7}$$

First we prove the following lemma.

LEMMA 4.1. *If $x \in \mathcal{D}_A$, $y \in H$, $\langle \mathcal{U}(t, r)\mathcal{A}(r)x, y\rangle$ is integrable on (s, t).*

Proof.

$$(\mathcal{A}(t)-\mathcal{A}(s))x = \mathcal{A}(t)(\mathcal{A}(s)^{-1}-\mathcal{A}(t)^{-1})\mathcal{A}(s)x \quad \text{for } x \in \mathcal{D}_A.$$

Thus $\langle(\mathcal{A}(t)-\mathcal{A}(s))x, y\rangle = \langle(\mathcal{A}(s)^{-1}-\mathcal{A}(t)^{-1})\mathcal{A}(s)x, \mathcal{A}^*(t)y\rangle$ for $x \in \mathcal{D}_A$, $y \in \mathcal{D}(\mathcal{A}^*(t))$. Hence

$$|\langle(\mathcal{A}(t)-\mathcal{A}(s))x, y\rangle| \leqq \|\mathcal{A}(s)^{-1}-\mathcal{A}(t)^{-1}\|\,\|\mathcal{A}(s)x\|\,\|\mathcal{A}^*(t)y\|$$

$$\leqq C_x\|\mathcal{A}(s)^{-1}-\mathcal{A}(t)^{-1}\|\|\mathcal{A}^*(t)y\|.$$

So by (4.2), $|\langle(\mathcal{A}(t)-\mathcal{A}(s))x, y\rangle| \to 0$ as $s \to t$ for $y \in \mathcal{D}(\mathcal{A}^*(t))$. But for each t, $\mathcal{A}(t)$ generates a semigroup, and so $\mathcal{A}^*(t)$ is also a closed linear operator with dense domain. Hence $\mathcal{A}(t)x$ is weakly continuous for $x \in \mathcal{D}_A$, so that $\langle \dot{\mathcal{U}}(t, \cdot)\mathcal{A}(\cdot)x, y\rangle$ is continuous on $[0, t)$ and hence integrable.

The above lemma together with (4.5) and the classical result corresponding to Property A.4 of the Appendix establishes that $\mathcal{U}(t, s)$ is a quasi-evolution operator. Thus the integral and differential forms of the Riccati equation have unique solutions.

We note that the assumption (4.7) is equivalent to requiring that the coefficients of the operator $\mathcal{A}(t)$ are uniformly bounded in t on T. In earlier papers on abstract evolution equations (for example, [11]) Kato imposed a slightly stronger set of conditions on the operator $\mathcal{A}(t)$ and then showed that the evolution operator $\mathcal{U}(t, s)$ not only satisfied (4.5) and (4.6) but was continuously differentiable in both arguments, so that in this case we may deduce that $\mathcal{U}(t, s)$ is a quasi-evolution operator without any extra assumptions.

When we consider the dual differential Riccati equation, it is easiest to prove that the dual evolution operator is both quasi and strong. For this we require the

assumption that

$$\sup_{t\in T}\|\mathscr{A}^*(t)y\|<C_y<\infty \quad \text{for } y\in\mathscr{D}_{A^*}. \tag{4.8}$$

Then it is easy to show that

$$\frac{\partial}{\partial t}\mathscr{Y}(t,s)x=\mathscr{A}^*(T-t)\mathscr{Y}(t,s)x \qquad \text{for all } x\in H,\quad t>s, \tag{4.9}$$

$$\frac{\partial}{\partial s}\mathscr{Y}(t,s)x=-\mathscr{Y}(t,s)\mathscr{A}^*(T-s)x \quad \text{for all } x\in H,\quad t>s, \tag{4.10}$$

where $\mathscr{Y}(t,s)=\mathscr{U}^*(T-s,T-t)$. Moreover the integrability condition follows as in Lemma 4.1, so that $\mathscr{U}(t,s)$ is both strong and quasi, so that the integral and differential forms of the dual Riccati equation have unique solutions.

4.2. Lions type evolution operators. In [13] Lions considers control problems for systems governed by parabolic partial differential equations. The operator $\mathscr{A}(t)$ is linked to a bilinear form $a(t;\varphi,\psi)$ on a Hilbert space V. Let V, H be Hilbert spaces such that H is identified with its dual. Then

$$V\subset H\subset V'.$$

Suppose that a family of bilinear forms on V are such that

$$a(t;\varphi,\psi) \quad \text{is measurable on } T \text{ for all } \varphi,\psi\in V; \tag{4.11}$$

$$|a(t;\varphi,\psi)|\leqq C\|\varphi\|_V\|\psi\|_V; \tag{4.12}$$

there exist λ, α such that $\alpha>0$ and

$$a(t;\varphi,\varphi)+\lambda\|\varphi\|_H^2\geqq\alpha\|\varphi\|_V^2 \quad \text{for all } \varphi\in V,\quad t\in T. \tag{4.13}$$

Then for each t it is possible to write

$$a(t;\varphi,\psi)=-\langle\mathscr{A}(t)\varphi,\psi\rangle_{V'V}$$

where $\langle\cdot,\cdot\rangle_{V'V}$ denotes the duality between V' and V. Lions shows that there is a unique solution in $W(0,T)$ of

$$\begin{aligned}\frac{dz}{dt}&=\mathscr{A}(t)z,\\ z(0)&=z_0\in H,\end{aligned} \tag{4.14}$$

where the equation is to be interpreted in the sense of distributions, and $W(0,T)$ is the Hilbert space

$$W(0,T)=\left\{x : x\in L_2(0,T;V),\frac{dx}{dt}\in L_2(0,T;V')\right\}$$

with norm

$$\|x\|^2_{W(0,T)}=\|x\|^2_{L_2(0,T;V)}+\left\|\frac{dx}{dt}\right\|^2_{L_2(0,T;V')}.$$

Moreover the solution depends continuously on the initial data, in the sense that the map $x_0 \to x(\cdot)$ from $H \to W(0, T)$ is continuous. By considering the equation

$$\frac{dz}{dt} = \mathcal{A}(t)z \quad \text{in } (s, t),$$

$$z(s) = z_0 \in H,$$

it is very easy to see that

$$z(t) = \mathcal{U}(t, s)z_0, \tag{4.15}$$

where $\mathcal{U}(t, s) \in \mathcal{L}(H)$, and has the evolution property (1.1). Using the continuous dependence on the initial data, it is easy to show that $\mathcal{U}(t, s)$ is strongly continuous in s and uniformly bounded on $\Delta(T)$. Since all $x \in W(0, T)$ are with modification on a set on measure zero, continuous from $[0, T] \to H$ it follows that $\mathcal{U}(t, s)$ is strongly continuous in t. Then $\mathcal{U}(t, s)$ is certainly a mild evolution operator, and so the integral form of the Riccati equation has a unique solution.

From the definitions it is easy to show that the solution of (4.15) must satisfy

$$\int_s^t a(\rho; x(\rho), \bar{x})\, d\rho = -\langle x(t), \bar{x}\rangle + \langle x(s), \bar{x}\rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product on H, and $\bar{x} \in V$.

$$\begin{aligned} \text{Now } a(\rho; x(\rho), \bar{x}) &= \langle \mathcal{A}(\rho)x(\rho), \bar{x}\rangle_{V'V} \\ &= -\langle x(\rho), \mathcal{A}^*(\rho)\bar{x}\rangle_{VV'}. \end{aligned}$$

But $\langle \cdot, \cdot \rangle_{VV'}$ is an extension of the inner product on H, so that if we define $\mathcal{D}(\mathcal{A}^*(t))$ by

$$\mathcal{D}(\mathcal{A}^*(t)) = \{v \in V : \mathcal{A}^*(t)v \in H\}, \tag{4.16}$$

then for $\bar{x} \in \mathcal{D}_{A^*}$, we have

$$a(\rho; x(\rho), \bar{x}) = -\langle x(\rho), \mathcal{A}^*(\rho)\bar{x}\rangle.$$

Thus

$$-\int_s^t \langle \mathcal{U}(\rho, s)x_0, \mathcal{A}^*(\rho)\bar{x}\rangle\, d\rho = -\langle \mathcal{U}(t, s)x_0, \bar{x}\rangle + \langle x_0, \bar{x}\rangle.$$

Setting $\rho = T - \gamma$, $t = T - s$ and $\mathcal{Y}(t, s) = \mathcal{U}^*(T - s, T - t)$ gives

$$\int_s^t \langle x_0, \mathcal{Y}(t, \gamma)\mathcal{A}^*(T - \gamma)\bar{x}\rangle\, d\gamma = \langle x_0, \mathcal{Y}(t, s)\bar{x}\rangle - \langle x_0, \bar{x}\rangle \quad \text{for } \bar{x} \in \mathcal{D}_{A^*}.$$

So that $\mathcal{Y}(t, s) = \mathcal{U}^*(T - s, T - t)$ is a quasi-evolution operator. In a similar manner, using the unique solution of the adjoint equation, it is possible to show that $\mathcal{U}(t, s)$ is a quasi-evolution operator, where

$$\mathcal{D}(\mathcal{A}(t)) = \{v \in V : \mathcal{A}(t)v \in H\}. \tag{4.17}$$

This enables us to differentiate the integral Riccati equation for both the control and filtering problems, but we are not able to say whether the solution of

the differential Riccati equation is unique or not, since we do not know if the evolution operator is strong.

4.3. Hyperbolic partial differential equations. In [21] Vinter considers the quadratic cost control problem for a class of hyperbolic partial differential equations, and allows for both boundary and distributed control. Because of the difficulties of boundary control, his best results are for distributed control, where he obtains a feedback optimal control, and a differential Riccati equation using the "Lions" approach. We shall show how these results can be obtained by applying the general theory of § 2. Of course the most important requirements are existence and uniqueness results.

Consider

$$\begin{aligned} &\frac{\partial y}{\partial t} = \sum_{i=1}^{m} A_i \frac{\partial y}{\partial x_i} + Ky + Bv, \\ &My|_{\partial\Omega} = 0, \qquad y(0) = y_0 \end{aligned} \tag{4.18}$$

on the spatial domain $\Omega = \{x \in \mathbb{R}^m;\ x_1 > 0\}$; $0 \leqq t \leqq T$, where A_i, K, M are C^∞ matrix-valued functions on $Q = \mathbb{R} \times \Omega$ and $\Sigma = \mathbb{R} \times \partial\Omega$, respectively. Denote by $C^\infty_{(0)}(Q; \mathbb{R}^k)$ the restriction of $C^\infty_{(0)}(\mathbb{R}^{m+1}; \mathbb{R}^k)$ to the closure of $T \times \Omega$. Then we define a strong solution $y \in L_2(Q; \mathbb{R}^n)$ to (4.18). If given $Bv \in L_2(Q; \mathbb{R}^n)$ and $y_0 \in L_2(\Omega; \mathbb{R}^n)$, there exists a sequence $\{y_n\}$ with $y_n \in C^\infty_{(0)}(Q; \mathbb{R}^n)$ such that

$$\begin{aligned} &\|y_n - y\|_{L_2(Q;\,\mathbb{R}^n)} \to 0, \\ &\|y_n - y\|_{L_2(\Sigma;\,\mathbb{R}^n)} \to 0, \\ &\left\| \frac{\partial}{\partial t} y_n - \sum_{i=1}^{m} A_i \frac{\partial y_n}{\partial x_i} - Ky_n - Bv \right\|_{L_2(Q;\,\mathbb{R}^n)} \to 0, \\ &\|My^n\|_{L_2(\Sigma;\,\mathbb{R}^k)} \to 0, \\ &\|y_n(0) - y_0\|_{L_2(\Omega;\,\mathbb{R}^n)} \to 0. \end{aligned} \tag{4.19}$$

Then under technical assumptions on A_i, which ensure that the system is strictly hyperbolic with noncharacteristic boundary and determinate boundary values, we can assert that [21] for given $Bv \in L_2(Q; \mathbb{R}^n)$, $y_0 \in L_2(\Omega; \mathbb{R}^n)$, (4.18) has a unique strong solution $y \in L_2(Q; \mathbb{R}^n)$. Furthermore, the map $t \to y(t)$ from $T \to L_2(\Omega; \mathbb{R}^n)$ is strongly continuous, and we have the estimate

$$\|y(t)\|_{L_2(\Omega;\,\mathbb{R}^n)} \leqq C \|y_0\|_{L_2(\Omega;\,\mathbb{R}^n)} \quad \text{for all } t \in T \tag{4.20}$$

and C is independent of t.

Similarly the following adjoint system has a unique strong solution $p \in L_2(Q; \mathbb{R}^n)$ with $p(t)$ well-defined and strongly continuous in t as an element of $L_2(\Omega; \mathbb{R}^n)$:

$$\begin{aligned} &-\frac{\partial p}{\partial t} = \sum_{i=1}^{m} A_j' \frac{\partial p}{\partial xj} + K'p - \sum_{j=1}^{m} \frac{\partial A_j}{\partial x_j} p, \\ &M'p|_{\partial\Omega} = 0, \\ &p(T) = 0. \end{aligned} \tag{4.21}$$

Moreover we have a similar estimate to (4.20).

It is easy to show that (4.18) and (4.21) are equivalent to (4.22), (4.23), which is a more appropriate formulation for our purposes.

$$\begin{aligned} &\dot{z}(t) = \mathcal{A}(t)z(t) + B(t)u(t), \\ &z(0) = z_0, \end{aligned} \tag{4.22}$$

$$\begin{aligned} &\dot{p}(t) = -\mathcal{A}^*(t)p(t), \\ &p(T) = p_1, \end{aligned} \tag{4.23}$$

where z_0, p_1, $z(t)$, $p(t) \in H = L_2(\Omega; \mathbb{R}^n)$ for each $t \in T$, $u \in L_2(T; H)$, $B \in \mathcal{B}_\infty(T; \mathcal{L}(H))$ and $\mathcal{A}(t)$ is a linear operator on H given by

$$(\mathcal{A}(t)h)(x) = \sum_{i=1}^{m} A_i(t, x)\frac{\partial h}{\partial x_i} + K(t, x)h \tag{4.24}$$

with domain

$$\mathcal{D}(\mathcal{A}(t)) = \left\{ \begin{matrix} h \in H : \mathcal{A}(t)h \in H & \text{and} \\ Mh|_{\partial\Omega} = 0 & \end{matrix} \right\}. \tag{4.25}$$

$\mathcal{A}^*(t)$ is then the H-adjoint of $\mathcal{A}(t)$ for each $t \in T$. Since the map $z_0 \to z(t)$ is continuous from H to H, we may write $z(t) = \mathcal{U}(t, s)z(s)$ for $0 \leqq s \leqq t \leqq T$. Then using the estimate (4.20) and the continuity of $z(t)$ in t, we can show in a manner similar to the previous example (4.2) that $\mathcal{U}(t, s)$ is a mild evolution operator on $\Delta(T)$. We can also show it is quasi, as follows.

First we show that the solution of (4.23) is

$$p(t) = \mathcal{U}^*(T, t)p_1.$$

To see this we use a result of [17] where it is shown that

$$\langle z(t_0), p(t_0)\rangle = \langle z(t_1), p(t_1)\rangle \quad \text{for } 0 \leqq t_0 \leqq t \leqq T.$$

But $z(t_1) = \mathcal{U}(t_1, t_0)z(t_0)$, hence

$$\langle z(t_0), p(t_0) - \mathcal{U}^*(t_1, t_0)p(t_1)\rangle = 0 \quad \text{for all } z_0 \in H.$$

Hence $p(t_0) = \mathcal{U}^*(t_1, t_0)p(t_1)$.

Now let $\{p_1^n\}$ be an approximating sequence in $C_{(0)}^\infty(Q; \mathbb{R}^n)$ to p_1. Then it is known [17] that the strong solution $p^{(n)}$ such that $p^n(t) = p_1^n$, belongs to $C_0^\infty(Q; \mathbb{R}^n)$. Thus for any $x \in L_2(\Omega; \mathbb{R}^n)$, we have

$$\int_s^t \langle \dot{p}^n(\rho) + \mathcal{A}^*(\rho)p^n(\rho), x\rangle \, d\rho = 0,$$

or

$$\int_s^t \cdot \langle \mathcal{A}^*(\rho)p^n(\rho), x\rangle \, d\rho + \langle p_1^n, x\rangle - \langle p^n(s), x\rangle = 0.$$

Now if $x \in \mathcal{D}_A$, we obtain

$$\int_s^t \langle p^n(\rho), \mathcal{A}(\rho)x\rangle \, d\rho + \langle p_1^n, x\rangle - \langle p^n(s), x\rangle = 0.$$

Applying the Lebesque dominated convergence theorem and using

$$p(\rho) = \mathcal{U}^*(t, \rho)p_1, \qquad p_1 \in H,$$

gives

$$\int_s^t \langle p_1, \mathcal{U}(t, \rho)\mathcal{A}(\rho)x\rangle \, d\rho = \langle p_1, \mathcal{U}(t, s)x\rangle - \langle p_1, x\rangle.$$

Thus $\mathcal{U}(t, \rho)$ is a quasi-evolution operator. Similarly it can be shown that $\mathcal{U}^*(T-\rho, T-t)$ is also quasi, so that there are unique solutions to the integral Riccati equations and hence to the optimal control and filtering problems. The Riccati equations may be differentiated, but we do not know whether the differential forms have unique solutions, since we are not able to prove that $\mathcal{U}(t, \rho)$ is a strong evolution operator.

4.4. Differential-delay systems. We now consider the class of linear differential-delay systems first considered by Delfour and Mitter in [9].

$$\begin{aligned} \frac{dx(t)}{dt} &= a_{00}(t)x(t) + \sum_{i=1}^{N} a_i(t)\begin{cases} x(t+\theta_i); & t+\theta_i \geqq 0 \\ h(t+\theta_i); & t+\theta_i < 0 \end{cases} \\ &\quad + \int_{-b}^{0} a_{01}(t, \theta)\begin{cases} x(t+\theta; & t+\theta \geqq 0 \\ h(t+\theta); & t+\theta < 0 \end{cases} d\theta, \\ x(0) &= h(0), \end{aligned} \tag{4.26}$$

where $t \in [0, T] = T$, X is a real separable Hilbert space, $a_{00} \in C(T; \mathcal{L}(X))$, $a_i \in C(\mathcal{L}(X))$, $a_{01} \in C(Tx(-b, 0); \mathcal{L}(X))$ and $-b < -\theta_N < \cdots < -\theta_0 = 0$.

This may be considered as abstract evolution equation on a Hilbert space $\mathcal{M}^2(-b, 0; X)$. The quotient space of $\mathcal{L}^2(-b, 0; X)$ generated by equivalence classes under the $\mathcal{M}^2$-norm:

$$\|y^2\|_{\mathcal{M}^2} = (\|y(0)\|_X^2 + \int_{-b}^{0} \|y(\theta)\|_X^2 \, d\theta)^{1/2}.$$

$\mathcal{M}^2(-b, 0; X)$ is isometrically isomorphic to $X \times L_2(-b, 0; X)$. (4.26) is now equivalent to the abstract evolution equation on $\mathcal{M}^2$:

$$\begin{aligned} \frac{dz(t)}{dt} &= \mathcal{A}(t)z(t), \\ z(0) &= h; \qquad h \in \mathcal{D}, \end{aligned} \tag{4.27}$$

where $\mathcal{A}(t)$ is a closed linear operator on $\mathcal{M}^2$ with domain $\mathcal{D}$:

$$\mathcal{D} = \left\{ \begin{array}{l} y \in \mathcal{M}^2(-b, 0; X), \text{ such that } y \text{ is absolutely continuous} \\ \text{and } \int_{-b}^{0} \left\| \frac{dy(\theta)}{d\theta} \right\|_X^2 d\theta < \infty \end{array} \right\}$$

and

$$(\mathcal{A}(t)h)(\theta) = \begin{cases} a_{00}(t)h(0) + \sum_{i=1}^{N} a_i(t)h(\theta_i) + \int_{-b}^{0} a_{01}(t, \theta)h(\theta) \, d\theta; & \theta = 0, \\ \frac{dh(\theta)}{d\theta}; & \theta \neq 0. \end{cases} \tag{4.28}$$

It is shown in [9] that $\mathcal{A}(t)$ generates a strong evolution operator $\mathcal{U}(t, s)$ on $\Delta(T)$. Since the coefficients a_{00}, a_i, and a_{01} are continuous, it is readily seen that $\sup_{t\in T}\|\mathcal{A}(t)h\| < C$ for $h \in \mathcal{D}$ and Lemma 1.1 may be applied to show that $\mathcal{U}(t, s)$ is also a quasi-evolution operator. Hence the integral and differential forms of the Riccati equation both have unique solutions.

For the dual Riccati equation, Theorem 3.1 ensures that we have a unique solution of the integral version, but the differential version proves more difficult.

LEMMA 4.2. *The differential version of the dual Riccati equation holds. It has a unique solution in the special case where $a_i(t)$ in (4.11) are time-invariant.*

Proof. By Theorem 3.2, the differential version is valid if $\|\mathcal{U}^*(t, r)\mathcal{A}^*(T-r)x\|$ is integrable and since $\mathcal{U}^*(t, r)$ is uniformly bounded in norm on $\Delta(T)$, we need only verify that $\sup_{t\in T}\|\mathcal{A}^*(t)h\| < \infty$ for $h \in \mathcal{D}(\mathcal{A}^*(t))$. From [19] we have that

$$\mathcal{D}(\mathcal{A}^*(t)) = \left\{ \begin{array}{l} h \in \mathcal{M}^2 : h(\theta) = z(\theta) + \sum_{i=1}^{N} a_i^*(t)h(0)\chi_i(\theta);\ \theta \neq 0 \\ \text{where } z \text{ and } z' \in L_2(-b, 0; X) \text{ with } z(-b) = 0 \end{array} \right\}.$$

$\chi_i(\theta)$ is the characteristic function of the interval $[\theta_i, 0)$ and * denotes the X-adjoint operation.

Then for $h \in \mathcal{D}(\mathcal{A}^*(t))$, we have

$$(\mathcal{A}^*(t)h)(0) = \begin{cases} a_{00}^*(t)h(0) + \sum_{i=1}^{N} a_i^*(t)h(0) + \int_{-b}^{0} z'(\theta)\, d\theta, & \theta = 0, \\ a_{01}^*(t, \theta)h(\theta) - \dfrac{dz}{d\theta}, & \theta \neq 0. \end{cases}$$

Again since a_{00}, a_{01} and a_i are continuous functions, we have that $\sup_{t\in T}\|\mathcal{A}^*(t)h\| < C$ for each $h \in \mathcal{D}(\mathcal{A}^*(t))$. So the differential version of the dual Riccati equation holds.

Consider now the time-invariant linear operator $\mathcal{A}$ defined by

$$(\mathcal{A}h)(0) = \begin{cases} \sum_{i=1}^{N} a_i h(\theta_i), & \theta = 0, \\ \dfrac{dh(\theta)}{d\theta}, & \theta \neq 0. \end{cases}$$

$\mathcal{A}$ is the infinitesimal generator of a strongly continuous semigroup. Consider $D \in \mathcal{B}_\infty(T; \mathcal{L}(\mathcal{M}^2))$ defined by

$$(D(t)h)(\theta) = \begin{cases} a_{00}(t)h(0) + \int_{-b}^{0} a_{01}(t, \theta)h(\theta)\, d\theta, & \theta = 0, \\ 0, & \theta \neq 0. \end{cases}$$

Then $\mathcal{A} + D(t)$ is just the special case of (4.28), where the a_i are time-invariant, and by Theorem 3.3, the differential version of the dual Riccati equation has a unique solution. This agrees with the results in [7] and [14].

Appendix. Bochner integration. Let H, K be Hilbert spaces and $T=[0, T]$, a real time interval. We consider Bochner integrals of H-valued functions of t and recall the definitions of strong measurability.

DEFINITION A.1. $x : T \to H$ is *strongly measurable* if there exists a sequence of finitely-valued functions $\{x_n(t)\}$ such that

$$\|x_n(t)-x(t)\| \to 0 \quad \text{as } n \to \infty \text{ a.e. on } T.$$

The following are standard results on Bochner integration which are relevant to our interests (see [10]).

PROPERTY A.1. If $x : T \to H$ is the limit almost everywhere of a sequence of strongly measurable functions, then $x(\cdot)$ is strongly measurable on T.

PROPERTY A.2. A function $x : T \to H$ is Bochner integrable if and only if x is strongly measurable and

$$\int_T \|x(t)\|\, dt < \infty.$$

We remark that for operator-valued functions there arise two kinds of measurability, depending on whether one refers to the uniform or the strong topology. One usually says that $D(\cdot) : T \to \mathcal{L}(H, K)$ is strongly measurable if $D(\cdot)x$ is strongly measurable for all $x \in H$.

In order to make sense of some of the integrals we have used in § 1, 2, we need the following lemma.

LEMMA A.1. *If $D(\cdot) : T \to \mathcal{L}(H, K)$ is strongly measurable on T and* $\operatorname{ess\,sup}_{t\in T}\|D(t)\| < \infty$, *we write $D \in \mathcal{B}_\infty(T; \mathcal{L}(H, K))$ and can easily show that if $f \in L_1(T; H)$, then $D(\cdot)f(\cdot) \in L_1(T; K)$.*

An immediate corollary is that

$$L_\infty(T; \mathcal{L}(H, K)) \subset \mathcal{B}_\infty(T; \mathcal{L}(H, K))$$

and if $D(\cdot) \in \mathcal{L}(H, K)$ is strongly continuous on T, then $D \in \mathcal{B}_\infty(T; \mathcal{L}(K, H))$. Other useful properties are as follows.

PROPERTY A.3. If $D(t) \in \mathcal{L}(H, K)$ and $D(t)x$ is weakly continuous on T, then $D \in \mathcal{B}_\infty(T; \mathcal{L}(H, K))$.

PROPERTY A.4 (Vinter [18]). Suppose that $f : [0, T] \to H$ is weakly continuous, and the weak right derivative $\partial^+ f(t)$ exists for *every* $t \in [0, T]$. Suppose $\partial^+ f(\cdot) \in L_1(T; H)$. Then

$$f(t) = f(0) + \int_0^t \partial^+ f(s)\, ds \quad \text{for each } t \in [0, T].$$

Many of the measurability difficulties we have encountered disappear when the Hilbert space H is separable. The reason for this is that strong measurability and weak measurability coincide in this case. That is a function $x : T \to H$ is strongly measurable if $\langle y, x(t)\rangle$ is measurable for all $y \in H$.

Acknowledgment. The authors would like to thank Dr. R. Vinter for several useful discussions and Professor J. L. Lions for informing them of a related paper by L. Tartar, *Sur l'étude directe d'équations non linéaires intervenant en théorie du contrôle optimal*, which is to appear in the Journal of Functional Analysis.

REFERENCES

[1] A BENSOUSSAN, *Filtrage Optimal des Systèmes Linéaires*, Dunod, Paris, 1971.

[2] A. BENSOUSSAN, M. DELFOUR AND S. K. MITTER, *Notes on Infinite Dimensional Systems*, to appear.

[3] R. F. CURTAIN, *Infinite-dimensional filtering*, this Journal, 13 (1975), pp. 89–104.

[4] ———, *The infinite-dimensional Riccati equation with applications to affine hereditary differential systems*, this Journal, 13 (1975), pp. 1130–1143.

[5] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Riccati equation*, J. Math. Anal. Appl., 47 (1974), pp. 43–57.

[6] R. F. CURTAIN, *Infinite dimensional estimation theory for linear systems*, Rep. 38, Control Theory Centre, Univ. of Warwick, Coventry, England, 1975.

[7] ———, *A Kalman–Bucy filtering theory for affine hereditary differential equations*, Control Theory, Numerical Methods and Computer Systems Modelling, no. 107, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1975.

[8] R. DATKO, *A linear control problem in abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[9] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972), pp. 298–327.

[10] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, vol. 31, American Mathematical Society, Providence, R.I., 1957.

[11] T. KATO, *Abstract evolution equations of parabolic type in Banach and Hilbert spaces*, Nagoya Math. J., 19 (1961), pp. 93–125.

[12] T. KATO AND H. TANABE, *On the abstract evolution equation*, Osaka Math. J., 14 (1962), pp. 107–133.

[13] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[14] S. K. MITTER AND R. B. VINTER, *Filtering for linear stochastic hereditary differential systems*, Control Theory, Numerical Methods and Computer-Systems Modelling, no. 107, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1975.

[15] R. S. PHILLIPS, *Perturbation theory for semigroups of linear operators*, Trans. Amer. Math. Soc., 74 (1953), pp. 199–221.

[16] A. J. PRITCHARD, *Stability and control of distributed parameter systems governed by wave equations*, IFAC Conf. on Distributed Parameter Systems, Banff, Canada, 1971.

[17] J. RAUCH, *L^2 is a continuable initial condition for Kreiss' mixed problems*, Comm. Pure Appl. Math., 25 (1972), pp. 265–285.

[18] R. B. VINTER, *Some results concerning perturbed evolution operators with applications to delay equations*, Dept. of Computing and Control Report, Imperial College, London, 1975.

[19] ———, *Representation of solutions of stochastic delay systems*, Imperial College Report, London, 1975.

[20] ———, *On the evolution of the state of linear differential delay equations in M^2, Properties of the generator*, Rep. ESL-R-541, Electronics Systems Lab., Mass. Inst. of Tech., Cambridge, Mass., 1974.

[21] ———, *Optimal control of non-symmetric hyperbolic systems in* n*-variables on the half-space*, Imperial College Report, London, 1974.

[22] K. YOSIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1966.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14. No. 6. November 1976

STABLE SETS AND STABLE POINTS OF SET-VALUED DYNAMIC SYSTEMS WITH APPLICATIONS TO GAME THEORY*

MICHAEL MASCHLER AND BEZALEL PELEG†

Abstract. Let X be a metric space. A dynamic system on X is a set-valued function φ from X to X which satisfies $\varphi(x) \neq \varnothing$ for $x \in X$. It generates φ-sequences:

$$x^{(t+1)} \in \varphi(x^{(t)}), \qquad t = 0, 1, 2, \cdots, x^{(0)} \in X.$$

We study the stability properties of such dynamic systems. Necessary and sufficient criteria for stability of sets and points are given. The main result is, essentially, that a subset of X is stable iff it is an inverse image of a Pareto minimal point of a vector-valued function which decreases along φ-sequences.

As a corollary we obtain a characterization of all stable sets and points of Stearns' transfer schemes as generalized nucleoli. In particular, the "lexicographic kernel", is always a stable set of the bargaining sets which may not include the nucleolus. All nonempty ε-cores are also stable sets of the bargaining sets.

1. Introduction. This paper is a contribution to the stability theory of dynamic systems with applications to game theory. Generally speaking, it has two novelties:

(i) It introduces vectorial Lyapunov functions. We exhibit in this paper that sometimes vectorial Lyapunov functions are more natural and perhaps easier to find than a single Lyapunov function; therefore, they are highly useful both for deducing general results and for treating particular systems.

(ii) We study stability properties of (discrete) *set-valued* dynamic systems. Such a system is defined by a set-valued function $\varphi: X \to 2^X$, from a metric space X into the space of its nonempty subsets. A *dynamic process*, or a trajectory, starting at $x^0 \in X$ is a sequence $x^0, x^1, \cdots$ having the property that $x^{i+1} \in \varphi(x^i)$, $i = 0, 1, \cdots$.

In the classical physical sciences, usually a process was completely determined by the initial conditions. For such sciences a theory for set-valued dynamic systems was quite superfluous. This is no longer the case for mathematical models for situations which involve human decisions. If the decision making person has more than one option, one cannot, in general, predict the "state of the world" at time $t = 1$ from the knowledge of the "state of the world" x^0 at time $t = 0$. All one can say is that the state at $t = 1$ will belong to a set of states $\varphi(x^0)$. The theory of set-valued dynamic systems should have applications to the social sciences, to control theory (see, e.g., Hermes (1970)) and to operations research (see, e.g., Zangwill (1969), who essentially calls such systems "autonomous algorithms"). Justman (1973) employs such systems to describe "iterative negotiations."

In this paper we address ourselves to the following basic problems:

1. Provide conditions that guarantee the existence of an "endpoint" (i.e., a point x satisfying $\varphi(x) = \{x\}$).

* Received by the editors June 16, 1975.

† Institute of Mathematics, The Hebrew University, Jerusalem, Israel.

2. Provide conditions that guarantee that each dynamic process converges.

3. Provide conditions that guarantee that the limit of a dynamic process is an endpoint.

4. Characterize closed stable sets of the system.

5. Characterize stable points of the system.

6. Under what conditions can one claim that a closed stable set consists of, or at least contains, stable points?

Problems 1–3 were already treated in Justman (1973), and the present paper is mainly concerned with problems 4–6 and applications.

A prime role in this investigation is played by the *nucleolus* of a vectorial Lyapunov function g, which is the inverse image of a Pareto minimal point of g. This concept is a somewhat sophisticated generalization of Schmeidler's nucleolus of a set with respect to a given game (1969).

A particular case of a dynamic system satisfying all our requirements is Stearns' transfer sequences which converge to the appropriate bargaining set of a given game (Stearns (1968); see also Billera (1972)). Stearns (1968) already observed that an endpoint of such a system may not be stable. Kalai, Maschler and Owen (1973) started a systematic investigation of asymptotically stable points in the various bargaining sets. They succeeded in showing that Schmeidler's nucleolus is a stable point for each system (and the only asymptotically stable point, in case it is isolated).

In this paper we characterize all the stable points and stable closed sets with respect to Stearns' systems which belong to the appropriate bargaining sets. They are nucleoli of appropriate Lyapunov functions. In particular, a new solution concept due to Gill Kalai, called the *lexicographic kernel* is shown to be a stable subset of the kernel. An example is provided of a game whose lexicographic kernel does not even contain the nucleolus of the same game. We also show that all nonempty ε-cores are stable sets for each bargaining set.

2. Dynamic systems. Let X be a metric space. A (set-valued) *dynamic system* on X is a set-valued function φ from X to X (i.e., a function from X to 2^X) which satisfies $\varphi(x) \neq \emptyset$ for all $x \in X$.

Let φ be dynamic system on X. A *φ-sequence* (starting at x^0) is a sequence (x^t), such that $x^0 \in X$ and $x^{t+1} \in \varphi(x^t)$, $t = 0, 1, 2, \cdots$. A φ-sequence is sometimes called a *trajectory* or a *dynamic process*. A point $x \in X$ is called an *endpoint* of φ (or a *critical-point*, or a *rest-point* of φ) if

$$\varphi(x) = \{x\}. \tag{2.1}$$

The set of all endpoints of φ will be denoted by $E\varphi$.

In this section we shall state conditions that guarantee the existence of endpoints, conditions that insure that each φ-sequence converges and conditions that guarantee that the limit points of converging φ-sequences are endpoints. Most results of this section, were obtained, or could be deduced from Justman (1973). We list them here in a way that is convenient for this paper both for the sake of completeness and in order to prepare the necessary tools for studying the stability properties of φ.

If $d(\cdot,\cdot)$ is the distance function of X, it is convenient to introduce the (generalized) real-valued function $f: X \to R_+ \cup \{\infty\}$ defined by

$$f(x) = \sup \left\{ \sum_{t=0}^{\infty} d(x^{t+1}, x^t) \mid (x^t) \text{ is a } \varphi\text{-sequence}, x^0 = x \right\}. \tag{2.2}$$

This function is the supremum of the lengths of all the trajectories that start at x. Obviously, it may take the value $+\infty$ and it is equal to 0 iff x is an endpoint.

LEMMA 2.1. *If φ is* l.s.c. (*lower semicontinuous*) *then f is* l.s.c.

Remark 2.2. Note that the lemma uses the term l.s.c. in two different meanings.[1]

Proof of Lemma 2.1. Let $x \in X$ and let $f(x) \geqq M$. Suppose that $y^k \in X$, $k = 1, 2, \cdots$, and $x = \lim_{k\to\infty} y^k$. We have to show that $\liminf_{k\to\infty} f(y^k) \geqq M$. Let $\varepsilon > 0$. There exist points $x^0 = x$, $x^1 \in \varphi(x^0), \cdots, x^T \in \varphi(x^{T-1})$ such that

$$\sum_{t=0}^{T-1} d(x^{t+1}, x^t) \geqq M - \varepsilon. \tag{2.3}$$

Since φ is l.s.c., there exist $T+1$ sequences $(y^{i,k})$, $i = 0, 1, \cdots, T$, where $y^{0,k} = y^k$, $y^{1,k} \in \varphi(y^{0,k}), \cdots, y^{T,k} \in \varphi(y^{T-1,k})$, $k = 1, 2, \cdots$, such that

$$\lim_{k\to\infty} y^{t,k} = x^t, \qquad t = 0, 1, \cdots, T. \tag{2.4}$$

It follows now from (2.3) and (2.4) that

$$\liminf_{k\to\infty} f(y^k) \geqq \lim_{k\to\infty} \sum_{t=0}^{T-1} d(y^{t+1,k}, y^{t,k}) \geqq M - \varepsilon.$$

Since ε was an arbitrary positive number, the proof of the lemma is complete.

Remark 2.3. *The function f monotonically decreases on trajectories. More precisely,*

$$y \in \varphi(x) \Rightarrow f(x) \geqq f(y) + d(x, y). \tag{2.5}$$

THEOREM 2.4. *Assume that X is not empty and compact. If φ is* l.s.c. *and for some $x \in X$, $f(x) < \infty$, then $E\varphi \neq \varnothing$.*

Proof. Let $A = \{y \in X \mid f(y) \leqq f(x)\}$. By Lemma 2.1, A is compact; hence, again by Lemma 2.1, there exists z in A such that $f(z) \leqq f(y)$ for all y in A. By Remark 2.3, z is an endpoint of φ.

Remark 2.5. *If φ is* l.s.c., *then $E\varphi$ is a closed set.*

Proof. $E\varphi = \{x \in X \mid f(x) \leqq 0\}$. In view of Lemma 2.1, this is a closed set. Obviously, if X is complete and $f(x) < \infty$ for some x in X, then each φ-sequence that starts at x must converge.

LEMMA 2.6. *If a bounded function $\psi: X \to R$ exists, which satisfies*

$$y \in \varphi(x) \Rightarrow \psi(x) - \psi(y) \geqq d(x, y), \tag{2.6}$$

then $f(x) < \infty$ for all $x \in X$.

---

[1] A real-valued function $f(x)$ is l.s.c. at x^0 if $x^n \to x^0 \Rightarrow \liminf_{n\to\infty} f(x^n) \geqq f(x^0)$. A set-valued function $\varphi(x)$ is l.s.c. at x^0 if $x^n \to x^0$ and $y^0 \in \varphi(x^0) \Rightarrow$ there exist $y^n \in \varphi(x^n)$, $n = 1, 2, \cdots$, such that $y^n \to y^0$.

Proof. Let M be a bound for ψ and let (x^n) be a φ-sequence, $x^0 = x$. Clearly, for each positive integer T,

$$\sum_{t=0}^{T} d(x^{t+1}, x^t) \leqq \sum_{t=0}^{T} [\psi(x^t) - \psi(x^{t+1})] \leqq 2M. \tag{2.7}$$

Consequently, $f(x) \leqq 2M < \infty$.

The function ψ can be regarded as a Lyapunov function whose existence guarantees convergence.

For $x \in X$ we shall now define the function

$$\rho(x) = \sup \{d(y, x) | y \in \varphi(x)\}. \tag{2.8}$$

$\rho(x)$ is the supremum of the lengths of single transition starting from x. The transition from x to y will be called *α-maximal* if $d(y, x) \geqq \alpha\rho(x)$. A φ-sequence will be called *maximal* if there exists $\alpha > 0$, such that the sequence contains infinitely many α-maximal transitions. Under quite general conditions, a converging maximal φ-sequence must converge to an endpoint.

THEOREM 2.7. *Let φ be* l.s.c. *If (x^t) is a converging maximal φ-sequence, $x^0 \in X$, then its limit z is an endpoint.*

Proof. Since φ is l.s.c., it follows that $\rho(x)$ is l.s.c. If $z \notin E\varphi$, then $\rho(z) > 0$; hence $\rho(x^t) > \frac{1}{2}\rho(z)$ for large enough t's. Consequently, $d(x^{t+1}, x^t) > (\alpha/2)\rho(z) > 0$ infinitely often and the sequence cannot converge.

Remark 2.8. It follows from Theorem 2.7 that Theorem 2.4 remains true if one replaces the requirement that X is compact by the requirement that it is complete.

3. Generalized nucleoli and stable sets. Let φ be a set-valued dynamic system defined on a metric space X.

DEFINITION 3.1. Let Q be a nonempty subset of X. Q is called *stable* w.r.t. (with respect to) φ if for every neighborhood U of Q there exists a neighborhood V of Q such that if $x \in V$ and (x^t) is a φ-sequence with $x^0 = x$, then $x^t \in U$, $t = 0, 1, 2, \cdots$.

DEFINITION 3.2. A point $x \in X$ is *stable* if $\{x\}$ is stable.

Remark 3.3. If $x \in X$ is stable, then $x \in E\varphi$. If Q is closed and stable, then $\varphi(Q) \subseteqq Q$; i.e., Q is an invariant set.

DEFINITION 3.4. Let $g(x) = (G_1(x), \cdots, G_m(x))$ be a vector of m real functions defined on X. A point a in R^m is called *Pareto-minimal* w.r.t. g if:

$$\text{there exists } x \in X \text{ such that } g(x) = a; \tag{3.1}$$

$$\text{if } y \in X \text{ and } G_i(y) \leqq a_i \text{ for all } i,\ i = 1, \cdots, m, \text{ then } g(y) = a. \tag{3.2}$$

DEFINITION 3.5. Let $g: X \to R^m$ and let a be a Pareto-minimal point w.r.t. g. The set

$$Nu(g, a) = \{x \in X | g(x) = a\}$$

is called the (generalized) *nucleolus*[2] of g w.r.t. a.

[2] Our definition generalizes that of Justman (1973). As we shall see in § 4, it also generalizes the nucleoli known in game theory.

DEFINITION 3.6. Let $g(x)=(G_1(x), \cdots, G_m(x))$ be defined on X. g is called *φ-monotone* if

$$x \in X \quad \text{and} \quad y \in \varphi(x) \Rightarrow G_i(x) \geqq G_i(y), \qquad i=1, \cdots, m. \tag{3.3}$$

g is called *strictly* φ-monotone if it is φ-monotone and satisfies, in addition,

$$x \in X,\ y \in \varphi(x) \quad \text{and} \quad y \neq x \Rightarrow G_k(x) > G_k(y) \quad \text{for some } k, \quad 1 \leqq k \leqq m, \quad k \equiv k(x, y). \tag{3.4}$$

φ-monotone vectorial functions will serve as Lyapunov functions.

Remark 3.7. *If g is strictly φ-monotone, then every nucleolus $Nu(g, a)$ is a subset of $E\varphi$.*

Our first result establishes, essentially, the stability of generalized nucleoli of φ-monotone l.s.c. vector functions.

THEOREM 3.8. *Let $g(x)=(G_1(x), \cdots, G_m(x))$ be defined on a compact metric space X. Let $Nu(g, a)$ be a generalized nucleolus. If the following conditions are satisfied*:

$$g \text{ is } \varphi\text{-monotone}; \tag{3.5}$$

$$G_i \text{ is l.s.c. on } X \text{ for } i=1, \cdots, m; \tag{3.6}$$

$$G_i \text{ is continuous on } Nu(g, a),\ i=1, \cdots, m; \tag{3.7}$$

then $Nu(g, a)$ is stable w.r.t. φ.

Remark 3.9. As $Nu(g, a)=\{x \in X | G_i(x) \leqq a_i,\ i=1, \cdots, m\}$, it follows from (3.6) that $Nu(g, a)$ is closed.

Proof of Theorem 3.8. Denote $Q=Nu(g, a)$. Let $U \supset Q$ be an open subset of X. $S=X-U$ is compact. If $y \in S$, then there exist $1 \leqq k \leqq m$ and a natural number r such that $G_k(y) > a_k + (1/r)$. Hence $y \in U_{k,r}$, where

$$U_{k,r}=\left\{x \in X \middle| G_k(x) > a_k + \frac{1}{r}\right\}.$$

As G_k is l.s.c., $U_{k,r}$ is open. As S is compact, there exists a finite collection $U_{k_1,r_1}, \cdots, U_{k_q,r_q}$ such that $S \subset \bigcup_{t=1}^{q} U_{k_t,r_t}$. For $1 \leqq t \leqq q$, let

$$V_{k_t,r_t}=\left\{x \in X \middle| G_{k_t}(x) < a_{k_t} + \frac{1}{r_t}\right\}.$$

As G_{k_t} is continuous on Q, V_{k_t,r_t} is a neighborhood of Q. Denote $V=\bigcap_{t=1}^{q} V_{k_t,r_t}$; then V is a neighborhood of Q which is contained in U. Now if $x \in V$ and (x^l) is a φ-sequence with $x^0=x$, then by (3.5) $G_{k_t}(x^l) < a_{k_t} + (1/r_t)$, $l=0, 1, 2, \cdots, 1 \leqq t \leqq q$. Hence $x^l \in V \subset U$, $l=0, 1, 2, \cdots$. This proves that Q is stable.

Remark 3.10. We can omit the requirement that X is compact if we know that, for some k_0, r_0,

$$\tilde{V}_{k_0,r_0} \equiv \left\{x \in X \middle| G_{k_0}(x) \leqq a_{k_0} + \frac{1}{r_0}\right\}$$

is compact. Indeed, instead of S, we can then cover $\tilde{V}_{k_0,r_0} - U$ by a finite number of $U_{k,r}$'s and then $V=\bigcap_{t=0}^{q} V_{k_t r_t}$ would be the required invariant neighborhood of Q which is contained in U.

The converse statement also holds:

THEOREM 3.11. *Let Q be a nonempty closed subset of X. If Q is stable, then there exists a (scalar)*[3] *function δ, continuous on Q and φ-monotone, such that $Q = Nu(\delta, 0)$.*

Proof. For $x \in X$, define

$$d(x) \equiv d(x, Q) \equiv \inf \{d(x, y) | y \in Q\},$$

and

$$\delta(x) \equiv \delta(x, Q) \equiv \sup \{d(x^t) | \text{where } (x^t) \text{ is a } \varphi\text{-sequence with } x^0 = x, t = 0, 1, 2, \cdots\}. \tag{3.8}$$

By (3.8) $\delta(x)$ is φ-monotone. Furthermore, since Q is invariant, (see Remark 3.3) $\delta(x) = 0$ iff $x \in Q$. Thus $Q = Nu(\delta, 0)$. To show that δ is continuous on Q, let $x \in Q$ and $\varepsilon > 0$. Let $U = \{z \in X | d(z) < \varepsilon\}$. Then U is open and $U \supset Q$. As Q is stable, there exists an open set V, $V \supset Q$, such that if $y \in V$ and (y^t) is a φ-sequence with $y^0 = y$, then $y^t \in U$, $t = 0, 1, 2, \cdots$. Thus $\delta(y) \leqq \varepsilon$ for $y \in V$. As $x \in Q$, $\delta(x) = 0$, and δ is continuous at x.

Remark 3.12. *If φ is* l.s.c., *then $\delta(x)$ (see* (3.8)) *is* l.s.c. The proof of Remark 3.12 is similar to that of Lemma 2.1; hence it will be omitted.

Remark 3.13. *If X is compact, Q is a nonempty closed invariant subset of X and $\delta(x)$ (see* (3.8)) *is continuous on Q, then Q is stable.* (Notice that δ need not be l.s.c. on X.) The proof of Remark 3.13 is straightforward and will be omitted.

DEFINITION 3.14. Let $g(x) = (G_1(x), \cdots, G_m(x))$ be defined on X. g is called *strongly* φ-monotone if it is φ-monotone and satisfies, in addition,

$$x \in X \quad \text{and} \quad y \in \varphi(x) \Rightarrow G_k(x) - G_k(y) \geqq d(x, y) \quad \text{for some } k,\ 1 \leqq k \leqq m,\ k \equiv k(x, y). \tag{3.9}$$

Remark 3.15. If g is strongly φ-monotone, then it is strictly φ-monotone.

THEOREM 3.16. *Let $g(x) = (G_1(x), \cdots, G_m(x))$ be strongly φ-monotone, and let $Nu(g, a)$ be a generalized nucleolus. If g satisfies also* (3.6) *and* (3.7), *then each point ξ in $Nu(g, a)$ is a stable point of φ.*

Proof. Let $\xi \in Nu(g, a)$. Define $g^*(x) = (G_1^*(x), \cdots, G_{m+1}^*(x))$ by $G_i^*(x) = G_i(x)$, $i = 1, \cdots, m$, and

$$G_{m+1}^*(x) = \sum_{i=1}^{m} G_i(x) + d(x, \xi). \tag{3.10}$$

By (3.9) and (3.10), g^* is φ-monotone. Furthermore, $G_i^*(x)$ is l.s.c. on X for $i = 1, \cdots, m+1$. Let $a^* \in R^{m+1}$ be defined by $a_i^* = a_i$, $i = 1, \cdots, m$, and $a_{m+1}^* = \sum_{i=1}^{m} a_i$. Then a^* is Pareto minimum for g^* and $\{\xi\} = Nu(g^*, a^*)$. By (3.7), $G_i^*(x)$ is continuous at ξ for $i = 1, \cdots, m+1$; consequently, by Theorem 3.8, ξ is stable.

Remark 3.17. The function g^* defined in the proof of Theorem 3.16 is actually strongly φ-monotone.

Under fairly general conditions a converse statement also holds; namely, every stable point of φ is a generalized nucleolus of a strongly φ-monotone

[3] We regard δ as a one-component vectorial function.

function which satisfies (3.6) and (3.7). We shall establish such a characterization in Corollary 3.23. In order to do so, let us proceed with the following definition.

DEFINITION 3.18. A function $\psi : X \to R^1$ is called a *valuation* of φ if:

$$\psi \text{ is continuous;} \tag{3.11}$$

$$x \in X \quad \text{and} \quad y \in \varphi(x) \Rightarrow \psi(x) - \psi(y) \geqq d(x, y). \tag{3.12}$$

Remark 3.19. *If X is compact and φ has a valuation ψ, then $f(x)$ is bounded on X.* The proof is an immediate consequence of Lemma 2.6.

Remark 3.20. Lower semicontinuity of φ (or of f) is *not* implied by the existence of a valuation ψ for φ. To see this let $X = [0, 1]$, and

$$\varphi(x) = \begin{cases} \{x\}, & 0 \leqq x < \frac{1}{2}, \\ \{x, x - \frac{1}{2}\}, & \frac{1}{2} \leqq x \leqq 1. \end{cases}$$

Then $\psi(x) = x$ is a valuation for φ, while φ is not l.s.c.

Remark 3.21. *If X is compact and φ has a valuation ψ, then $E\varphi \neq \varnothing$.*

Proof. Let $y \in X$ satisfy $\psi(y) \leqq \psi(x)$ for all $x \in X$. By (3.12) $y \in E\varphi$.

THEOREM 3.22. *Let $x \in E\varphi$. If f* (see (2.2)) *is continuous at x, then x is stable. If x is stable and φ has a valuation ψ, then f is continuous at x.*

Proof. Assume that f is continuous at x. Let $\varepsilon > 0$ be given. Choose $0 < \delta < \varepsilon/2$ such that if $y \in X$ and $d(x, y) < \delta$, then $f(y) < \varepsilon/2$. Thus if $y \in X$, $d(x, y) < \delta$ and (y^t) is a φ-sequence with $y^0 = y$, then $d(y, y^t) \leqq f(y) < \varepsilon/2$. Hence $d(y^t, x) \leqq d(y^t, y) + d(y, x) < \varepsilon$, $t = 0, 1, 2, \cdots$. Thus x is stable. To prove the second part of the theorem let ψ be a valuation of φ and assume that x is stable. Let $\varepsilon > 0$ be given. There exists a $\delta_1 > 0$ such that if $y \in X$ and $d(y, x) < \delta_1$, then $|\psi(y) - \psi(x)| < \varepsilon/2$. As x is stable, there exists a $\delta > 0$ such that if $y \in X$, $d(y, x) < \delta$ and (y^t) is a φ-sequence with $y^0 = y$, then $d(y^t, x) < \delta_1$, $t = 0, 1, 2, \cdots$. Now if $y \in X$, $d(y, x) < \delta$ and (y^t) is a φ-sequence with $y^0 = y$, then for $T \geqq 0$,

$$\sum_{t=0}^{T} d(y^{t+1}, y^t) \leqq \sum_{t=0}^{T} \{\psi(y^t) - \psi(y^{t+1})\}$$

$$= \psi(y) - \psi(y^{(T+1)}) < \psi(x) + \frac{\varepsilon}{2} - \left(\psi(x) - \frac{\varepsilon}{2}\right) = \varepsilon.$$

Thus $f(y) \leqq \varepsilon$. Since $f(x) = 0$ (as $x \in E\varphi$, see Remark 3.3) this proves the continuity of f at x.

COROLLARY 3.23. *Let X be compact and let φ have a valuation. If ξ is a stable point of φ, then there exists a (scalar) function g, continuous at ξ and strongly φ-monotone, such that $\{\xi\} = Nu(g, 0)$. If, in addition, φ is* l.s.c., *then g is* l.s.c.

Proof. Choose $g(x) = 2f(x) + d(x, \xi)$ for $x \in X$.

We close this section with the following, somewhat surprising, result.

THEOREM 3.24. *Let X be compact. Assume that φ has a valuation ψ and that φ is lower semicontinuous. If Q is closed and stable, then Q contains a stable point.*

Proof. As Q is stable, $Q \neq \varnothing$ (see Definition 3.1). As Q is a closed subset of X it is compact; hence ψ takes its minimum on Q, say at $\xi \in Q$. Form the function

$$g(x) = (\delta(x), \psi(x) + d(x, \xi)) \qquad \text{(see (3.8)).}$$

Let $a=(0, \psi(\xi))$. Then $\{\xi\}=Nu(g, a)$. By Theorem 3.11, δ is continuous on Q, and, in particular, at ξ. By Remark 3.12, δ is l.s.c. on X. $\psi(x)+d(x, \xi)$ is continuous and φ-monotone (see (3.12)). Hence by Theorem 3.8, ξ is stable.

4. Applications to the stability theory of the bargaining sets. We consider a *game* $(N; v)$ described by a set of *players* $N=\{1, \cdots, n\}$, and a *characteristic function* $v: 2^N \to R^1$. We assume that v is 0–1 normalized, i.e., $v(\{i\})=0$, $i \in N$, and $v(N)=1$.

Denote:

$$X(N)=\left\{x \mid x \in R^n, x_i \geqq 0, i \in N \text{ and } \sum_{i \in N} x_i = 1\right\}. \tag{4.1}$$

For $x \in X(N)$ and a *coalition* S (namely, a subset of N) we define the *excess* $e(S, x)$ of S w.r.t. x to be

$$e(S, x)=v(S)-\sum_{i \in S} x_i. \tag{4.2}$$

For each ordered pair $i, j \in N$, $i \neq j$, and for each $x \in X(N)$, we define the *maximum surplus* $s_{ij}(x)$ of i against j w.r.t. x to be

$$s_{ij}(x)=\max \{e(S, x) \mid i \in S, j \notin S\}. \tag{4.3}$$

For $x \in X(N)$ and $i, j \in N$, $i \neq j$, we denote

$$k_{ij}(x)=\max (0, \min (x_j, \tfrac{1}{2}(s_{ij}(x)-s_{ji}(x)))). \tag{4.4}$$

By a *demand function* $D=\{d_{ij}\}$, we mean a collection of functions $d_{ij}: X(N) \to R^1$, one for each pair $i, j \in N$, $i \neq j$, which satisfy

$$0 \leqq d_{ij}(x) \leqq k_{ij}(x); \tag{4.5}$$

(4.6) $d_{ij}(x)$ is l.s.c. on $X(N)$ (when, e.g., $X(N)$ is considered as a subset of l_∞^n).

Remark 4.1. All the following results of this section could be obtained for a compact subset X of l_∞^n, instead of $X(N)$. However, in order to keep the presentation as simple as possible, we restrict ourselves to $X(N)$.

Let D be a demand function. Following Stearns (1968), we define the *bargaining set* M_D by

$$M_D=\{x \in X(N) \mid d_{ij}(x)=0 \text{ for all } i, j \in N, i \neq j\}. \tag{4.7}$$

The *kernel* is equal to $K=M_{D^*}$, where $D^*=\{k_{ij}\}$.

Let $D=\{d_{ij}\}$ be a demand function and $i, j \in N$, $i \neq j$. Let $x \in X(N)$ and $\alpha \geqq 0$. We say that y results from x by a *D-bounded transfer* (of size α from j to i) if:

$$y_i = x_i + \alpha; \tag{4.8}$$

$$y_j = x_j - \alpha; \tag{4.9}$$

$$y_l = x_l, \qquad l \neq i, j; \tag{4.10}$$

$$\alpha \leqq d_{ij}(x). \tag{4.11}$$

As $\alpha \leqq d_{ij}(x) \leqq k_{ij}(x) \leqq x_j$, it follows that $y \in X(N)$. Also every D-bounded transfer is a D^*-bounded transfer.

For each demand function D, we define the dynamic system φ_D on $X(N)$ by:

$$\varphi_D(x) = \{y | y \text{ results from } x \text{ by a } D\text{-bounded transfer}\}. \tag{4.12}$$

Remark 4.2. *φ_D is* l.s.c. *$M_D = E\varphi_D$. Also $\varphi_D(x) \subset \varphi_{D^*}(x)$ for all $x \in X(N)$.* The proof of Remark 4.2 is straightforward; hence, it will be omitted.

For $x \in X(N)$, let $\theta(x)$ be the vector in R^{2^n} whose components are the numbers $e(S, x)$, $S \subseteq N$, arranged in nonincreasing order.

DEFINITION 4.3. The *nucleolus point* of a game $(N; v)$ is a point x in $X(N)$ for which $\theta(x)$ is lexicographically minimum, i.e., the point x in $X(N)$ such that $\theta(x)_L \leqq \theta(y)$ for all y in $X(N)$. Here ${}_L\leqq$ is the lexicographic order in R^{2^n}.

The nucleolus point was introduced in Schmeidler (1969), who proved that there exists exactly one such point.

We now introduce two strongly φ_{D^*}-monotone functions. The first one is related to the nucleolus point.

Example 4.4. For $x \in X(N)$ let $\theta(x)$ be, as before, the vector in R^{2^n} whose components are the numbers $e(S, x)$, $S \subseteq N$, arranged in nonincreasing order. Define $g(x) = (G_1(x), \cdots, G_{2^n}(x))$ by

$$G_k(x) = \sum_{t=1}^{k} 2^{k-t}\theta_t(x), \qquad k = 1, \cdots, 2^n. \tag{4.13}$$

Claim 4.5. *If a is the lexicographic minimum of $g(x)$ on $X(N)$, then $Nu(g, a)$ consists exactly of the nucleolus point.*

Proof. Let $\nu \in Nu(g, a)$. Then $g(\nu) {}_L\leqq g(x)$ for all $x \in X(N)$. Hence it follows from (4.13) that $\theta(\nu) {}_L \leqq \theta(x)$ for all $x \in X(N)$. Hence by definition, ν is the nucleolus point of $(N; v)$.

Claim 4.6. *$g(x)$* (see (4.13)) *is strongly φ_{D^*}-monotone.*

Proof. Let $x \in X(N)$ and let $y \in \varphi_{D^*}(x)$. There exist $i, j \in N$, $i \neq j$, such that y results from x by a D^*-bounded transfer from j to i. If the size of the transfer is 0, then, clearly, $G_k(x) - G_k(y) = \|x - y\| = 0$, $k = 1, \cdots, 2^n$. Assume that it is positive. For $z \in R^n$, denote $\|z\| = \max_{1 \leqq t \leqq n} |z_t|$. Let S be a coalition such that $i \in S$, $j \notin S$ and $s_{ij}(x) = e(S, x)$. Clearly, $e(S, x) - e(S, y) = \|x - y\|$ and for each R, $R \subseteq N$, $e(R, x) - e(R, y) \geqq -\|x - y\|$.

We can enumerate all coalitions in such a way that $\theta(y) = (e(R_1, y), \cdots, e(R_{2^n}, y))$ and if $S = R_q$, then either $q = 1$ or $e(R_l, y) > e(S, y)$ whenever $l < q$.

By (4.4) and (4.5), $s_{ij}(y) = e(S, y) \geqq s_{ji}(y)$; therefore, all coalitions R_l, $l < q$, either contain $\{i, j\}$ or are disjoint from $\{i, j\}$; consequently,

$$e(R_l, y) = e(R_l, x) \quad \text{whenever } l < q. \tag{4.14}$$

Let $1 \leqq k \leqq 2^n$; then

$$G_k(x) \geqq \sum_{t=1}^{k} 2^{k-t} e(S_t, x), \tag{4.15}$$

where $(e(S_1, x) \cdots e(S_k, x))$ is any permutation of the first k components of $\theta(x)$. (See (4.13) and Hardy, Littlewood and Polya (1933, p. 261).) Consequently,

$$G_k(x) \geqq \sum_{t=1}^{k} 2^{k-t} e(R_t, x), \tag{4.16}$$

because $(R_1, R_2, \cdots, R_k)$ is obtained from a certain permutation $(S_1, S_2, \cdots, S_k)$ by replacing some coalitions by coalitions of lower excesses at x.

Thus

$$G_k(x)-G_k(y)\geqq \sum_{t=1}^{k} 2^{k-t}(e(R_t, x)-e(R_t, y)). \tag{4.17}$$

It follows from (4.14) that $G_k(x)-G_k(y)=0$ for $k<q$ and if $k\geqq q$, then

$$G_k(x)-G_k(y)\geqq 2^{k-q}\|x-y\|-\sum_{t=q+1}^{k} 2^{k-t}\|x-y\|=\|x-y\|. \tag{4.18}$$

In particular, $G_{2^n}(x)-G_{2^n}(y)\geqq\|x-y\|$ and g is strongly φ_{D^*}-monotone.

COROLLARY 4.7. *$G_{2^n}(x)$ is a valuation function for each φ_D.*

Proof. If $x\in X(N)$ and $y\in\varphi_D(x)$, then, by (4.5), $y\in\varphi_{D^*}(x)$. Hence $G_{2^n}(x)-G_{2^n}(y)\geqq\|x-y\|$. Clearly, $G_{2^n}(x)$ is a continuous function.

COROLLARY 4.8 (Kalai, Maschler and Owen (1973)). *The nucleolus is a stable point of every bargaining set M_D.*

Proof. Let D be a demand function. By Claim 4.6, g (see (4.13)) is φ_{D^*}-monotone. Hence because $\varphi_D(x)\subset\varphi_{D^*}(x)$ for all $x\in X(N)$, g is φ_D-monotone. Clearly, g is continuous on X. By Claim 4.5, the nucleolus is a generalized nucleolus of g. Hence by Theorem 3.8, the nucleolus is a stable point w.r.t. φ_D; i.e., a stable point of M_D.

Example 4.9. For $x\in X(N)$ let $\theta^*(x)$ be the vector in $R^{n(n-1)}$ whose components are the numbers $s_{ij}(x)$, $i, j\in N$, $i\neq j$, arranged in nonincreasing order. The lexicographic kernel of $(N; v)$, $LK(N; v)$, is defined by

$$LK(N; v)=\{x\in X(N)\,|\,\theta^*(x)\ {}_L\!\geqq\ \theta^*(y)\quad \text{for all } y\in X(N)\}, \tag{4.19}$$

where ${}_L\!\geqq$ denotes here the lexicographic order of $R^{n(n-1)}$. It was first suggested by Gill Kalai.[4] As for the nucleolus, we define $g^*(x)=(G_1^*(x), \cdots, G_{n(n-1)}^*(x))$ by:

$$G_k^*(x)=\sum_{t=1}^{k} 2^{k-t}\theta_t^*(x), \qquad k=1, \cdots, n(n-1). \tag{4.20}$$

The following claims hold:

Claim 4.10. *Let a^* be the lexicographic minimum of $g^*(x)$ on $X(N)$. Then $LK(N; v)=Nu(g^*, a^*)$.*

Claim 4.11. *$g^*(x)$ is strongly φ_{D^*}-monotone.*

Claim 4.12. *$G_{n(n-1)}^*(x)$ is a valuation function for each φ_D.*

Claim 4.13. *Each point of $LK(N; v)$ is a stable point of every bargaining set M_D.*

The proofs of Claims 4.10–4.13 are similar to those of Claims 4.5 and 4.6 and Corollaries 4.7 and 4.8, respectively; hence they will be omitted.

We now present an example of a game for which the nucleolus point *does not* belong to the lexicographic kernel.

Example 4.14. Let $N=\{1, 2, 3, 4, 5, 6, 7, 8\}$. Let $v(N)=1$, $v(\{1, 2, 3, 4\})=v(\{5, 6, 7, 8\})=v(\{1, 2, 5, 6\})=v(\{3, 4, 7, 8\})=100$, $v(\{1, 2, 3, 4, 5, 6\})=10$, $v(S)=1$ if S is a two-player coalition, and $v(S)=0$ otherwise. Then the nucleolus

[4] Written communication.

point of $(N; v)$, $Nu(N; v) = (0, 0, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0)$, while $LK(N; v) = \{(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})\}$.

We close this section with the observation that every nonempty strong ε-core is a stable set w.r.t. φ_{D^*}; hence, every nonempty strong ε-core is a stable set of every bargaining set M_D.

DEFINITION 4.15. Let $(N; v)$ be a game and ε be a real number. The *strong ε-core* C_ε of $(N; v)$ is defined by[5]

$$C_\varepsilon = \{x \in X(N) | e(S, x) \leqq \varepsilon \text{ for all } S \subset N, S \neq \varnothing, N\}. \tag{4.21}$$

THEOREM 4.16. *Every nonempty strong ε-core is stable* w.r.t. φ_{D^*}.

Proof. Let $C_\varepsilon \neq \varnothing$. Define, for $x \in X(N)$, the (scalar) function

$$\bar{g}(x) = \max [\max \{e(S, x) - \varepsilon | S \subset N, S \neq \varnothing, N\}, 0].$$

Then $\bar{g}(x)$ is continuous and φ_{D^*}-monotone. Indeed, $\max \{e(S, x) - \varepsilon | S \subset N, S \neq \varnothing, N\} = \max \{s_{ij}(x) | i, j \in N, i \neq j\} - \varepsilon$ (see (4.3)). By Claim 4.11, it is nonincreasing along trajectories. The same property certainly is shared by the constant function 0; therefore $\bar{g}(x)$ is φ_{D^*}-monotone.

Furthermore, $C_\varepsilon = Nu(\bar{g}(x), 0)$; hence by Theorem 3.8, C_ε is stable w.r.t. φ_{D^*}.

REFERENCES

L. J. BILLERA (1972), *Global stability in n-person games*, Trans. Amer. Math. Soc., 172, pp. 45–56.

G. H. HARDY, J. E. LITTLEWOOD AND G. POLYA (1933), *Inequalities*, Cambridge University Press, London.

H. HERMES (1970), *The generalized differential equation $\dot{x} \in R(t, x)$*, Advances in Math., 4, pp. 149–169.

M. JUSTMAN (1973), *Regulative frameworks for iterative negotiations*, Institute of Mathematics, The Hebrew University, Jerusalem; Internat. J. Game Theory, to appear.

G. KALAI, M. MASCHLER AND G. OWEN (1973), *Asymptotic stability and other properties of trajectories and transfer sequences leading to bargaining sets*, Rep. Department of Operations Research, Stanford University; Internat. J. Game Theory, to appear.

D. SCHMEIDLER (1969), *The nucleolus of a characteristic function game*, SIAM J. Appl. Math., 17, pp. 1163–1170.

R. E. STEARNS (1968), *Convergent transfer schemes for n-person games*, Trans. Amer. Math. Soc., 134, pp. 449–459.

W. I. ZANGWILL (1969), *Non-linear Programming*, Prentice-Hall, Englewood Cliffs, N.J.

[5] The customary definition requires in (4.16) $x(N) = v(N)$ instead of $x \in X(N)$. With the customary definition, Theorem 4.16 still holds if one replaces "strong ε-core" by "the intersection of the strong ε-core with $X(N)$".

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 6, November 1976

INVARIANTS AND CANONICAL FORMS UNDER DYNAMIC COMPENSATION*

W. A. WOLOVICH AND P. L. FALB†

Abstract. This paper is concerned with the development of a complete abstract invariant as well as a set of canonical forms under dynamic compensation for linear systems characterized by proper, rational transfer matrices. More specifically, it is shown that one can always associate with any proper rational transfer matrix, $T(s)$, a special lower left triangular matrix, $\xi_T(s)$, called the interactor. This matrix is then shown to represent an abstract invariant under dynamic compensation which, together with the rank of $T(s)$, represents a complete abstract invariant. A set of canonical forms under dynamic compensation is also developed along with appropriate dynamic compensation.

1. Introduction. The primary purpose of this paper will be to exhibit invariants and a set of canonical forms for linear dynamical systems which are equivalent under dynamic compensation. Critical to this purpose will be the introduction of a special lower triangular matrix $\xi_T(s)$ associated with any $T(s)$ in S and called the "interactor" of T. The role of the interactor in resolving the closely allied questions of exact model matching and inverse systems is also displayed.

Section 2 contains a precise description of equivalence under dynamic compensation as well as some elementary properties of this notion. The interactor $\xi_T(s)$ is introduced in § 3 and is shown to be an abstract invariant under dynamic compensation in § 4. The main results on invariants and canonical forms are also established in § 4 and some final observations are made in § 5.

2. Dynamic compensation. We begin with some definitions.

DEFINITION 2.1. Let S be the set of all proper $(p \times m)$ transfer matrices of full rank $r(=\min\{p, m\})$ with first r rows, $T_r(s)$, of rank r. Let S_+, S_- be the subsets of S given by

$$S_+ = \{T(s) \in S \mid T(s) \quad \text{is } p \times m \quad \text{with } m-p \geqq 0\},$$
$$S_- = \{T(s) \in S \mid T(s) \quad \text{is } p \times m \quad \text{with } m-p < 0 \text{ and } T_m(s) \text{ of rank } m\},$$

respectively.

We observe that $S_+ \cap S_- = \varnothing$ and that $S_+ \cup S_- = S$.

DEFINITION 2.2. Let $T(s)$ be a given $p \times m$ element of S. Then, any $m \times k$ transfer matrix $T_c(s)$ in S is called a *dynamic compensator of* $T(s)$.

The operation of $T_c(s)$ can be represented in "open loop" form by the block diagram of Fig. 1, where $y(s) = T(s)\,u(s)$ represents the (Laplace transform of the) zero state dynamical behavior of the given system and $u(s) = T_c(s)\,v(s)$ that of the compensator.

Since $T(s)T_c(s)$ is again a proper transfer matrix, it is readily shown that $T_c(s)$ can be represented in terms of input dynamics and state feedback [1], [2]. More

* Received by the editors September 5, 1974, and in revised form August 20, 1975.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This work was supported in part by the Air Force Office of Scientific Research under AFOSR 71-2078C and in part by the National Science Foundation under Eng. 73-03846A01.

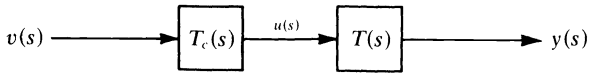


FIG. 1

precisely, if $T(s)=R(s)P^{-1}(s)$ with $R(s)$ and $P(s)$ relatively right prime polynomial matrices [3], then there are polynomial matrices $F(s)$, $G(s)$ and $L(s)$ such that $T_c(s)=P(s)P_c^{-1}(s)L(s)$ is proper with $P_c(s)=G(s)P(s)-F(s)$. Thus, the operation of $T_c(s)$ can be represented in "closed loop" form by the block diagram of Fig. 2, with $z(s)$ the (Laplace transform of the) partial state of the given system, $F(s)z(s)$ the state feedback "part", and $G(s)^{-1}L(s)$ the input dynamic "part" of

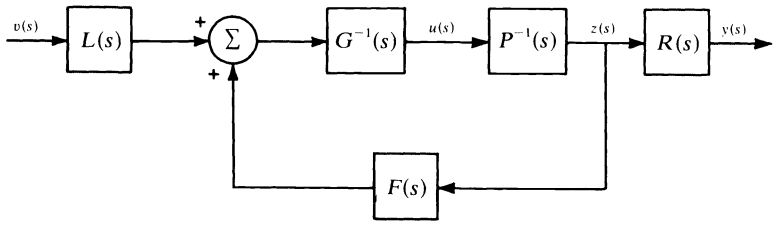


FIG. 2

$T_c(s)$ (see [1], [2]). Knowledge of the portions of a dynamical system which remain unaltered (or invariant) under a particular form of compensation is ultimately tied to a number of important questions of control system analysis and synthesis such as model matching and decoupling.

DEFINITION 2.3. If $T_1(s)$ and $T_2(s)$ are elements of S, then $T_1(s)$ and $T_2(s)$ are equivalent under dynamic compensation if

$$\begin{aligned} T_1(s)T_{1c}(s) &= T_2(s), \\ T_2(s)T_{2c}(s) &= T_1(s) \end{aligned} \tag{2.4}$$

for some $T_{1c}(s)$ and $T_{2c}(s)$ in S.

If $T_1(s)$ and $T_2(s)$ are equivalent under dynamic compensation, we write $T_1(s)E_dT_2(s)$. It is clear that E_d defines an equivalence relation on S. The main purpose of this paper is the characterization of the orbits of this equivalence relation by the determination of invariants and a set of canonical forms.

The following elementary observations are required because we deal with proper transfer matrices of different dimensions.

OBSERVATION 2.5. *If $T_1(s)E_dT_2(s)$, then $T_1(s)$ and $T_2(s)$ have the same rank.*

OBSERVATION 2.6. *S_+ and S_- are stable under dynamic compensation (i.e., if $T_1(s)\in S_+$, say, and $T_1(s)E_dT_2(s)$, then $T_2(s)\in S_+$).*

OBSERVATION 2.7. *If $T_1(s)$ is a $p \times m$ element of S_- and $T_1(s)E_dT_2(s)$, then $T_2(s)$ is also a $p \times m$ element of S_-, and both $T_{1c}(s)$ and $T_{2c}(s)$ are nonsingular.*

3. The interactor $\xi_T(s)$. Let $T(s)$ be an element of S. We shall, in this section, determine a unique nonsingular, lower left triangular polynomial matrix $\xi_T(s)$ associated with $T(s)$ and called the interactor of $T(s)$. The constructive procedure which will be outlined is similar to that given by Silverman [4] in the case of state space representations, although unlike Silverman's algorithm, it does yield a unique $\xi_T(s)$ for transfer matrix representations.

LEMMA 3.1. *Let $T(s)$ be an $m \times m$ element of S. Then there is a unique, nonsingular $(m \times m)$, lower left triangular polynomial matrix $\xi_T(s)$ of the form*

$$\xi_T(s) = H_T(s)\,\mathrm{diag}\,[s^{f_1}, \cdots, s^{f_m}], \tag{3.2}$$

where

$$H_T(s) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ h_{21}(s) & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ & & \cdots & 0 \\ h_{m1}(s) & h_{m2}(s) & \cdots & 1 \end{bmatrix} \tag{3.3}$$

and $h_{ij}(s)$ is divisible by s (or is zero) such that

$$\lim_{s\to\infty} \xi_T(s)T(s) = K_T \tag{3.4}$$

with K_T nonsingular.

Proof. We first prove the existence of such a $\xi_T(s)$. It is well-known [2] that $T(s)$ can always be factored as the product $R(s)P^{-1}(s)$ with $R(s)$, $P(s)$ relatively right prime polynomial matrices and $P(s)$ column proper. Let $\partial_i(P) = d_i$, $i = 1, \cdots, m$, denote the column degrees of $P(s)$ and let $\sum_{i=1}^{m} d_i = n$. Now, det $R(s)$ is a nonzero polynomial of degree q (since $T(s)$ is nonsingular) with $q \leqq n$ (since $T(s)$ is proper). There are unique integers μ_i, $i = 1, \cdots, m$, such that

$$\lim_{s\to\infty} s^{\mu_i} T_i(s) = \tau_i, \qquad i = 1, \cdots, m, \tag{3.5}$$

where $T_i(s)$ is the ith row of $T(s)$ and τ_i is both finite and nonzero. We define the first row $\xi_T(s)_1$ of $\xi_T(s)$ by

$$\xi_T(s)_1 = (s^{\mu_1}, 0, \cdots, 0) \tag{3.6}$$

so that

$$\lim_{s\to\infty} \xi_T(s)_1 T(s) = \xi_1 = \tau_1. \tag{3.7}$$

If τ_2 is linearly independent of ξ_1, then we set

$$\xi_T(s)_2 = (0, s^{\mu_2}, 0, \cdots, 0) \tag{3.8}$$

so that

$$\lim_{s\to\infty} \xi_T(s)_2 T(s) = \xi_2 = \tau_2. \tag{3.9}$$

On the other hand, if τ_2 and ξ_1 are linearly dependent so that $\tau_2 = \alpha_1^1 \xi_1$ with $\alpha_1^1 \neq 0$, then we let

$$\tilde{\xi}_T^1(s)_2 = s^{\mu_2^1}[(0, s^{\mu_2}, 0, \cdots, 0) - \alpha_1^1 \xi_T(s)_1], \tag{3.10}$$

where μ_2^1 is the unique integer for which $\lim_{s\to\infty} \tilde{\xi}_T^1(s)_2 T(s) = \tilde{\xi}_2^1$ is both finite and nonzero. If $\tilde{\xi}_2^1$ is linearly independent of ξ_1, then we set

$$\xi_T(s)_2 = \tilde{\xi}_T^1(s)_2 \tag{3.11}$$

and note that

$$\lim_{s\to\infty} \xi_T(s)_2 T(s) = \tilde{\xi}_2^1 \tag{3.12}$$

is linearly independent of ξ_1. If not, then $\tilde{\xi}_2^1 = \alpha_1^2 \xi_1$ and we let

$$\tilde{\xi}_T^2(s)_2 = s^{\mu_2^2}[\tilde{\xi}_T^1(s)_2 - \alpha_1^2 \xi_T(s)_1], \tag{3.13}$$

where μ_2^2 is the unique integer for which $\lim_{s\to\infty} \tilde{\xi}_T^2(s)_2 T(s) = \tilde{\xi}_2^2$ is both finite and nonzero. If $\tilde{\xi}_2^2$ and ξ_1 are linearly independent, then we set $\xi_T(s)_2 = \tilde{\xi}_T^2(s)_2$ and if not, we repeat the procedure until either linear independence is obtained or $\mu_1 + \mu_2^k = n - q$.[1] In case $\mu_1 + \mu_2^k = n - q$, set $f_3 = 0, \cdots, f_m = 0$ and the corresponding $h_{ij} = 0$. The remaining rows of $\xi_T(s)$ are defined recursively in an entirely analogous manner. In other words, we obtain either (i) a matrix $\xi_T(s)$ of the form (3.2) such that (3.4) is satisfied or (ii) $\xi_T(s)_1, \cdots \xi_T(s)_r$, $r \leqq m$, such that $\lim_{s\to\infty} \cdot \xi_T(s)_j T(s) = \xi_j$ with $\xi_1, \cdots, \xi_r$ linearly independent and $\sum_1^r f_i = n - q$. In case (ii), we set $f_{r+1} = 0, \cdots, f_m = 0$ and the corresponding $h_{ij} = 0$ to obtain $\xi_T(s)$. *If* $r = m$, then

$$\lim_{s\to\infty} \xi_T(s) T(s) = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_m \end{bmatrix} = K_T \tag{3.14}$$

is nonsingular since the ξ_i are linearly independent. If $r < m$, then $\xi_T(s) T(s) = \xi_T(s) R(s) P^{-1}(s)$ is a proper transfer matrix as each step produces a proper transfer matrix. But then

$$\partial_i(\xi_T R) = \bar{d}_i \leqq d_i = \partial_i(P) \tag{3.15}$$

so that

$$\sum_{i=1}^m \partial_i(\xi_T R) = \sum_{i=1}^m \bar{d}_i \leqq \sum d_i = n.$$

However,

$$\begin{aligned} &\text{degree } (\det \xi_T R) = \text{degree } (\det \xi_T) = \text{degree } (\det R) \\ &\qquad = n - q + q, \end{aligned} \tag{3.16}$$

[1] Note $\mu_1 + \mu_2^k$ cannot exceed $n - q$ as $\xi_1, \tilde{\xi}_2^k$ are finite and nonzero.

which implies $\sum_{i=1}^{m} \partial_i(\xi_T R) = \sum_{i=1}^{m} \bar{d}_i \geqq n$. It follows that $\bar{d}_i = d_i$ and hence, that $\xi_T(s)R(s)$ is column proper with the same column degrees as $P(s)$. Thus,

$$\lim_{s\to\infty} \xi_T(s)R(s)P^{-1}(s) = K_T \tag{3.17}$$

is nonsingular [2].

We now prove uniqueness. Let $\xi_T(s) = H_T(s)\,\mathrm{diag}\,[s^{f_1}, \cdots, s^{f_m}]$ and $\hat{\xi}_T(s) = \hat{H}_T(s)\,\mathrm{diag}\,[s^{\hat{f}_1}, \cdots, s^{\hat{f}_m}]$ satisfy (3.4). Then, $\xi_T R$ and $\hat{\xi}_T R$ are column proper with $\partial_i(\xi_T R) = \partial_i(P) = \partial_i(\hat{\xi}_T R)$. It follows [2] that

$$[\xi_T R][\hat{\xi}_T R]^{-1} = H_T(s)\,\mathrm{diag}\,[s^{f_1-\hat{f}_1}, \cdots, s^{f_m-\hat{f}_m}]\hat{H}_T^{-1}(s),$$

$$[\hat{\xi}_T R][\xi_T R]^{-1} = \hat{H}_T(s)\,\mathrm{diag}\,[s^{\hat{f}_1-f_1}, \cdots, s^{\hat{f}_m-f_m}]H_T^{-1}(s)$$

are both proper. Since $H_T(s)$ and $\hat{H}_T(s)$ are of the form (3.3), $f_i = \hat{f}_i$ for $i = 1, \cdots, m$. Now, both $H_T(s)$ and $\hat{H}_T(s)$ are unimodular, lower left triangular matrices with diagonal entries 1. Moreover, $H_T(s)\hat{H}_T^{-1}(s) = U(s)$ is unimodular, proper and satisfies $\lim_{s\to\infty} U(s) = L$ with L nonsingular. Since each $h_{ij}(s)$ is divisible by s (or is zero) and each $\hat{h}_{ij}(s)$ is divisible by s (or is zero), $U(s) = I$ and $H_T(s) = \hat{H}_T(s)$. The proof of the lemma is now complete.

LEMMA 3.18. *Let $T(s)$ be a $p \times m$ element of S_+ with $p < m$. Then there is a unique $p \times p$ matrix $\xi_T(s)$ of the form*

$$\xi_T(s) = H_T(s)\,\mathrm{diag}\,[s^{f_1}, \cdots, s^{f_p}], \tag{3.19}$$

where

$$H_T(s) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ h_{21}(s) & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ h_{p1}(s) & h_{p2}(s) & \cdots & 1 \end{bmatrix} \tag{3.20}$$

and $h_{ij}(s)$ is divisible by s (or is zero) such that

$$\lim_{s\to\infty} \xi_T(s)T(s) = K_T \tag{3.21}$$

with K_T of rank p.

Proof. Let $T(s) = R(s)P^{-1}(s)$ with $R(s)$, $P(s)$ relatively right prime and $P(s)$ column proper. Then the $p \times m$ matrix $R(s)$ is of rank p and there are row vectors $r_{p+1}, \cdots, r_m$ (with polynomial entries) such that

$$R_e(s) = \begin{bmatrix} R \\ r_{p+1} \\ \vdots \\ r_m \end{bmatrix} \tag{3.22}$$

is nonsingular and $T_e(s) = R_e(s)P^{-1}(s)$ is proper (i.e., is an element of S). By virtue of Lemma 3.1, there is a $\xi_{T_e}(s)$ of the form (3.2) such that $\lim_{s\to\infty} \xi_{T_e}(s)T_e(s) = K_e$

is nonsingular. Let

$$\xi_{T_e}(s)=\left[\begin{array}{c|c}\xi_T(s) & O_{p,m-p}\\ \hline X_{m-p,p} & X_{m-p,m-p}\end{array}\right] \tag{3.23}$$

where $\xi_T(s)$ is a $p\times p$ matrix and $X_{i,j}$ is an $i\times j$ matrix. Then $\xi_T(s)$ is necessarily of the form (3.19) and

$$\lim_{s\to\infty}\xi_T(s)R(s)P^{-1}(s)=K_T \tag{3.24}$$

with K_T of rank p. The uniqueness of $\xi_{T_e}(s)$ implies that $\xi_T(s)$ is unique.

LEMMA 3.25. *Let $T(s)$ be a $p\times m$ element of S, and let $T_m(s)$ denote the nonsingular matrix consisting of the first m rows of $T(s)$. Then there is a unique $p\times p$ matrix $\xi_T(s)$ of the form*

$$\xi_T(s)=\begin{bmatrix}\xi_{T_m}(s) & 0\\ -\gamma_1(s) & \gamma_2(s)\end{bmatrix} \tag{3.26}$$

where $\gamma_1(s)$, $\gamma_2(s)$ are relatively left prime and $\gamma_2(s)$ is a nonsingular lower left triangular matrix in Hermite normal form [3] *with monic diagonal entries such that*

$$\lim_{s\to\infty}\xi_T(s)T(s)=K_T \tag{3.27}$$

with K_T a constant matrix of rank m whose final $p-m$ rows are zero.

Proof. Let $T(s)=R(s)P^{-1}(s)$ with $R(s)$, $P(s)$ relatively right prime. Then

$$R(s)=\begin{bmatrix}R_m(s)\\ R_{p-m}(s)\end{bmatrix} \tag{3.28}$$

so that $T_m(s)=R_m(s)P^{-1}(s)$. Since $T_m(s)$ is nonsingular, $R_m(s)$ is nonsingular. It follows that there is a (unique) pair of polynomial matrices $\gamma_1(s)$, $\gamma_2(s)$ such that

$$R_{p-m}(s)R_m^{-1}(s)=\gamma_2^{-1}(s)\gamma_1(s), \tag{3.29}$$

where $\gamma_1(s)$, $\gamma_2(s)$ are relatively left prime polynomial matrices and $\gamma_2(s)$ is a nonsingular lower left triangular matrix in Hermite normal form with monic diagonal entries. However, (3.29) implies that

$$\gamma_2(s)R_{p-m}(s)-\gamma_1(s)R_m(s)=0. \tag{3.30}$$

Since $\xi_{T_m}(s)$ is unique by Lemma 3.1 and $\gamma_2^{-1}(s)\gamma_1(s)$ represents a unique factorization (since the Hermite normal form of $\gamma_2(s)$ is unique) of $R_{p-m}(s)R_m^{-1}(s)$, the matrix $\xi_T(s)$ exists and is unique.

DEFINITION 3.31. If $T(s)$ is an element of S, then $\xi_T(s)$ is called the *interactor* of $T(s)$.

We note that the interactor is defined for all proper transfer matrices in S.

We illustrate the construction of $\xi_T(s)$ in the following two examples.

Example 3.32. Let

$$T(s)=\begin{bmatrix}\dfrac{1}{s+1} & \dfrac{1}{s+2} & \dfrac{1}{s+3}\\ 0 & \dfrac{1}{s+3} & 1\end{bmatrix}.$$

Then $f_1 = 1$, $f_2 = 0$ and $\xi_T(s)_1 = (s \quad 0)$ so that $\lim_{s\to\infty} \xi_T(s)_1 T(s) = \xi_1 = (1 \quad 1 \quad 1) = \tau_1$. Now, $\tau_2 = (0 \quad 0 \quad 1)$ is linearly independent of ξ_1 and so, $\xi_T(s)_2 = (0 \quad 1)$. Thus,

$$\xi_T(s) = \begin{bmatrix} s & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \lim_{s\to\infty} \xi_T(s) T(s) = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = K_T$$

with K_T a constant matrix of rank 2.

Example 3.33. Let

$$T(s) = \begin{bmatrix} \dfrac{1}{s+1} & \dfrac{1}{s+2} \\ \dfrac{1}{s+3} & \dfrac{1}{s+4} \end{bmatrix}.$$

Then $f_1 = 1, f_2 = 1$ and $\tau_1 = (1 \quad 1), \tau_2 = (1 \quad 1)$. So $\xi_T(s)_1 = (s \quad 0)$ and τ_2 is linearly dependent on ξ_1 with $\tau_2 = 1 \cdot \xi_1$. Thus, $\tilde{\xi}^1_T(s)_2 = s[(0 \quad s) - (s \quad 0)] = (-s^2 \quad s^2)$ and $\lim_{s\to\infty} \tilde{\xi}^1_T(s)_2 T(s) = \tilde{\xi}^1_2 = (-2 \quad -2) = -2 \cdot \xi_1$. Since $\tilde{\xi}^1_2$ depends linearly on ξ_1, we continue by setting $\tilde{\xi}^2_T(s)_2 = s[(-s^2 \quad s^2) + 2(s \quad 0)] = (-s^3 \quad +2s^2 \quad s^3)$. Then $\tilde{\xi}^2_2 = (6 \quad 8)$ is not linearly dependent on ξ_1 and so,

$$\xi_T(s) = \begin{bmatrix} s & 0 \\ -s^3 + 2s^2 & s^3 \end{bmatrix},$$

$$\lim_{s\to\infty} \xi_T(s) T(s) = \begin{bmatrix} 1 & 1 \\ 6 & 8 \end{bmatrix} = K_T$$

with K_T a constant matrix of rank 2.

4. Invariants and canonical forms. We begin with some lemmas.

LEMMA 4.1. *Let $T(s)$ be a $p \times m$ element of S_+. Then there is a (not necessarily proper) $m \times p$ transfer matrix $\theta_T(s)$ such that* (i) $T(s)\theta_T(s) = I_p$ *and* (ii) $\theta_T(s)\xi_T^{-1}(s)$ *is proper.*

Proof. If $p = m$, then let $\theta_T(s) = T^{-1}(s)$. Since $\lim_{s\to\infty} \xi_T(s) T(s) = K_T$ is nonsingular, $\lim_{s\to\infty} T^{-1}(s)\xi_T^{-1}(s) = K_T^{-1}$ is nonsingular. It therefore follows that $T^{-1}(s)\xi_T^{-1}(s)$ is proper.

If $p < m$, we append $m - p$ row vectors from the standard basis to $T(s)$ to obtain a proper nonsingular $m \times m$ transfer matrix

$$T_e(s) = \begin{bmatrix} T(s) \\ E \end{bmatrix}. \tag{4.2}$$

Then

$$\xi_{T_e}(s) = \begin{bmatrix} \xi_T(s) & 0_{p,m-p} \\ 0_{m-p,p} & I_{m-p} \end{bmatrix}. \tag{4.3}$$

Let $\theta_T(s)$ be the $m \times p$ transfer matrix consisting of the first p columns of $T_e^{-1}(s)$.

Then $T(s)\theta_T(s) = I_p$. Since $T_e^{-1}(s)\xi_{T_e}^{-1}(s)$ is proper and

$$T_e^{-1}(s) = [\theta_T(s) \quad *_{-m,m-p}],$$
$$\xi_{T_e}^{-1}(s) = \begin{bmatrix} \xi_T^{-1}(s) & 0_{p,m-p} \\ 0_{m-p,p} & I_{m-p} \end{bmatrix}, \tag{4.4}$$

it follows that $\theta_T(s)\xi_T^{-1}(s)$ is proper.

THEOREM 4.5. *Let $T_1(s)$ be a $p \times m$ element of S_+ and let $T_2(s)$ be a $p \times q$ element of S_+. Then there is an element $T(s)$ of S such that*

$$T_1(s)T(s) = T_2(s) \tag{4.6}$$

if and only if $\xi_{T_1}(s)\xi_{T_2}^{-1}(s)$ is proper.

Proof. If $\xi_{T_1}(s)\xi_{T_2}^{-1}(s)$ is proper, then $\xi_{T_1}(s)T_2(s) = [\xi_{T_1}(s)\xi_{T_2}^{-1}(s)][\xi_{T_2}(s)T_2(s)]$ is proper. Hence, $T(s) = \theta_{T_1}(s)T_2(s) = [\theta_{T_1}(s)\xi_{T_1}^{-1}(s)][\xi_{T_1}(s)T_2(s)]$ is proper. But $T_1(s)T(s) = [T_1(s)\theta_{T_1}(s)]T_2(s) = I_pT_2(s) = T_2(s)$.

On the other hand, if there is an element $T(s)$ of S such that $T_1(s)T(s) = T_2(s)$, then $\xi_{T_1}(s)T_2(s) = [\xi_{T_1}(s)T_1(s)]T(s)$ is proper. But $\xi_{T_1}(s)\xi_{T_2}^{-1}(s) = \xi_{T_1}(s)I_p\xi_{T_2}^{-1}(s) = \xi_{T_1}(s)[T_2(s)\theta_{T_2}(s)]\xi_{T_2}^{-1}(s) = [\xi_{T_1}(s)T_2(s)][\theta_{T_2}(s)\xi_{T_2}^{-1}(s)]$ is then a proper transfer matrix.

It is of interest to note that Theorem 4.5 represents a direct resolution to the question of existence of solutions to the well-known model matching problem, which has recently been expanded somewhat and termed the "minimal design problem".

COROLLARY 4.7. *$\xi_T(s)$ is an abstract invariant for E_d on S_+.*

Proof. Suppose that $T_1(s)E_dT_2(s)$ with $T_1(s) \in S_+$ (hence, by Observation 2.6, $T_2(s) \in S_+$). Then Theorem 4.5 implies that both $\xi_{T_1}(s)\xi_{T_2}^{-1}(s)$ and $\xi_{T_2}(s)\xi_{T_1}^{-1}(s)$ are proper $p \times p$ transfer matrices. In view of the uniqueness part of the proof of Lemma 3.1, it follows that $\xi_{T_1}(s) = \xi_{T_2}(s)$.

LEMMA 4.8. *The invariant $\xi_T(s)$ is complete on S_+.*

Proof. Let $T_1(s)$ and $T_2(s)$ be elements of S_+ such that $\xi_{T_1}(s) = \xi_{T_2}(s)$. If G is a constant $m \times p$ matrix such that $\xi_{T_1}G$ is nonsingular, then $\xi_{T_1}(s)T_1(s)G$ is a $p \times p$ element of S with $\lim_{s\to\infty} \xi_{T_1}(s)T_1(s)G = \xi_{T_1}G$ nonsingular. Thus, $[\xi_{T_1}(s)T_1(s)G]^{-1}$ is in S and so is

$$T_{1c}(s) = G[\xi_{T_1}(s)T_1(s)G]^{-1}\xi_{T_2}(s)T_2(s). \tag{4.9}$$

Since $\xi_{T_1}(s) = \xi_{T_2}(s)$, $T_{1c}(s) = G[T_1(s)G]^{-1}T_2(s)$ and

$$T_1(s)T_{1c}(s) = [T_1(s)G][T_1(s)G]^{-1}T_2(s) = T_2(s). \tag{4.10}$$

Similarly, there is a $T_{2c}(s)$ is S with $T_2(s)T_{2c}(s) = T_1(s)$ and so, $T_1(s)E_dT_2(s)$.

We note that Theorem 4.5 has a number of interesting consequences, such as Corollary 4.7, as well as the following corollaries.

COROLLARY 4.11. *Let $T(s)$ be an element of S_+. Then $T(s)$ has a proper right inverse $T_{ri}(s)$ if and only if $\xi_T(s) = I$.*

Proof. $T(s)T_{ri}(s) = I$ if and only if $\xi_T(s)\xi_I^{-1}(s)$ is proper. But $\xi_I(s) = I_p$ and $\xi_T(s)$ is proper if and only if $\xi_T(s) = I$.

COROLLARY 4.12. *Let $T(s)$ be an element of S_-. Then $T(s)$ has a proper left inverse if and only if $\xi_{T'_m}(s) = I$ (where $T'(s)$ is the transpose of $T(s)$).*

We return now to our study of invariants and canonical forms.

DEFINITION 4.13. *If $T(s) \in S$, let ρ_T denote the rank of $T(s)$, and let $S_-, q = \{T(s) \in S_- | \rho_T = q\}$.*

LEMMA 4.14. *$\xi_T(s)$ is an abstract invariant for E_d on S_-.*

Proof. Suppose that $T_1(s)E_dT_2(s)$ with $T_1(s)$, $T_2(s)$ in S_-. Then, clearly, $T_{1m}(s)E_dT_{2m}(s)$ and so, by Corollary 4.7, $\xi_{T_{1m}}(s) = \xi_{T_{2m}}(s)$.

By virtue of Observation 2.7, since $[-\gamma_{21}(s), \gamma_{22}(s)]T_1(s)T_{1c}(s) = [-\gamma_{21}(s), \gamma_{22}(s)]T_2(s) = 0$, we have

$$-\gamma_{21}(s)R_{1m}(s) + \gamma_{22}(s)R_{1p-m}(s) = 0. \tag{4.15}$$

But then $\gamma_{22}^{-1}(s)\gamma_{21}(s) = \gamma_{12}^{-1}(s)\gamma_{11}(s)$ are both relatively left prime factorizations of the *same* transfer matrix with both $\gamma_{22}(s)$ and $\gamma_{12}(s)$ lower left triangular matrices in (unique) Hermite normal form with monic diagonal entries. Thus, $\gamma_{22}(s) = \gamma_{12}(s)$ and $\xi_{T_1}(s) = \xi_{T_2}(s)$.

LEMMA 4.16. *$\xi_T(s)$ is complete on S_-, q.*

Proof. Let $T_1(s)$ and $T_2(s)$ be elements of S_-, q such that $\xi_{T_1}(s) = \xi_{T_2}(s)$. Then $\xi_{T_{1q}}(s) = \xi_{T_{2q}}(s)$ and

$$\begin{aligned} T_{1c}(s) &= [\xi_{T_{1q}}(s)T_{1q}(s)]^{-1}[\xi_{T_{2q}}(s)T_{2q}(s)] \\ &= T_{1q}^{-1}(s)T_{2q}(s) \end{aligned} \tag{4.17}$$

is an element of S. But

$$\begin{aligned} \xi_{T_1}(s)T_1(s)T_{1c}(s) &= \begin{bmatrix} \xi_{T_{1q}}(s)T_{1q}(s) \\ 0 \end{bmatrix} T_{1q}^{-1}(s)T_{2q}(s) \\ &= \begin{bmatrix} \xi_{T_{2q}}(s) \\ 0 \end{bmatrix} T_{2q}(s) \\ &= \xi_{T_2}(s)T_2(s) = \xi_{T_1}(s)T_2(s) \end{aligned} \tag{4.18}$$

and so, $T_1(s)T_{1c}(s) = T_2(s)$. Similarly, there is a $T_{2c}(s)$ in S with $T_2(s)T_{2c}(s) = T_1(s)$ and so, $T_1(s)E_dT_2(s)$. We now state the main result of this paper.

THEOREM 4.19. *Let ψ be the function on S given by*

$$\psi(T(s)) = (\rho_T, \xi_T(S)). \tag{4.20}$$

Then ψ is a complete abstract invariant for equivalence under dynamic compensation.

Proof. By virtue of Observation 2.5, Corollary 4.7 and Lemma 4.14, ψ is an invariant.

As for the completeness of ψ, it will be sufficient in view of Lemmas 4.8 and 4.16 to show that if $\psi(T_1(s)) = \psi(T_2(s))$, then either $T_1(s)$ and $T_2(s)$ are in S_+ or $T_1(s)$ and $T_2(s)$ are in S_-. Suppose that $\psi(T_1(s)) = \psi(T_2(s))$ and that (say) $T_1(s)$ is a $p_1 \times m_1$ element of S_+ and $T_2(s)$ is a $p_2 \times m_2$ element of S_-. Then $\rho_{T_1} = p_1 \leqq m_1$ and $\rho_{T_2} = m_2 < p_2$. But $p_1 = m_2 < p_2$ and so the $p_1 \times p_1$ matrix $\xi_{T_1}(s)$ could not equal the $p_2 \times p_2$ matrix $\xi_{T_2}(s)$. Thus, the theorem is proved.

It is important to note that this theorem establishes the fact that any two dynamical systems (with transfer matrices in S) are equivalent under dynamic compensation if and only if their transfer matrices have equal rank and their interactors are equal.

DEFINITION 4.21. A subset C of S is called *a set of canonical forms for S under E_d* if, for each $T(s)$ in S, there is a unique $C_T(s)$ in C with $T(s)E_dC_T(s)$.

Let $C_+ = \{\xi_T^{-1}(s) | T(s) \in S_+\}$ and

$$C_- = \left\{ \begin{bmatrix} \xi_{T_m}^{-1}(s) \\ T_{p-m}(s)T_m^{-1}(s)\xi_{T_m}^{-1}(s) \end{bmatrix} | T(s) \in S_- \right\}.$$

We then have the following theorem.

THEOREM 4.22. *If $C = C_+ \cup C_-$, then C is a set of canonical forms for S under E_d.*

Proof. If $T(s) \in S_+$, we set

$$T_c(s) = G[\xi_T(s)T(s)G]^{-1},$$

where G is any $m \times p$ constant matrix such that $\xi_T G$ is nonsingular. Then $T(s)T_c(s) = \xi_T^{-1}(s)$ and $T(s) = \xi_T^{-1}(s)T_c(s)^{-1}$ so that $T(s)E_d\xi_T(s)^{-1}$.

If $T(s) \in S_-$, we set

$$T_c(s) = T_m^{-1}(s)\xi_{T_m}^{-1}(s)$$

so that $T(s)T_c(s) = C_T(s)$ and $T(s) = C_T(s)T_c(s)^{-1}$ (i.e. $T(s)E_dC_T(s)$).

Example 4.23. Let $T_1(s)$ be the element of S given by

$$T_1(s) = \begin{bmatrix} \dfrac{1}{s+2} & 0 \\ \dfrac{2}{s+3} & \dfrac{s+1}{s+3} \\ \dfrac{1}{(s+2)(s+3)} & \dfrac{1}{s+3} \\ \dfrac{2s+7}{(s+2)(s+3)} & \dfrac{1}{s+3} \end{bmatrix},$$

and let $T_2(s)$ be the element of S given by

$$T_2(s) = \begin{bmatrix} \dfrac{1}{s+1} & 0 \\ \dfrac{s^2+6s+7}{(s+1)(s+3)} & 1 \\ \dfrac{s+4}{(s+1)(s+3)} & \dfrac{1}{s+1} \\ \dfrac{3s+10}{(s+1)(s+3)} & \dfrac{1}{s+1} \end{bmatrix}.$$

Are $T_1(s)$ and $T_2(s)$ equivalent under dynamic compensation? Since both $T_1(s)$ and $T_2(s)$ are of rank 2, we need only examine $\xi_{T_1}(s)$ and $\xi_{T_2}(s)$. Since

$$\lim_{s\to\infty}\begin{bmatrix} s & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \dfrac{1}{s+2} & 0 \\ \dfrac{2}{s+3} & \dfrac{s+1}{s+3} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and

$$\lim_{s\to\infty}\begin{bmatrix} s & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \dfrac{1}{s+1} & 0 \\ \dfrac{s^2+6s+7}{(s+1)(s-3)} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

we have $\xi_{T_1,2}(s) = \xi_{T_2,2}(s)$. To determine the remaining rows of $\xi_{T_1}(s)$ and $\xi_{T_2}(s)$, we note that

$$T_1(s) = \begin{bmatrix} R_{11}(s) \\ R_{12}(s) \end{bmatrix} P_1^{-1}(s),$$

$$T_2(s) = \begin{bmatrix} R_{21}(s) \\ R_{22}(s) \end{bmatrix} P_2^{-1}(s),$$

where

$$R_{11}(s) = \begin{bmatrix} 1 & 0 \\ 1 & s+1 \end{bmatrix}, \quad R_{12}(s) = \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix}, \quad P_1(s) = \begin{bmatrix} s+2 & 0 \\ -1 & s+3 \end{bmatrix}$$

and

$$R_{21}(s) = \begin{bmatrix} s+3 & 0 \\ s^2+6s+7 & s+1 \end{bmatrix}, \quad R_{22}(s) = \begin{bmatrix} s+4 & 1 \\ 3s+10 & 1 \end{bmatrix},$$

$$P_2(s) = \begin{bmatrix} (s+1)(s+3) & 0 \\ 0 & s+1 \end{bmatrix},$$

respectively. Thus,

$$R_{12}(s)R_{11}^{-1}(s) = \begin{bmatrix} -\dfrac{1}{s+1} & \dfrac{1}{s+1} \\ \dfrac{2s+1}{s+1} & \dfrac{1}{s+1} \end{bmatrix} = \gamma_{12}^{-1}(s)\gamma_{11}(s),$$

$$R_{22}(s)R_{21}^{-1}(s) = \begin{bmatrix} -\dfrac{1}{s+1} & \dfrac{1}{s+1} \\ \dfrac{2s+1}{s+1} & \dfrac{1}{s+1} \end{bmatrix} = \gamma_{22}^{-1}(s)\gamma_{21}(s),$$

where

$$\gamma_{12}(s)=\gamma_{22}(s)=\begin{bmatrix} s+1 & 0 \\ -1 & 1 \end{bmatrix},$$
$$\gamma_{11}(s)=\gamma_{21}(s)=\begin{bmatrix} -1 & 1 \\ 2 & 0 \end{bmatrix}$$

and

$$\xi_{T_1}(s)=\xi_{T_2}(s)=\begin{bmatrix} s & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & -1 & s+1 & 0 \\ -2 & 0 & -1 & 1 \end{bmatrix}.$$

In other words, $T_1(s)E_dT_2(s)$.

The unique[2] dynamic compensator, $T_{1c}(s)$, which equates the two systems is now given by (4.17), i.e.,

$$T_{1c}(s)=T_{1q}^{-1}(s)T_{2q}(s)$$
$$=\begin{bmatrix} \dfrac{s+2}{s+1} & 0 \\ \dfrac{s+3}{s+1} & \dfrac{s+3}{s+1} \end{bmatrix},$$

and $C_{T_1}(s)=C_{T_2}(s)$, the canonical form for both $T_1(s)$ and $T_2(s)$ is given by

$$C_{T_1}(s)=C_{T_2}(s)=\begin{bmatrix} \dfrac{1}{s} & 0 \\ 0 & 1 \\ \dfrac{-1}{s^2+s} & \dfrac{1}{s+1} \\ \dfrac{2s+1}{s^2+3s} & \dfrac{1}{s+1} \end{bmatrix}.$$

5. Concluding remarks. We have now exhibited a complete abstract invariant, $\psi(T(s))=(\rho_T, \xi_T(s))$ for transfer matrix equivalence under dynamic compensation; i.e. we have shown that for systems characterized by full rank, proper transfer matrices, the rank and the interactor determine equivalence under dynamic compensation.

The relevance of this observation with respect to the question of exact model matching and proper right inverses was also shown.

In establishing completeness, explicit expressions are obtained for the requisite dynamic compensators. We further determined a set of canonical forms for the

[2] One can readily establish that the dynamic compensators which equate two equivalent systems in S_- are unique.

class of systems considered and developed the explicit compensators which produce the canonical forms.

Subsequent investigations will build on the results presented here, and will employ the interactor to resolve numerous related questions; e.g. the development of complete abstract invariants for system equivalence under state feedback compensation, the derivation of new and direct procedures for (dynamically and triangularly) decoupling systems via both dynamic and state feedback compensation, and more efficient resolutions to model matching via both stable and minimal order compensation (the minimal design problem).

REFERENCES

[1] P. FALB AND W. A. WOLOVICH, *Linear Systems and Invariants*, Springer-Verlag, Berlin, to appear.

[2] W. A. WOLOVICH, *Linear Multivariable Systems*, Springer-Verlag, Berlin, 1974.

[3] C. C. MACDUFFEE, *An Introduction to Abstract Algebra*, John Wiley, New York, 1940.

[4] LEONARD M. SILVERMAN, *Inversion of multivariable linear systems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 270–276.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 6, November 1976

CRITERIA FOR FUNCTION SPACE CONTROLLABILITY OF LINEAR NEUTRAL SYSTEMS*

MARC Q. JACOBS† AND C. E. LANGENHOP‡

Abstract. Necessary and sufficient conditions for the exact state controllability of the linear autonomous differential difference equation of neutral type, $\dot{x}(t)=A_{-1}\dot{x}(t-h)+A_0x(t)+A_1x(t-h)+Bu(t)$, are given for the Sobolev state space $W_2^{(1)}([-h,0],R^n)$. In particular when B is an $n\times 1$ matrix, it is shown that the controllability of the above n-dimensional system on the interval $[0,\tau]$, $\tau>nh$, is equivalent to rank $[B,A_{-1}B,\cdots,A_{-1}^{n-1}B]=n$ and that a certain two point boundary value problem for a related homogeneous ordinary differential equation have only the trivial solution. Practical criteria based thereon entail only elementary computations involving the coefficient matrices $[A_{-1},A_0,A_1,B]$ but these computations can be tedious when $n>3$. The condition that the two point boundary value problem have only the trivial solution is often equivalent to a much simpler condition: $K(\lambda)\mathscr{S}_\lambda^n\neq 0$ for all complex λ, where $\mathscr{S}_\lambda^n=[1,e^{-\lambda h},\cdots,e^{-(n-1)\lambda h}]^T$, and $K(\lambda)$ is an $n\times n$ matric polynomial of degree $n-1$ which is constructed from the matrix $[A_{-1},A_0,A_1,B]$. This equivalence for the general case is still an open question. It is shown that the collection of controllable neutral systems form an open, dense subset of the collection of all neutral systems of the type considered. This is in marked contrast with the situation for retarded systems. It is also proved (for general B) that when the matrix, $[B,A_{-1}B,\cdots,A_{-1}^{n-1}B]$, has rank n, the solution operator, $u\to x_\tau(\cdot,0,u)$, for quite general neutral systems has closed range and finite deficiency. This often turns out to be an adequate substitute for a controllability assumption.

1. Introduction. In this paper we examine two fundamental questions regarding the attainable set (see (2.3) below) $\mathscr{A}(\tau)\subset W_2^{(1)}([-h,0],R^n)$ of a controlled linear system of autonomous neutral functional differential equations. The first is that of characterizing the controllable systems (i.e., those systems for which $\mathscr{A}(\tau)$ coincides with the Sobolev state space $W_2^{(1)}([-h,0],R^n)$). The second question is the related one of establishing conditions which assure that $\mathscr{A}(\tau)$ is closed in $W_2^{(1)}([-h,0],R^n)$. Answers to both these questions are important in the solution of so-called optimal settling problems for hereditary systems [2], [7], [19], [21]. Often the property that $\mathscr{A}(\tau)$ is closed is exactly what is needed to establish that the given variational problem is normal [2], [4], [21], [22].

The question of controllability is examined only for neutral differential-difference equations of the form

$$\dot{x}(t)=A_{-1}\dot{x}(t-h)+A_0x(t)+A_1x(t-h)+Bu(t), \tag{1.1}$$

where h is a positive constant, the A_i, $i=\pm 1,0$, are $n\times n$ constant real matrices, and B is an $n\times m$ constant real matrix. For this restricted class of neutral systems we can obtain very explicit and computationally effective criteria for checking the controllability. In § 2 we establish a preliminary necessary condition for (1.1) to be controllable on $[0,\tau]$, $\tau>h$ (see § 2 for the terminology). This condition takes the

* Received by the editors June 16, 1975, and in revised form January 14, 1976. This work was supported by the National Science Foundation under Grants GP-33882, NSF MPS72 04695 A03 and NSF GF-37298.

† Department of Mathematics, University of Missouri, Columbia, Missouri 65201.

‡ Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901.

form

$$\text{rank}\,[B, A_{-1}B, \cdots, A_{-1}^{p-1}B] = n \tag{1.2}$$

(p is defined in (2.9) below), and rank $G(\tau - h) = n$ where $G(\tau - h)$ is the controllability Gramian defined in (2.7). These two conditions are independent for general $\tau > h$, but when $\tau > nh$, (1.2) implies that rank $G(\tau - h) = n$.

It is known from [4] that the controllability of (1.1) on $[0, \tau]$ depends on τ if $\tau \leqq nh$ while it is independent of τ, if $\tau > nh$. In § 4 we give necessary and sufficient conditions for (1.1) to be controllable on $[0, \tau]$, $\tau > nh$. In this section we restrict ourselves to scalar controllers u (i.e., B is $n \times 1$). The machinery needed for the general analysis is available in [4], but the simplicity of the results tends to get obscured when the general $n \times m$ matrix B is treated. Moreover, in view of (1.2) and Theorem 6, p. 86 of [24] there is only a small loss of generality. A number of examples illustrating the use of the controllability conditions are given in § 6. The result for 2-dimensional systems which is given in Example 6.1 was reported earlier in [20]. Some of the tools used in § 4 were developed in [4], and are extensions of Minjuk's work [27] on the operational calculus (see also [32]). We also show that controllability of (1.1) on $[0, \tau]$, $\tau > nh$, is a generic property of such systems (B is $n \times 1$); i.e., the set of controllable systems (1.1) on $[0, \tau]$, $\tau > nh$, is an open and dense subset of all systems (1.1).

In § 5 it is noted that the controllable systems (1.1) (B is $n \times 1$) are linearly equivalent to a certain canonical system where A_{-1} is a companion matrix and $B = [0, 0, \cdots, 0, 1]^*$. The equivalence is constructive and these canonical systems are considerably simpler to work with computationally. In this section we discuss another useful necessary condition for the controllability of canonical systems. In addition, we give a Leverrier type of algorithm for recursively generating the operator $K(D)$ used in deciding whether a system is controllable.

As a by-product of our study of the basic system (1.1) (where B is $n \times m$, $m \geqq 1$) we obtain some results on the question of whether $\mathscr{A}(\tau)$ is closed. These results apply to very general systems.

Let $t \mapsto x(t, 0, u)$ denote the solution of (1.1) on $[0, \tau]$ ($u \in L_2([0, \tau], R^m)$) satisfying $x(t) = 0$, $-h \leqq t \leqq 0$. In § 3 it is proved that (1.2) is a necessary and sufficient condition that the solution operator, $u \to x_\tau(\cdot, 0, u)$, for the difference equation (1.1) with $A_1 = A_0 = 0$ to have closed range and finite deficiency in $W_2^{(1)}([-h, 0], R^n)$. This property is shown to persist for quite general neutral systems

$$\frac{d}{dt}\mathscr{D}(x_t) = L(x_t) + Bu(t), \tag{1.3}$$

where $\mathscr{D}$ is a Hale-type difference operator [10], [17] which is a sufficiently small or sufficiently smooth perturbation of the operator $\mathscr{D}_0 \phi = \phi(0) - A_{-1}\phi(-h)$ and L is any continuous linear operator with range R^n and domain the space of continuous functions on $[-h, 0]$ into R^n with the norm of uniform convergence.

It is noted that for retarded systems (1.3) with $\mathscr{D}(\phi) = \phi(0)$ a necessary and sufficient condition for controllability is rank $B = n$. Thus the controllability theory for neutral systems is markedly different (the uncontrollable neutral systems are meagre) from the retarded systems. This might be anticipated from

the analogy between hyperbolic partial differential equations with boundary control and controlled neutral differential equations (see [7] and the references therein). Russell [34] and others have developed a rather complete controllability theory for hyperbolic systems with boundary control in contrast to parabolic systems (analogous to retarded systems) where one usually obtains density results. There is a good survey of earlier work on controllability of retarded systems in [1] and [13]. Gabasov and Kirillova [13] also give some results for the Euclidean controllability of neutral systems. We establish a related result, our Corollary 4.1, below. Minjuk and Stepanjuk [28] and Olbrot [29] have resolved the problem of null controllability for retarded systems. In [20] the authors have used the operational techniques in § 4 to give simple necessary and sufficient conditions for two-dimensional retarded systems to be null controllable.

2. Preliminary necessary conditions for controllability. For a Hilbert space H we denote the inner product by $\langle x, y\rangle$ or by $\langle x, y\rangle_H$ where confusion may otherwise arise. The norm on H is, of course, $\|x\| = \langle x, x\rangle^{1/2}$, $x \in H$. Statements concerning measures, integrals, etc., will refer to Lebesgue measure on R unless explicitly stated to the contrary. For $E \subset R$, E measurable, $L_2(E, R^p)$ denotes all measurable functions $u : E \to R^p$ such that $\int_E \|u(t)\|^2\, dt < \infty$, two such functions u, v being considered the same if they are equal almost everywhere (a.e.) on E. With the inner product $\langle u, v\rangle_{L_2} = \int_E \langle u(t), v(t)\rangle_{R^p}\, dt$ the set $L_2(E, R^p)$ is a Hilbert space.

If $x : [\alpha, \beta] \to R^p$ is absolutely continuous, we define $(Dx)(t) = \dot{x}(t) = (dx(t)/dt)$ a.e. on $[\alpha, \beta]$. Higher powers of the operator D are defined inductively by $D^{k+1} = DD^k$ with domain equal to all $x : [\alpha, \beta] \to R^p$ such that $D^k x$ is absolutely continuous on $[\alpha, \beta]$. By D^0 we understand the identity, $(D^0 x)(t) = x(t)$, $t \in [\alpha, \beta]$. The Sobolev spaces $W_2^{(\nu)}([\alpha, \beta], R^p)$, $\nu = 1, 2, \cdots$, are defined as the collection of all functions $x : [\alpha, \beta] \to R^p$ such that $D^{\nu-1}x$ is absolutely continuous on $[\alpha, \beta]$ and $D^\nu x \in L_2([\alpha, \beta], R^p)$ with inner product given by

$$\langle x, y\rangle_{W_2^{(\nu)}} = \sum_{i=0}^{\nu-1} \langle D^i x(\alpha), D^i y(\alpha)\rangle_{R^p} + \langle D^\nu x, D^\nu y\rangle_{L_2}. \tag{2.1}$$

With the inner product as in (2.1) there is an isometry between $W_2^{(\nu)}([\alpha, \beta], R^p)$ and $(R^p)^\nu \times L_2([\alpha, \beta], R^p)$ given by

$$x \mapsto (x(\alpha), Dx(\alpha), \cdots, D^{\nu-1}x(\alpha), D^\nu x), \qquad x \in W_2^{(\nu)}([\alpha, \beta], R^p).$$

If $A : H_1 \to H_2$ is a linear mapping between two Hilbert spaces, then A^* denotes the adjoint linear mapping and we define $\operatorname{Ker} A = \{x \in H_1 | Ax = 0\}$ and $\operatorname{Im} A = \{Ax | x \in H_1\}$. If H is a Hilbert space and $M \subset H$, then $M^\perp = \{h \in H | \forall m \in M, \langle h, m\rangle = 0\}$. For a matrix A we use A^* for the conjugate transpose.

We remind the reader of the customary notation for the "states" of systems governed by a functional differential equation, viz., if $x : [\tau_0 - h, \tau_1] \to R^n$, $h > 0$, then for $t \in [\tau_0, \tau_1]$ the symbol x_t denotes the function on $[-h, 0]$ defined by $x_t(\theta) = x(t + \theta)$, $\theta \in [-h, 0]$.

Let A_i, $i = \pm 1, 0$, be constant real $n \times n$ matrices and B a constant real $n \times m$ matrix. We consider the differential-difference equation of neutral type given by

$(h>0)$

$$\dot{x}(t)=A_{-1}\dot{x}(t-h)+A_0x(t)+A_1x(t-h)+Bu(t). \tag{2.2}$$

Aside from the control variable u this system is autonomous so no generality is lost by taking the initial time as 0. Let $\mathscr{I}=[0,\tau]$ be a given interval with $\tau>h$. As in [4] we say that (2.2) is *controllable on* $\mathscr{I}$ if for each $\phi,\psi\in W_2^{(1)}([-h,0],R^n)$ there is a $u\in L_2(\mathscr{I},R^m)$ such that $x_\tau(\cdot\,,\phi,u)=\psi$. Here $t\mapsto x(t,\phi,u)$ is the solution of (2.2) for $t\geqq 0$ using control u and initial data $x_0=\phi$. Similarly, (2.2) is *Euclidean controllable* on $\mathscr{I}$ ($\tau>h$ is not required for this) if for each $\phi\in W_2^{(1)}([-h,0],R^n)$ and each $\xi\in R^n$ there is a $u\in L_2(\mathscr{I},R^m)$ such that $x(\tau,\phi,u)=\xi$. The attainable set of states for (2.2) at time τ starting from $\phi\equiv 0$ is

$$\mathscr{A}(\tau)=\{\psi=x_\tau(\cdot\,,0,u)\,|\,u\in L_2([0,\tau],R^m)\}. \tag{2.3}$$

Clearly $\mathscr{A}(\tau)\subseteq W_2^{(1)}([-h,0],R^n)$ with equality if and only if (2.2) is controllable on $\mathscr{I}=[0,\tau]$. It will be convenient later to use the following notations:

$$\begin{aligned}&\text{(i)}\quad X=W_2^{(1)}([-h,0],R^n),\text{ the state space for (2.2),}\\ &\text{(ii)}\quad U=L_2(\mathscr{I},R^m),\text{ the admissible controls, }\mathscr{I}=[0,\tau],\\ &\text{(iii)}\quad \tilde{\phi}(t)=\phi(t-\tau),\ t\in[\tau-h,\tau],\text{ if }\phi\in X,\\ &\text{(iv)}\quad \tilde{X}=W_2^{(1)}([\tau-h,\tau],R^n)=\{\tilde{\phi}\,|\,\phi\in X\}.\end{aligned} \tag{2.4}$$

From the variation of constants formula for (2.2) we have

$$x(t,\phi,u)=x(t,\phi,0)+\int_0^t\Phi(t-s)Bu(s)\,ds,\qquad t\in\mathscr{I}, \tag{2.5}$$

where $\Phi(t)$ is the $n\times n$ transition matrix. That is,

$$\Phi(t)=I_n+\Phi(t-h)A_{-1}+\int_0^t\Phi(r)A_0\,dr+\int_0^{t-h}\Phi(r)A_1\,dr,\qquad t>0, \tag{2.6}$$

(I_n is the $n+n$ identity matrix) and $\Phi(t)=0$ for $t<0$ (cf. [8], [17]). The *Euclidean controllability Gramian* for (2.2) is defined for $t\geqq 0$ by

$$G(t)=\int_0^t\Phi(t-s)BB^*\Phi^*(t-s)\,ds=\int_0^t\Phi(s)BB^*\Phi^*(s)\,ds. \tag{2.7}$$

The proof of the following propositions is easy and will be omitted.

PROPOSITION 2.1. *In order that* (2.2) *be Euclidean controllable on* $[0,\tau]$ *it is necessary and sufficient that* $G(\tau)$ *have rank* n.

For any $n\times n$ matrix A and $n\times m$ matrix B we define the $n\times\nu m$ matrix

$$C_\nu[A,B]=[B,AB,\cdots,A^{\nu-1}B] \tag{2.8}$$

for integers $\nu\geqq 1$. For any $t\geqq 0$ we let

$$p(t)=[t/h] \tag{2.9}$$

where $[\,\cdot\,]$ denotes the greatest integer function. We may now prove a necessary condition for controllability of (2.2).

PROPOSITION 2.2 *Let* $\mathscr{I}=[0,\tau]$ *and* $p=p(\tau)$. *If* (2.2) *is controllable on* $\mathscr{I}$, *then* $\tau>h$, rank $G(\tau-h)=n$ *and* rank $C_p[A_{-1},B]=n$.

Proof. It is obvious that if (2.2) is controllable on $\mathscr{I}$, then $\tau>h$ and that (2.2) is Euclidean controllable on $[0,t]$ for each $t\in[\tau-h,\tau]$. Hence rank $G(t)=n$ for

$t \in [\tau - h, \tau]$ by Proposition 2.1. (Since rank $G(t)$ is a nondecreasing integer valued function on t, it is clear, in any case, that rank $G(\tau - h) = n$ implies rank $G(t) = n$ for $t \in [\tau - h, \tau]$.) Now let $C_p = C_p[A_{-1}, B]$ and suppose rank $C_p < n$. Then there is a $y \in R^n$, $y \neq 0$, such that

$$y^* C_p = 0. \tag{2.10}$$

Since $ph \leqq \tau < (p+1)h$, the interval $[\tau - h, ph]$ has positive length. Let $\psi \in X$. Since we are assuming (2.2) is controllable on $\mathscr{I}$ there is a $u \in U$ such that $x_\tau(\cdot, 0, u) = \psi$. Hence from (2.2) we have

$$\dot{\tilde{\psi}}(t) - A_0 \tilde{\psi}(t) = A_{-1}\dot{x}(t-h) + A_1 x(t-h) + Bu(t) \quad \text{a.e. on } [\tau - h, \tau], \tag{2.11}$$

where $x(t) = x(t, 0, u)$. But also from (2.2),

$$\begin{aligned} \dot{x}(t-kh) = {} & A_{-1}\dot{x}(t-(k+1)h) + A_0 x(t-kh) \\ & + A_1 x(t-(k+1)h) + Bu(t-kh) \end{aligned} \tag{2.12}$$

for $k = 1, \cdots, p-1$, $t \in [\tau - h, \tau]$ a.e. Using the fact that $x(t) = 0$ for $t \leqq 0$, we may combine (2.11) and (2.12) to get

$$\dot{\tilde{\psi}}(t) - A_0 \tilde{\psi}(t) = \sum_{i=1}^{p-1} \Lambda_i x(t-ih) + C_p \begin{bmatrix} u(t) \\ u(t-h) \\ \vdots \\ u(t-(p-1)h) \end{bmatrix} \tag{2.13}$$

a.e. on $[\tau - h, ph]$, where

$$\Lambda_1 = A_1 + A_{-1}A_0, \qquad \Lambda_{i+1} = A_{-1}\Lambda_i, \quad i = 1, 2, \cdots. \tag{2.14}$$

Now multiply both sides of (2.13) by y^* and use (2.10) to get

$$y^*[\dot{\tilde{\psi}}(t) - A_0 \tilde{\psi}(t)] = \sum_{i=1}^{p-1} y^* \Lambda_i x(t-ih) \quad \text{a.e. on } [\tau - h, ph]. \tag{2.15}$$

For any $f \in L_2([\tau - h, \tau], R^n)$ the equation $\dot{\tilde{\psi}}(t) - A_0 \tilde{\psi}(t) = f(t)$ has a solution $\tilde{\psi} \in \tilde{X}$. Hence by appropriate choice of ψ, the left side of (2.15) can be taken to be any function in $L_2([\tau - h, ph], R)$ since $y \neq 0$. The assumption that rank $C_p < n$ has thus led to the conclusion that an arbitrary element of $L_2([\tau - h, ph], R)$ is equal a.e. on $[\tau - h, ph]$ to a continuous function (the right side of (2.15)). This is a contradiction so rank $C_p = n$.

When $\tau > nh$ the condition rank $C_p = n$ in Proposition 2.2 is equivalent to rank $C_n = n$. This follows in the usual way from the Cayley-Hamilton theorem. It would appear that whether (2.2) is controllable on $\mathscr{I} = [0, \tau]$ depends on the length of the interval $\mathscr{I}$ and not just on the coefficient matrices A_i, $i = \pm 1, 0$ and B. This is, indeed, the case but this does not persist when $\tau > nh$. (See Corollary 5.1 of

[4] where it is shown that $\mathscr{A}(\tau)$ is constant for $\tau > nh$.) Accordingly, we confine our search for necessary and sufficient conditions for controllability of (2.2) on $\mathscr{J}$ to the situation $\tau > nh$. It will, of course, also be necessary then to assume that rank $C_n = n$. For future reference we state the following corollary to Proposition 2.2.

COROLLARY 2.1. *If* (2.2) *is controllable on* $\mathscr{J} = [0, \tau]$, *and if* $\tau > nh$, *then* rank $G(\tau - h) = n$ *and* rank $C_n[A_{-1}, B] = n$.

Remark 2.1. It is natural to conjecture that the two conditions in the conclusion to Corollary 2.1 are also sufficient for controllability on $\mathscr{J}$ when $\tau > nh$. However, the system ($n = 2$)

$$\begin{aligned}\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} &= \begin{bmatrix}0 & 1\\ \gamma_{-1} & \delta_{-1}\end{bmatrix}\begin{bmatrix}\dot{x}_1(t-h)\\ \dot{x}_2(t-h)\end{bmatrix} + \begin{bmatrix}0 & 1\\ \gamma_0 & \delta_0\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix}\\ &\quad + \begin{bmatrix}0 & -1\\ \gamma_1 & \delta_1\end{bmatrix}\begin{bmatrix}x_1(t-h)\\ x_2(t-h)\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix}u(t),\end{aligned} \tag{2.16}$$

where γ_i, δ_i, $i = \pm 1, 0$ are real numbers, satisfies both these conditions. In fact rank $C_2[A_0, B] = 2$ and this implies that rank $G(\tau - h) = 2$ if $\tau > 2h$ (see Remark 3.3 in [2]). The system (2.16) is not controllable on $[0, \tau]$, $\tau > 2h$ (see [20]). This is easily shown using the results developed later in this paper (cf. the examples in § 6).

We recall that (2.2) is Euclidean controllable on $[0, \tau]$ for each $\tau > 0$ if and only if rank $C_n[A_0, B] = n$ (cf. Lemma 3.3 of [4] and Remark 3.3 of [2]). Consequently rank $G(\tau - h) = n$ and rank $C_p[A_{-1}, B] = n$ ($p = p(\tau)$) are, in general, independent conditions. However, we shall see below (Corollary 4.1) that if $\tau > nh$, then rank $C_n[A_{-1}, B] = n$ implies rank $G(\tau - h) = n$. In § 3 we develop some additional important consequences of the condition rank $C_p[A_{-1}, B] = n$.

Remark 2.2. The proof of Proposition 2.2 contains a result which is worth emphasizing. If a function $\psi \in X$ is also in $\mathscr{A}(\tau)$ and $\tau > h$, then there is a $u \in U$ and an $x \in W_2^{(1)}([0, \tau], R^n)$ such that $x_\tau = \psi$ so

$$\dot{\tilde{\psi}}(t) = \begin{cases} A_0\tilde{\psi}(t) + \sum_{i=1}^{p} \Lambda_i x(t-ih) + C_{p+1}[A_{-1}B]\begin{bmatrix}u(t)\\ \vdots\\ u(t-ph)\end{bmatrix} \\ \qquad\qquad \text{a.e. on } [ph, \tau], \\ A_0\tilde{\psi}(t) + \sum_{i=1}^{p-1} \Lambda_i x(t-ih) + C_p[A_{-1}, B]\begin{bmatrix}u(t)\\ \vdots\\ u(t-(p-1)h)\end{bmatrix} \\ \qquad\qquad \text{a.e. on } [\tau - h, ph], \end{cases} \tag{2.17}$$

where the Λ_i are given by (2.14). From this we see that there is an intimate connection between neutral systems (2.2) and retarded systems with delayed control action (see also [3], [6], [26]).

3. A semi-Fredholm property of the solution operator $x_\tau(\cdot, 0, u)$. Let H_1 and H_2 be Hilbert spaces and $\mathscr{T}: H_1 \to H_2$ a bounded linear operator. According

to Kato [23, p. 230] the operator $\mathscr{T}$ is *semi-Fredholm* if Im $\mathscr{T}$ is closed and either the nullity of $\mathscr{T}$ ($=$ null $\mathscr{T}=$ dim ker $\mathscr{T}$) or the deficiency of $\mathscr{T}$ ($=$ def $\mathscr{T}=$ dim $H_2/\text{Im}\ \mathscr{T}$) is finite. It follows from results in [23] (the proof of Theorem 5.26, p. 238, and Theorem 5.13, p. 234) that if $\mathscr{T}$ has closed range and finite deficiency and $\mathscr{C}: H_1 \to H_2$ is a compact linear operator, then $\mathscr{T}+\mathscr{C}$ has closed range and finite deficiency. This is also given explicitly in [35, p. 129] and [33, p. 316]. Moreover, if $\mathscr{T}$ has finite deficiency, then $\mathscr{T}$ must have closed range [23, p. 230].

Since we need be concerned only with solutions of (2.2) specified by zero initial data (i.e. $x(t)=0$, $t\in[-h, 0]$), it is convenient to extend these to all of $(-\infty, 0]$ and to extend u as well so that $x(t)=0$, $u(t)=0$, $t\leqq 0$. Accordingly, we define $W_{2,0}^{(\nu)}(\tau, R^\mu)$, μ a positive integer and ν a nonnegative integer, to be the collection of all $x:(-\infty, \tau]\to R^\mu$ such that $x(t)=0$ for $t\leqq 0$ and the restriction $x|_{[0,\tau]}$ is in $W_2^{(\nu)}([0, \tau], R^\mu)$. Here we adopt the convention $W_2^{(0)}([0, \tau], R^\mu)=L_2([0, \tau], R^\mu)$. For $f\in W_{2,0}^{(\nu)}(\tau, R^\mu)$ define the shift operator S by

$$(Sf)(t)=f(t-h), \qquad t\leqq\tau. \tag{3.1}$$

We define S^0 to be the identity operator on $W_{2,0}^{(\nu)}(\tau, R^\mu)$ and inductively using (3.1), we take $S^{k+1}=SS^k$, $k=0, 1, 2, \cdots$. For any integer $q\geqq 1$ define $\mathscr{S}^q: W_{2,0}^{(\nu)}(\tau, R^\mu)\to W_{2,0}^{(\nu)}(\tau, (R^\mu)^q)$ by

$$\mathscr{S}^q f=\begin{bmatrix} S^0 f \\ S^1 f \\ \vdots \\ S^{q-1} f \end{bmatrix}, \qquad f\in W_{2,0}^{(\nu)}(\tau, R^\mu). \tag{3.2}$$

We will also write $\mathscr{S}^q=[I_\mu S^0, I_\mu S, \cdots, I_\mu S^{q-1}]^*$ where I_μ is the $\mu\times\mu$ identity matrix.

Consider now the special case of (2.2) in which $A_0=A_1=0$, i.e. for $t\geqq 0$

$$\dot{x}(t)=A_{-1}\dot{x}(t-h)+Bu(t). \tag{3.3}$$

If $\mathscr{T}_d$ denotes the solution operator for (3.3), i.e. $\mathscr{T}_d u=x_\tau(\cdot, 0, u)$, then $\mathscr{T}_d: U\to X$ is a bounded linear operator. With $x=x(\cdot, 0, u)$ and u interpreted as the corresponding extensions to functions in $W_{2,0}^{(1)}(\tau, R^n)$ and $W_{2,0}^{(0)}(\tau, R^m)$, respectively, the relation (3.3) holds a.e. on $(-\infty, \tau]$.

PROPOSITION 3.1. *The solution operator $\mathscr{T}_d$ corresponding to* (3.3) *has closed range and finite deficiency if and only if* rank $C_p[A_{-1}, B]=n$ *where* $p=p(\tau)=[\tau/h]$.

Proof. Let $p=p(\tau)=[\tau/h]$ and $C_k=C_k[A_{-1}, B]$. By Remark 2.2 we have that if $\psi\in\mathscr{A}(\tau)$ and u is such that $x_\tau(\cdot, 0, u)=\psi$, then

$$\dot{\psi}(t)=\begin{cases} C_{p+1}\mathscr{S}^{p+1}u(t), & \text{a.e. } t\in[ph, \tau], \\ C_p\mathscr{S}^p u(t), & \text{a.e. } t\in[\tau-h, ph]. \end{cases} \tag{3.4}$$

If rank $C_p<n$, then def $\mathscr{T}_d=+\infty$. This follows at once from (3.4) and the fact that $[\tau-h, ph]$ has positive length. Conversely, suppose rank $C_p=n$. For arbitrary

$\psi \in X$ we define $u \in W_{2,0}^{(0)}(\tau, R^m)$ by

$$\begin{aligned} \mathcal{S}^{p+1}u(t) &= C_{p+1}^{+}\dot{\psi}(t), \quad \text{a.e. } t \in [ph, \tau], \\ \mathcal{S}^{p}u(t) &= C_{p}^{+}\dot{\psi}(t), \quad \text{a.e. } t \in [\tau - h, ph], \\ u(t) &= 0, \qquad t \leqq 0, \end{aligned} \tag{3.5}$$

where C_k^+ denotes the generalized inverse of C_k, $k = p, p+1$ (cf. [25, p. 163]). Both C_p and C_{p+1} have rank n so $C_k^+ = C_k^*(C_k C_k^*)^{-1}$, $k = p, p+1$. The relations (3.5) define $u(t)$ a.e. on $(-\infty, \tau]$. Now let x denote the solution in $W_{2,0}^{(1)}(\tau, R^m)$ on $(-\infty, \tau]$ to (3.3) for the controller u defined by (3.5). From the analysis leading to (2.17) it follows that

$$\dot{x}(t) = \begin{cases} C_{p+1}\mathcal{S}^{p+1}u(t), & \text{a.e. } t \in [ph, \tau], \\ C_p \mathcal{S}^p u(t), & \text{a.e. } t \in [\tau - h, ph]. \end{cases} \tag{3.6}$$

But rank $C_p = n$ so $C_k C_k^+ = I_n$, $k = p, p+1$, and (3.5) and (3.6) imply

$$\dot{x}(t) = \dot{\psi}(t), \quad \text{a.e. } t \in [\tau - h, \tau]. \tag{3.7}$$

Hence $\psi = x_\tau + c$ for some constant function $c \in X$. This proves that

$$X = \mathcal{A}(\tau) + \mathcal{F} = \operatorname{Im} \mathcal{T}_d + \mathcal{F}, \tag{3.8}$$

where $\mathcal{A}(\tau)$ is given by (2.3) ($A_0 = A_1 = 0$) and

$$\mathcal{F} = \{c \in X \mid Dc = 0 \quad \text{a.e.} \quad \text{on } [-h, 0]\}. \tag{3.9}$$

But $\dim \mathcal{F} = n$ so $\mathcal{T}_d$ has finite deficiency which in turn implies that $\operatorname{Im} \mathcal{T}_d = \mathcal{A}(\tau)$ is closed in X.

Only in exceptional circumstances will $\operatorname{rank} C_p[A_{-1}, B] = n$ imply the stronger conclusion that $\mathcal{T}_d(U) = X$. In fact, (3.3) requires

$$x(t) = A_{-1}x(t-h) + B\int_0^t u(s)\, ds, \qquad t \leqq \tau,$$

for $u \in W_{2,0}^{(0)}(\tau, R^m)$ and $x \in W_{2,0}^{(1)}(\tau, R^n)$. In particular, if $\mathcal{T}_d u = \psi$, then we must have

$$\psi(0) - A_{-1}\psi(-h) = B\int_0^\tau u(s)\, ds. \tag{3.10}$$

Clearly (3.10) is possible for arbitrary $\psi \in X$ only if $\operatorname{rank} B = n$. On the other hand, it is obvious that $\operatorname{Im} \mathcal{T}_d = X$ if $\operatorname{rank} B = n$. Hence we have

PROPOSITION 3.2. *The solution operator $\mathcal{T}_d$ corresponding to* (3.3) *satisfies* $\operatorname{Im} \mathcal{T}_d = X$ *if and only if* $\operatorname{rank} B = n$.

This result is analogous to the controllability condition for linear retarded systems given in [19, Lem. 3.1] and [4, Thm. 3.1].

Remark 3.1. It is of interest to note what functions $\psi \in X$ could be missing from $\operatorname{Im} \mathcal{T}_d$. Now it was shown in [4, Corollary 5.1] that for any system (2.2) the attainable sets from zero initial data satisfy

$$\mathcal{A}(\tau_1) \subseteq \mathcal{A}(\tau_2) \quad \text{if } \tau_2 \geqq \tau_1 \geqq 0$$

with equality holding if $\tau_1 > nh$. Thus to get some idea of the functions missing from $\operatorname{Im} \mathcal{T}_d$ for the special case (3.3) we assume $\tau = (n+1)h$ and rank $C_{n+1} = n$, $C_{n+1} = C_{n+1}[A_{-1}, B]$. Define the controller $u \in W_{2,0}^{(0)}((n+1)h, R^m)$ by $u(t) = 0$, $t \leqq 0$, and

$$\mathcal{S}^{n+1}u(t) = C_{n+1}^{+}\dot{\tilde{\psi}}(t), \quad \text{a.e. } t \in [nh, (n+1)h]. \tag{3.11}$$

We may then write

$$\mathcal{S}^{n+1}u(t-h) = J\mathcal{S}^{n+1}u(t), \qquad t \leqq (n+1)h, \tag{3.12}$$

where J is the $(n+1)m \times (n+1)m$ nilpotent matrix given by

$$J = \begin{bmatrix} 0 & I_m & 0 & \cdots & 0 \\ 0 & 0 & I_m & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & I_m \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{3.13}$$

It follows that for $i = 0, 1, 2, \cdots, n$,

$$\mathcal{S}^{n+1}u(t-ih) = J^i C_{n+1}^{+}\dot{\tilde{\psi}}(t), \quad \text{a.e. } t \in [nh, (n+1)h]. \tag{3.14}$$

Now, as in the proof of Proposition 2.2, we get for this u that the solution $x \in W_{2,0}^{(1)}((n+1)h, R^n)$ of (3.3) satisfies

$$\dot{x}(t) = C_k \mathcal{S}^k u(t), \quad \text{a.e. } t \in [(k-1)h, kh], \quad k = 1, \cdots, n+1,$$

where $C_k = C_k[A_{-1}, B]$. Since $u(t) = 0$ for $t \leqq 0$ this may be written

$$\dot{x}(t) = C_{n+1}\mathcal{S}^{n+1}u(t), \quad \text{a.e. } t \in [(k-1)h, kh].$$

Integrating this from 0 to $t \in [nh, (n+1)h]$, we have $(x(0) = 0)$

$$\begin{aligned} x(t) &= \int_0^t \dot{x}(s)\,ds = \sum_{i=1}^{n} \int_{(n-i)h}^{(n+1-i)h} C_{n+1}\mathcal{S}^{n+1}u(s)\,ds + \int_{nh}^{t} C_{n+1}\mathcal{S}^{n+1}u(s)\,ds \\ &= \sum_{i=1}^{n} \int_{nh}^{(n+1)h} C_{n+1}\mathcal{S}^{n+1}u(s-ih)\,ds + \int_{nh}^{t} C_{n+1}C_{n+1}^{+}\dot{\tilde{\psi}}(s)\,ds. \end{aligned}$$

Using (3.14) and (3.11), we get

$$x(t) = \sum_{i=1}^{n} C_{n+1}J^i C_{n+1}^{+}(\tilde{\psi}((n+1)h) - \tilde{\psi}(nh)) + \tilde{\psi}(t) - \tilde{\psi}(nh), \tag{3.15}$$

since rank $C_{n+1} = n$ and thus

$$C_{n+1}C_{n+1}^{+} = I_n. \tag{3.16}$$

Inasmuch as J in (3.13) is nilpotent $(J^{n+1} = 0)$ we have

$$\sum_{i=0}^{n} J^i = (I - J)^{-1} \tag{3.17}$$

in which I is the $(n+1)m \times (n+1)m$ identity. Here $\tau = (n+1)h$ so $\tilde{\psi}(nh) = \psi(-h)$ and $\tilde{\psi}((n+1)h) = \psi(0)$. Hence $\psi = x_\tau \in \mathscr{A}(\tau)$ if

$$\sum_{i=1}^{n} C_{n+1} J^i C_{n+1}^{+}(\psi(0) - \psi(-h)) - \psi(-h) = 0.$$

Using (3.17), we get

$$C_{n+1}[I-J]^{-1} C_{n+1}^{+}(\psi(0) - \psi(-h)) = \psi(0) \tag{3.18}$$

as a sufficient condition for ψ to belong to $\operatorname{Im} \mathscr{T}_d$

We consider now a perturbation of (3.3):

$$\dot{x}(t) = A_{-1}\dot{x}(t-h) + L(x_t) + Bu(t), \tag{3.19}$$

where L is a continuous linear mapping into R^n with domain $C([-h, 0], R^n)$, the space of continuous functions on $[-h, 0]$ into R^n with the norm of uniform convergence. If $\tilde{\mathscr{T}}$ is the solution operator corresponding to (3.19) (initial data $\phi = 0$), then both $\mathscr{T}_d$ and $\tilde{\mathscr{T}}$ are bounded linear mappings from U to X (cf. [17]). Let $\Phi_d(t)$ and $\tilde{\Phi}(t)$ be the transition matrices for the homogeneous systems ($u = 0$) corresponding to (3.3) and (3.19) respectively. Thus Φ_d is calculated from (2.6) with $A_0 = A_1 = 0$ and for $\tilde{\Phi}$ there is a parallel theory explained in [17]. In particular, if $\tilde{x}(t) = x(t, 0, u)$ for (3.19) and $x_d(t) = x(t, 0, u)$ for (3.3), $u \in U$, then for $0 \leqq t \leqq \tau$ we have

$$\tilde{x}(t) = \int_0^t \tilde{\Phi}(t-s)Bu(s)\, ds, \tag{3.20}$$

$$x_d(t) = \int_0^t \Phi_d(t-s)Bu(s)\, ds. \tag{3.21}$$

From (2.6) one readily obtains

$$\Phi_d(t) = \sum_{i=0}^{k-1} A_{-1}^i, \quad (k-1)h \leqq t < kh, \quad k = 1, 2, \cdots, \tag{3.22}$$

whose discontinuities are at worst simple jumps at integer multiples of h. The same applies to $\tilde{\Phi}$. We may now establish

THEOREM 3.1. *If $\tau > h$ and* $\operatorname{rank} C_p[A_{-1}, B] = n$ *where $p = p(\tau) = [\tau/h]$, then the solution operator $\tilde{\mathscr{T}}: U \to X$ corresponding to* (3.19) ($\tilde{\mathscr{T}}u = \tilde{x}_\tau(\cdot\,, 0, u)$) *has closed range and finite deficiency.*

Proof. This follows from Proposition 3.1 and the comments on semi-Fredholm operators at the beginning of this section if we show that $\mathscr{C} = \tilde{\mathscr{T}} - \mathscr{T}_d$ is compact. To this end let $u^\nu \in U = L_2([0, \tau], R^m)$ and suppose $u^\nu \to u$ weakly (in U) as $\nu \to \infty$. The solutions of (3.19) and (3.3) corresponding to u^ν will be denoted by $\tilde{x}^\nu$ and x_d^ν. They are obtained from (3.20) and (3.21), respectively, by replacing u by u^ν. It follows from these formulas that $\tilde{x}^\nu(t) \to \tilde{x}(t)$ and $x_d^\nu(t) \to x_d(t)$ for each $t \in [0, \tau]$. Moreover, since the sequence $\{u^\nu\}$ converges weakly in U, there is a $M_1 > 0$ such that $\|u^\nu\|_{L_2} \leqq M_1$, $\nu = 1, 2, \cdots$ (cf. [11]). It then follows from (3.20) and (3.21) with u replaced by u^ν that $\|\tilde{x}^\nu\|_{L_2} \leqq M_2$ and $\|x_d^\nu\|_{L_2} \leqq M_2$ for some $M_2 > 0$. Thus $\tilde{x}^\nu, x_d^\nu$ and u^ν are uniformly bounded in L_2-norm ($L_2 = L_2([0, \tau], R^\mu)$ where $\mu = n$ or m). Using $\tilde{x}^\nu(t) = x_d^\nu(t) = 0$, $t \leqq 0$, and (3.19) and

(3.3), respectively, one may then show that there is $M_3>0$ such that $\|\dot{\tilde{x}}^\nu\|_{L_2}\leqq M_3$ and $\|\dot{x}_d^\nu\|_{L_2}\leqq M_3$. This is accomplished by working along $[0,\tau]$ by a finite sequence of subintervals $[(k-1)h, kh]$, $k=1,2,\cdots$. It follows that both sequences $\{\tilde{x}^\nu\}$ and $\{x_d^\nu\}$ are uniformly bounded and equicontinuous on $[0,\tau]$ so by the Arzela-Ascoli theorem each has a subsequence which is uniformly convergent on $[0,\tau]$. But these sequences converge pointwise on $[0,\tau]$ so this convergence must then also be uniform on $[0,\tau]$. It follows that $L(\tilde{x}_t^\nu)\to L(\tilde{x}_t)$ for $t\leqq\tau$ since L is continuous on $C([-h,0],R^n)$ endowed with the uniform norm. Applying the same principle used in obtaining (2.17), we get

$$\dot{\tilde{x}}^\nu(t)=\begin{cases}\sum_{i=0}^{p} A_{-1}^i[L(\tilde{x}_{t-ih}^\nu)+Bu^\nu(t-ih)], & \text{a.e. } t\in[ph,\tau],\\ \sum_{i=0}^{p-1} A_{-1}^i[L(\tilde{x}_{t-ih}^\nu)+Bu^\nu(t-ih)], & \text{a.e. } t\in[\tau-h,ph].\end{cases}\tag{3.23}$$

Similarly (or apply (3.6)),

$$\dot{x}_d^\nu(t)=\begin{cases}\sum_{i=0}^{p} A_{-1}^i Bu^\nu(t-ih), & \text{a.e. } t\in[ph,\tau],\\ \sum_{i=0}^{p-1} A_{-1}^i Bu^\nu(t-ih), & \text{a.e. } t\in[\tau-h,ph].\end{cases}\tag{3.24}$$

From (3.23) and (3.24) it follows that the sequence $D\mathscr{C}u^\nu(t)=\dot{\tilde{x}}^\nu(t)-\dot{x}_d^\nu(t)$ converges pointwise on $[\tau-h,\tau]$ to a continuous function since $\tilde{x}_t$ is a continuous function of t on $t\leqq\tau$. Moreover, the boundedness of L and the fact that the sequence $\{\tilde{x}^\nu\}$ is uniformly bounded on $[0,\tau]$ imply for some M_4 that $\|D\mathscr{C}u^\nu(t)\|\leqq M_4$ for all ν and all $t\in[\tau-h,\tau]$. An application of the Lebesgue dominated convergence theorem then shows that $D\mathscr{C}(u^\nu)$ converges in $L_2([\tau-h,\tau],R^n)$. Since $\mathscr{C}(u^\nu)(\tau-h)=\tilde{x}^\nu(\tau-h)-x_d(\tau-h)$ converges in R^n, it follows that $\mathscr{C}$ is compact and the theorem is proved. (The argument used in this proof is the same type as that of Reid in [31].)

COROLLARY 3.1. *Let $\mathscr{A}(\tau)$ be the attainable set of* (2.2) *as defined in* (2.3) *and suppose $\tau>nh$. Then $\mathscr{A}(\tau_1)=\mathscr{A}(\tau_2)$ if $\tau_1,\tau_2\geqq\tau$ and if* rank $C_n[A_{-1},B]=n$, *then $\mathscr{A}(\tau)$ is closed in X with* dim $\mathscr{A}(\tau)^\perp$ *finite.*

Proof. The first part was mentioned earlier (Corollary 5.1 of [4]). The last part follows from Theorem 3.1 by taking $L(\phi)=A_0\phi(0)+A_1\phi(-h)$ in (3.19).

Let $C=C([-h,0],R^n)$ with the uniform norm as before and let $B(C,R^n)$ denote the collection of all bounded linear operators on C to R^n. Of particular interest is the subset B_0 of all $\mathscr{D}\in B(C,R^n)$ such that

$$\mathscr{D}\phi=\phi(0)-\int_{-h}^{0} d\mu(\theta)\phi(\theta)$$

where $\theta\to\mu(\theta)$ is an $n\times n$ matrix-valued function of bounded variation on $[-h,0]$ which is continuous at $\theta=0$ and satisfies $\mu(0)=0$. We now consider a fixed $\mathscr{D}=\mathscr{D}_0+\mathscr{D}_1$ where $\mathscr{D}_1\in B_0$ and

$$\mathscr{D}_0\phi=\phi(0)-A_{-1}\phi(-h),\qquad \phi\in C.$$

Let $\mathcal{T}_{\mathcal{D}}$ denote the solution operator (cf. [17]) $u \mapsto x_\tau(\cdot, 0, u)$ for the system

$$\frac{d}{dt}\mathcal{D}(x_t) = L(x_t) + Bu(t). \tag{3.25}$$

Then $\mathcal{T}_{\mathcal{D}} = \tilde{\mathcal{T}} + A_{\mathcal{D}}$ where $A_{\mathcal{D}} : U \to X$ is a bounded linear operator such that for some $\kappa > 0$,

$$\|A_{\mathcal{D}}u\| \leq \kappa \|\mathcal{D}_1\| \, \|u\|, \qquad u \in U, \tag{3.26}$$

and $\tilde{\mathcal{T}}$ is the solution operator in Theorem 2.1.

COROLLARY 3.2. *If* rank $C_p[A_{-1}, B] = n$, $p = p(\tau)$, *then there is a* $\delta > 0$ *such that if*

$$\|\mathcal{D}_1\| < \delta,$$

then the solution operator $\mathcal{T}_{\mathcal{D}}$ *has closed range and finite deficiency in* X.

Proof. This is an immediate consequence of Theorem 3.1, inequality (3.26), and a result of Kato [23, Thm. 5.22, p. 236].

COROLLARY 3.3. *Let* $\mathcal{D}_1(\phi)(\theta) = \int_{-h}^{0} d\lambda(\theta)\phi(\theta)$, *where* $\lambda \in W_2^{(2)}([-h, 0], M_n)$ (M_n *the space of* $n \times n$ *matrices over the reals*) *and let* $\mathcal{D} = \mathcal{D}_0 + \mathcal{D}_1$. *Then* $\mathcal{T}_{\mathcal{D}}$ *has closed range and finite deficiency.*

Proof. Note that $\mathcal{T}_{\mathcal{D}}(u) = x_\tau(\cdot, 0, u)$, $u \in U$, is the solution operator for the equation ($\lambda'(\theta) = d\lambda(\theta)/d\theta$)

$$\begin{aligned}\dot{x}(t) &= A_{-1}\dot{x}(t-h) + \int_{-h}^{0} \lambda'(\theta)\dot{x}(t+\theta)\, d\theta + L(x_t) + Bu(t) \\ &= A_{-1}\dot{x}(t-h) + \lambda'(0)x(t) - \lambda'(-h)x(t-h) \\ &\quad - \int_{-h}^{0} \lambda''(\theta)x(t+\theta)\, d\theta + L(x_t) + Bu(t).\end{aligned}$$

Thus, under the given hypotheses, (3.25) can be written in a form to which Theorem 2.1 applies.

COROLLARY 3.4. *Let* $\mathcal{D} \in B_0$, *and suppose the solution operator* $u \mapsto x_\tau(\cdot, 0, u)$, $u \in U$, *for the difference equation*

$$\frac{d}{dt}\mathcal{D}x_t = Bu(t), \quad \text{a.e. } t \in [0, \tau],$$

has closed range and finite deficiency. If $L \in B(C, R^n)$, *then the solution operator* $u \mapsto x_\tau(\cdot, 0, u)$, $u \in U$ *for*

$$\frac{d}{dt}\mathcal{D}x_t = L(x_t) + Bu(t), \quad \text{a.e. } t \in [0, \tau],$$

also has closed range and finite deficiency.

The proof of this is essentially the same as the proof of Theorem 3.1 and is omitted.

Remark 3.2. We note that if (2.2) is the n-dimensional first order equation corresponding to the nth order scalar neutral equation, then the condition rank $C_n[A_{-1}, B] = n$ is not satisfied. For this system $u \mapsto x_\tau(\cdot, 0, u)$ has closed range but the deficiency is not finite (cf. Example 3.1 of [2]).

4. Necessary and sufficient conditions for controllability. In §§ 2 and 3 we defined the differentiation operator D, the shift operator S, and the function spaces $W_{2,0}^{(\nu)}(\tau, R^\mu)$ for integers $\nu \geqq 0, \mu \geqq 1$. For $\nu \geqq 1$ we may take $W_{2,0}^{(\nu)}(\tau, R^\mu)$ as a common domain for the operators S and D. In this setting S and D commute and each commutes with multiplication by a scalar (element in R). Note that $\alpha \in R$, as an operator on functions in $W_{2,0}^{(\nu)}(\tau, R^\mu)$ so that $D\alpha = \alpha D$, must be distinguished from the constant function with value α for which, of course, $(D\alpha)(t) = 0$. The operators D, S and multiplication by a scalar all commute with the coordinate projections; that is, if $x \in W_{2,0}^{(\nu)}(\tau, R^\mu)$ and $(x)_j = x_j$ denotes the jth coordinate of x, $j = 1, \cdots, \mu$, then $(Sx)_j = Sx_j$ and $(Dx)_j = Dx_j$, $j = 1, \cdots, \mu$.

Now let $\mathcal{P}_\mu(S, D)$ be the ring of all $\mu \times \mu$ matric polynomials in two indeterminates S and D. By treating S and D as scalars the elements of $\mathcal{P}_\mu(S, D)$ are identified in the standard way with the $\mu \times \mu$ matrices over the polynomials in S and D with real coefficients. We denote by $C_0^{(\infty)}(\tau, R^\mu)$ the class of all $x \in W_{2,0}^{(\nu)}(\tau, R^\mu)$ such that $D^k x$ exists for integers $k \geqq 0$ and we let $\mathfrak{A}_\mu(S, D)$ denote the algebra of operators obtained from the elements of $\mathcal{P}_\mu(S, D)$ by respectively identifying the indeterminates S and D with the operator of shift and differentiation on $C_0^{(\infty)}(\tau, R^\mu)$. If $Q(D, S) \in \mathfrak{A}_\mu(S, D)$ has degree k in D, then $Q(D, S)$ is also a bounded linear operator from $W_{2,0}^{(k+r)}(\tau, R^\mu)$ into $W_{2,0}^{(r)}(\tau, R^\mu)$, $k, r = 0, 1, 2, \cdots$. Since $C_0^{(\infty)}(\tau, R^\mu)$ is dense in each $W_{2,0}^{(\nu)}(\tau, R^\mu)$ it follows that any polynomial identities in $\mathcal{P}_\mu(S, D)$ are operator identities in $\mathfrak{A}_\mu(S, D)$ if the domain, $W_{2,0}^{(\nu)}(\tau, R^\mu)$, of the operators is chosen so that they are all bounded. In particular, because S and D commute as operators with the coordinate projections, the elements in $\mathfrak{A}_\mu(S, D)$ may be identified with corresponding matrices of operators as was done for the elements of $\mathcal{P}_\mu(S, D)$.

With these remarks in mind it is clear that the solution $x(\cdot, 0, u)$ of (2.2) is the restriction to $[-h, \tau]$ of the solution $x \in W_{2,0}^{(1)}(\tau, R^n)$ of the equation

$$(I_n D - A_{-1} SD - A_0 - A_1 S)x = Bu$$

wherein u denotes the extension to $W_{2,0}^{(0)}(\tau, R^m)$ of the specified control function (that is, we extend the control function u by the condition $u(t) = 0$, $t < 0$). We now define $Q(D, S)$ by the equation

$$Q(D, S) = I_n D - A_{-1} SD - A_0 - A_1 S \tag{4.1}$$

and let

$$P(D, S) = \text{adj } Q(D, S) \tag{4.2}$$

in which "adj" denotes the transposed matrix of cofactors. We have the following basic relation between these two operators:

$$Q(D, S)P(D, S) = P(D, S)Q(D, S) = I_n \det Q(D, S).$$

Moreover,

$$P(D, S) = \sum_{j=0}^{n-1} P_j(D)S^j = \sum_{j=0}^{n-1} \hat{P}_j(s)D^j \tag{4.3}$$

where the $n \times n$ matric polynomials $P_j(D)$, $\hat{P}_j(S)$ are at most of degree $n-1$ in their arguments. We have used these relations in [4] to study the attainable sets

$\mathcal{A}(\tau)$ for equation (2.2). *From here on in this and in the remaining sections we confine our attention to the case where B is $n \times 1$ (that is, the control u is a scalar function.)* In this case our results have an appealing simplicity. Although the same ideas work for B being $n \times m$ with $m > 1$, the theorems tend to become untidy in these cases.

Using the polynomials $P_j(D)$ in (4.3), we define

$$K(D) = [P_0(D)B, P_1(D)B, \cdots, P_{n-1}(D)B]. \tag{4.4}$$

Here the multiplications $P_j(D)B$ are understood in the operator context and *not* with B as a constant function on which $P_j(D)$ acts. Let J_μ denote the $\mu \times \mu$ nilpotent matrix

$$J_\mu = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{4.5}$$

and let

$$\tilde{J}_\mu = [0, I_{\mu-1}] \tag{4.6}$$

denote the $(\mu - 1) \times \mu$ matrix obtained from J_μ by deleting the last row. We may then state Theorem 5.2 in [4] in the following form:

THEOREM 4.1. *Let B be $n \times 1$ and $\tau > nh$. If $\psi \in X$, then there is $u \in U_0 = W_{2,0}^{(0)}(\tau, R)$ such that the corresponding solution $x \in W_{2,0}^{(1)}(\tau, R^n)$ of*

$$Q(D, S)x = Bu \quad (a.e.\ on\ (-\infty, \tau]) \tag{4.7}$$

satisfies

$$x_\tau = \psi \tag{4.8}$$

if and only if there is a function $\omega \in W_2^{(n)}([\tau - h, \tau], R^n)$ such that

$$K(D)\omega = \tilde{\psi} \quad on\ [\tau - h, \tau] \tag{4.9}$$

and

$$\tilde{J}_n D^i \omega(\tau) = \tilde{J}_n J_n^* D^i \omega(\tau - h), \qquad i = 0, 1, \cdots, n-1. \tag{4.10}$$

(*We note that $\tilde{J}_n J_n^* = [I_{n-1}, 0]$.*)

Theorem 4.1 is an extension of a result in [27]. In [32, Chap. 5] there is a similar algebraic approach to solving periodic boundary value problems for retarded functional differential equations.

Now let the operator $K(D)$ in (4.4) be written in polynomial form as

$$K(D) = \sum_{i=0}^{n-1} K_i D^{n-1-i} \tag{4.11}$$

where the K_i, $i = 0, 1, \cdots, n-1$ are $n \times n$ real matrices (constants). We then have

LEMMA 4.1. *The matrix K_0 in* (4.11) *and the matrix $C_n[A_{-1}, B]$ are column equivalent, hence, have the same rank.*

Proof. There is a well-known algorithm for computing $\operatorname{adj}(\lambda I_n - A)$, λ complex and A an $n \times n$ matrix over the complex numbers. This may be found in [12], [14] or [36] and we restate it as follows:

$$
\begin{aligned}
&\text{(a)} \quad \operatorname{adj}(\lambda I_n - A) = \sum_{i=1}^{n} \lambda^{n-i} F_i, \\
&\text{(b)} \quad F_1 = I_n, \qquad \theta_1 = -\operatorname{tr} A, \\
&\text{(c)} \quad F_{i+1} = AF_i + \theta_i I_n, \\
&\qquad \theta_i = -(1/i)\operatorname{tr}(AF_i), \qquad i = 1, \cdots, n-1.
\end{aligned} \tag{4.12}
$$

Now the coefficient K_0 in (4.11) is determined completely by the polynomial $\hat{P}_{n-1}(S)$ in (4.3). Since $\hat{P}_{n-1}(S)$ is of degree at most $n-1$ we may write

$$
\hat{P}_{n-1}(S) = \sum_{i=0}^{n-1} M_i S^i \tag{4.13}
$$

for some constant $n \times n$ matrices M_i, $i = 0, 1, \cdots, n-1$. From the definitions of $\hat{P}_{n-1}(S)$ and K_0 in (4.3) and (4.11), respectively, it follows that

$$
K_0 = [M_0 B, M_1 B, \cdots, M_{n-1} B]. \tag{4.14}
$$

From (4.1) and (4.2) we see that $\hat{P}_{n-1}(S) = \operatorname{adj}(I_n - A_{-1}S)$. Thus taking $\lambda = 1$ and $A = A_{-1}S$ in (4.12) we may calculate the matrices M_i. Hence

$$
\operatorname{adj}(I_n - A_{-1}S) = \sum_{i=1}^{n} F_i \tag{4.15}
$$

and

$$
F_{i+1} = S(A_{-1}F_i - I_n(1/i)\operatorname{tr}(A_{-1}F_i)) \tag{4.16}
$$

for $i = 1, \cdots, n-1$, with $F_1 = I_n$. It follows that

$$
F_i = M_{i-1} S^{i-1}, \qquad i = 1, \cdots, n, \tag{4.17}
$$

and for some real constants $\alpha_1, \cdots, \alpha_{n-1}$,

$$
F_i = (A_{-1}^{i-1} - \sum_{k=1}^{i-1} \alpha_{i-k} A_{-1}^{k-1}) S^{i-1}, \qquad i = 1, \cdots, n. \tag{4.18}
$$

Combining equations (4.14), (4.17) and (4.18), we get

$$
K_0 = [B, A_{-1}B - \alpha_1 B, \cdots, A_{-1}^{n-1}B - \sum_{k=1}^{n-1} \alpha_{n-k} A_{-1}^{k-1} B]
$$

from which it is obvious that K_0 is column equivalent to $C_n[A_{-1}, B]$.

Our goal in this section is to develop necessary and sufficient conditions for controllability of (2.2) on $\mathcal{I} = [0, \tau]$, $\tau > nh$, so in view of Proposition 2.2 we may assume rank $C_n[A_{-1}, B] = n$ in our subsequent analysis. By Lemma 4.1 then K_0 in (4.11) is invertible so (4.9) can be written

$$
D^{n-1}\omega + \sum_{i=1}^{n-1} K_0^{-1} K_i D^{n-1-i}\omega = K_0^{-1}\tilde{\psi}. \tag{4.19}
$$

This may be converted to an equivalent first order linear system. First, let $\delta = n(n-1)$ and let Ω be the $\delta \times 1$ vector given by

$$\Omega = \begin{bmatrix} \omega \\ D\omega \\ \vdots \\ D^{n-2}\omega \end{bmatrix}. \tag{4.20}$$

Now let G denote the $\delta \times \delta$ matrix given in

$$G = \begin{bmatrix} 0 & I_n & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & I_n \\ \hat{K}_{n-1} & \hat{K}_{n-2} & \dots & \hat{K}_1 \end{bmatrix}, \tag{4.21}$$

where

$$\hat{K}_i = -K_0^{-1}K_i, \qquad i = 1, \cdots, n-1. \tag{4.22}$$

Finally let

$$\hat{\psi} = K_0^{-1}\tilde{\psi}, \qquad \psi \in X, \tag{4.23}$$

and let $\tilde{B}$ be the $\delta \times n$ matrix given by

$$\tilde{B} = [0, \cdots, 0, I_n]^*. \tag{4.24}$$

Using the notation given in (4.20) through (4.24), we may write (4.9) (or (4.19)) in the form

$$\dot{\Omega}(t) = G\Omega(t) + \tilde{B}\hat{\psi}(t) \quad \text{on } [\tau - h, \tau]. \tag{4.25}$$

To formulate the boundary conditions equivalent to (4.10) we introduce the $\delta \times \delta$ matrix

$$\mathcal{J} = \operatorname{diag}[J_n, \cdots, J_n] \tag{4.26}$$

and the $(n-1)^2 \times \delta$ matrix

$$\tilde{\mathcal{J}} = \operatorname{diag}[\tilde{J}_n, \cdots, \tilde{J}_n], \tag{4.27}$$

where $\tilde{J}_n$ is given in (4.6) with $\mu = n$ and J_n, as in (4.5), is the nilpotent matrix from which $\tilde{J}_n$ is obtained. We now partition the matrices $\mathcal{J}$ and $\tilde{\mathcal{J}}$ as

$$\mathcal{J} = \begin{bmatrix} \mathcal{J}_1 \\ \vdots \\ \mathcal{J}_{n-1} \end{bmatrix}, \qquad \tilde{\mathcal{J}} = \begin{bmatrix} \tilde{\mathcal{J}}_1 \\ \vdots \\ \tilde{\mathcal{J}}_{n-1} \end{bmatrix}, \tag{4.28}$$

where the $\mathcal{J}_i$ and $\tilde{\mathcal{J}}_i$, $i=1,\cdots,n-1$, are $n\times\delta$ and $(n-1)\times\delta$ matrices, respectively. The boundary conditions (4.10) may now be written in terms of Ω as

$$\tilde{\mathcal{J}}\Omega(\tau)=\tilde{\mathcal{J}}\mathcal{J}^*\Omega(\tau-h), \tag{4.29a}$$

$$\tilde{\mathcal{J}}_{n-1}\dot{\Omega}(\tau)=\tilde{\mathcal{J}}_{n-1}\mathcal{J}^*\dot{\Omega}(\tau-h). \tag{4.29b}$$

Recalling that (2.2) is controllable on $[0,\tau]$ if and only if $\mathcal{A}(\tau)=X$, we may combine Theorem 4.1, Proposition 2.2 and Lemma 4.1 to get the following:

LEMMA 4.2. *Let B be $n\times 1$ and $\tau>nh$. The system* (2.2) (*i.e.*, $Q(D,S)x=Bu$) *is controllable on $\mathcal{I}=[0,\tau]$ and only if* rank $C_n[A_{-1},B]=n$ *and for every $\psi\in X$ the two point boundary value problem* (4.25), (4.29) *has a solution Ω.*

The homogeneous equation corresponding to (4.25) is

$$\dot{\Omega}(t)=G\Omega(t) \tag{4.30}$$

and the controllability criteria can be given in another familiar way in terms of this equation. The precise statement is this:

THEOREM 4.2. *Let B be $n\times 1$ and $\tau>nh$. In order that* (2.2) (*i.e.*, $Q(D,S)x=Bu$) *be controllable on* $[0,\tau]$ *it is necessary and sufficient that* rank $C_n[A_{-1},B]=n$ *and the homogeneous problem* (4.30), (4.29) *have only the trivial solution* ($\Omega(t)\equiv 0$).

Proof. Suppose first that (2.2) is controllable on $[0,\tau]$. Then rank $C_n[A_{-1},B]=n$ and (4.25), (4.29) must have a solution Ω for every $\psi\in X$. For a given $\psi\in X$, the boundary conditions (4.29) may be written, using (4.25), in the form

$$\tilde{\mathcal{J}}\Omega(\tau)-\tilde{\mathcal{J}}\mathcal{J}^*\Omega(\tau-h)=0, \tag{4.31a}$$

$$\tilde{\mathcal{J}}_{n-1}(G\Omega(\tau)-\mathcal{J}^*G\Omega(\tau-h))=\tilde{\mathcal{J}}_{n-1}(\mathcal{J}^*\tilde{B}\hat{\psi}(\tau-h)-\tilde{B}\hat{\psi}(\tau)). \tag{4.31b}$$

Now define $\delta\times\delta$ matrices M, N by

$$M=\begin{bmatrix}\tilde{\mathcal{J}}\\ \tilde{\mathcal{J}}_{n-1}G\end{bmatrix},\qquad N=\begin{bmatrix}\tilde{\mathcal{J}}\mathcal{J}^*\\ \tilde{\mathcal{J}}_{n-1}\mathcal{J}^*G\end{bmatrix}, \tag{4.32}$$

and a $\delta\times 1$ matrix $\Gamma(\psi)$ by

$$\Gamma(\psi)=\begin{bmatrix}0\\ \tilde{\mathcal{J}}_{n-1}(\mathcal{J}^*\tilde{B}\hat{\psi}(\tau-h)-\tilde{B}\hat{\psi}(\tau))\end{bmatrix}. \tag{4.33}$$

With these the boundary conditions (4.31) become

$$M\Omega(\tau)-N\Omega(\tau-h)=\Gamma(\psi). \tag{4.34}$$

Let $\Omega^0=\Omega(\tau-h)\in R^\delta$ and write the solution of (4.25) as

$$\Omega(t)=e^{G(t-\tau+h)}\Omega^0+\int_{\tau-h}^{t}e^{G(t-s)}\tilde{B}\hat{\psi}(s)\,ds. \tag{4.35}$$

The boundary condition (4.34) thus requires that Ω^0 satisfy the linear equation

$$(Me^{Gh}-N)\Omega^0=\Gamma(\psi)-M\int_{\tau-h}^{\tau}e^{G(\tau-s)}\tilde{B}\hat{\psi}(s)\,ds. \tag{4.36}$$

Our aim now is to show that the right-hand side of (4.36) can be any vector in R^{δ} by appropriate choice of $\psi \in X$. Let us define

$$\gamma(\psi) = \tilde{\mathscr{J}}_{n-1}(\mathscr{J}^* \tilde{B}\hat{\psi}(\tau - h) - \tilde{B}\hat{\psi}(\tau)).$$

One can readily show that

$$\gamma(\psi) = \tilde{J}_n J^* K_0^{-1}\psi(-h) - \tilde{J}_n K_0^{-1}\psi(0). \tag{4.37}$$

If we define the mapping $L : L_2([\tau - h, \tau], R^n) \to R^{\delta}$ by

$$Lv = \int_{\tau-h}^{\tau} e^{G(\tau - s)}\tilde{B}v(s)\, ds, \tag{4.38}$$

then (4.36) may be written

$$(Me^{Gh} - N)\Omega^0 = \Gamma(\psi) - ML\hat{\psi} = \begin{bmatrix} -\tilde{\mathscr{J}}L\hat{\psi} \\ \gamma(\psi) - \tilde{\mathscr{J}}_{n-1}GL\hat{\psi} \end{bmatrix}. \tag{4.39}$$

Now consider the set $X(y) = \{\psi \in X | \psi(-h) = \psi(0) = y\}$ where $y \in R^n$. For each y this is dense in $L_2([-h, 0], R^n)$. Both Γ and L are linear transformations so the image of X under the mapping T defined by $T\psi = \Gamma(\psi) - ML(\hat{\psi})$ is a subspace of R^{δ} and hence is closed. On the other hand given $v \in L_2([\tau - h, \tau], R^n)$ and any $y \in R^n$ there exists a sequence $\psi^{\nu} \in X(y)$ such that $\hat{\psi}^{\nu}$ converges to v in L_2-norm. Since L is continuous it follows (cf. (4.39) and (4.37)) that Im $T|_X$ contains the set

$$\left\{ \begin{bmatrix} -\tilde{\mathscr{J}}L(v) \\ (\tilde{J}_n J^* - \tilde{J}_n)y - \tilde{\mathscr{J}}_{n-1}GL(v) \end{bmatrix} \middle| \, y \in R^n, v \in L_2([\tau - h, \tau], R^n) \right\}.$$

From (4.21) and (4.24) one sees that rank $C_{n-1}[G, \tilde{B}] = \delta$ which implies that Im $L = R^{\delta}$. Since $\tilde{\mathscr{J}}$ is $(n-1)^2 \times \delta$ with rank $(n-1)^2$ and $\tilde{J}_n J^* - \tilde{J}_n$ is $(n-1) \times n$ with rank $n-1$ it follows that

$$\text{Im } T|_X = R^{\delta}. \tag{4.40}$$

But (4.36) requires

$$\text{Im } (Me^{Gh} - N) \supset \text{Im } T|_X. \tag{4.41}$$

We conclude from (4.40) and (4.41) that

$$\text{rank } (Me^{Gh} - N) = \delta. \tag{4.42}$$

Now the homogeneous problem (4.30), (4.29) has a nontrivial solution if and only if there is an $\Omega^0 \neq 0$, $\Omega^0 \in R^{\delta}$ such that

$$(Me^{Gh} - N)\Omega^0 = 0. \tag{4.43}$$

This is clearly incompatible with (4.42) so controllability of $Q(D, S)x = Bu$ implies that (4.30), (4.29) have only the trivial solution. On the other hand, suppose rank $C_n[A_{-1}, B] = n$ and that the homogeneous problem (4.30), (4.29) has only the trivial solution. From (4.35) and (4.36) with $\psi = 0$ it is then evident that (4.43) can have only the trivial solution $\Omega^0 = 0$. This implies (4.42) so (4.36) can be solved for $\Omega^0 \in R^{\delta}$ for every $\psi \in X$; that is, $Q(D, S)x = Bu$ is controllable on $[0, \tau]$, $\tau > nh$. The theorem is proved.

Remark 4.1. Theorem 4.2 can be recast in an equivalent form in terms of the homogeneous boundary value problem corresponding to (4.9) and (4.10). Thus let B be $n \times 1$ and $\tau > nh$. In order that $Q(D, S)x = Bu$ be controllable on $[0, \tau]$ it is necessary and sufficient that rank $C_n[A_{-1}, B] = n$ and the homogeneous problem

$$K(D)\omega(t) = 0, \qquad t \in [\tau - h, \tau], \tag{4.44}$$

and (4.10) have only the trivial solution $\omega(t) \equiv 0$.

COROLLARY 4.1. *If B is $n \times 1$ and* rank $C_n[A_{-1}, B] = n$, *then the system $Q(D, S)x = Bu$ is Euclidean controllable on $\mathcal{I} = [0, \tau]$ provided $\tau > (n-1)h$.*

Proof. By assumption $\tau + h > nh$ so the calculations in the proof of Theorem 4.2 can be used with τ there replaced by $\tau + h$. If $y, z \in R^n$ and we define $X(y, z) = \{\psi \in X | \psi(-h) = y, \psi(0) = z\}$, then $X(y, z)$ is dense in $L_2([-h, 0], R^n)$. Consequently, by reasoning similar to that given in the proof of Theorem 4.2, it follows that

$$\{L\hat{\psi} | \psi \in X(y, z)\} = \{L\hat{\psi} | \psi \in L_2([-h, 0], R^n\},$$

where L is defined in (4.38). Hence for any $\psi^0, \psi^1 \in R^n$ there is a $\bar{\psi} \in X(\psi^0, \psi^1)$ such that $L\hat{\bar{\psi}} = 0$. From the definition of $\Gamma(\psi)$ in (4.33) one may show that for any $\psi^0 \in R^n$ there is a $\psi^1 \in R^n$ such that $\Gamma(\psi) = 0$ for all $\psi \in X(\psi^0, \psi^1)$. In particular, $\Gamma(\bar{\psi}) = 0$. Hence Ω^0 can be taken to be zero in (4.36), and then $\bar{\psi}, \Omega^0 = 0$ is a solution to (4.39). We have proved that for every $\psi^0 \in R^n$ there is a $\bar{\psi} \in X$ with $\bar{\psi}(-h) = \psi^0$ for which there is a solution of the boundary value problem (4.25) and (4.29) with ψ replaced by $\bar{\psi}$ and τ replaced by $\tau + h$. Hence $\bar{\psi} \in \mathcal{A}(\tau + h)$ and $Q(D, S)x = Bu$ is Euclidean controllable on $[0, \tau]$.

The proof of Theorem 4.2 has also established the following.

COROLLARY 4.2. *Let B be $n \times 1$ and $\tau > nh$. In order that $Q(D, S)x = Bu$ be controllable on $[0, \tau]$ it is necessary and sufficient that* rank $C_n[A_{-1}, B] = n$ *and*

$$\det [Me^{Gh} - N] \neq 0 \tag{4.45}$$

where G, M, N are given in (4.21) *and* (4.32).

A property of the points of a topological space $(E, \mathcal{T})$ will be said to be *generic* if the set of points of E having the property is a dense open subset of E. Let $\mathcal{L}$ denote the set of quadruples of real matrices of the form $L = (A_{-1}, A_0, A_1, B)$ where the A_i, $i = \pm 1, 0$ are $n \times n$ and B is $n \times 1$. With each $L \in \mathcal{L}$ we associate the corresponding control system

$$Q_L(D, S)x = Bu, \qquad L = (A_{-1}, A_0, A_1, B) \in \mathcal{L}, \tag{4.46}$$

where

$$Q_L(D, S) = (I_n - A_{-1}S)D - A_0 - A_1 S.$$

Define

$$\|L\|_{\mathcal{L}} = \|A_{-1}\| + \|A_0\| + \|A_1\| + \|B\|, \tag{4.47}$$

where the norms on the right are any convenient matrix norms. With addition of quadruples in $\mathcal{L}$ and multiplication by real scalars defined in the obvious natural way $\mathcal{L}$ is then a normed linear space of dimension $3n^2 + n$.

COROLLARY 4.3. *The controllability of* $Q_L(D, S)x = Bu$ *on* $[0, \tau]$, $\tau > nh$, *is a generic property of the space* $\mathscr{L}$.

Proof. Let $\mathscr{L}^n$ denote the collection of all $L \in \mathscr{L}$, $L = (A_{-1}, A_0, A_1, B)$ such that rank $C[A_{-1}, B] = n$. Then $\mathscr{L}^n$ is a dense open subset of $\mathscr{L}$. This result is given in [24, p. 100]. Since $\mathscr{L}^n$ is open and dense in $\mathscr{L}$ it suffices to prove that the set $\mathscr{L}_c$ of all $L \in \mathscr{L}^n$ such that $Q_L(D, S)x = Bu$ is controllable on $[0, \tau]$, $\tau > nh$, is an open dense subset of $\mathscr{L}^n$. From the proof of Lemma 4.1 we find that $K_0 = C_n[A_{-1}, B]$ $(I_n + \alpha)$ where α is an upper triangular $n \times n$ matrix with diagonal elements zero and elements above the diagonal depending continuously on A_{-1}. Hence at points $L \in \mathscr{L}^n$ the mappings $L \mapsto M_L$, $L \mapsto N_L$, $L \mapsto G_L$ are continuous. Here G_L is defined in (4.21) and M_L and N_L in (4.32). Consequently, $L \mapsto d(L)$, $L \in \mathscr{L}^n$, is continuous where we define $d(L)$ by

$$d(L) = \det (M_L e^{G_L h} - N_L).$$

Hence $\{L \in \mathscr{L}^n | d(L) \neq 0\}$ is open in $\mathscr{L}^n$. But by Corollary 4.2 this is precisely the set $\mathscr{L}_c$ so $\mathscr{L}_c$ is open in $\mathscr{L}^n$. The density of $\mathscr{L}_c$ in $\mathscr{L}^n$ is somewhat more involved. Suppose $\mathscr{L}_c$ is not dense in $\mathscr{L}^n$. Then there is $L^0 \in \mathscr{L}^n$ and $\varepsilon > 0$ such that $S_\varepsilon = \{L \in \mathscr{L} | \|L - L^0\|_{\mathscr{L}} \leqq \varepsilon\} \subset \mathscr{L}^n$ and $d(L) = 0$ for all $L \in S_\varepsilon$ by Corollary 4.2. If we define $L^0(z) = L^0 + z(L - L^0)$ with $z \in R$, $L \in S_\varepsilon$, then $d(L^0(z)) = 0$ for $-1 \leqq z \leqq 1$. From the construction of G_L, M_L, N_L we see that $d(L^0(z))$ is a meromorphic function of z which is zero on the real interval $[-1, 1]$. Hence we must in fact have $d(L^0(z)) = 0$, $z \in R$, $L \in S_\varepsilon$. Now define $H = (J_n, J_n, 0, e_n)$ where $e_n = [0, \cdots, 0, 1]^* \in R^n$ and J_n is defined in (4.5). One may readily verify that $C_n[J_n, e_n]$ has rank n so $H \in \mathscr{L}^n$. Moreover, in § 6 we show that the system $Q_H(D, S) = e_n u$ is controllable on $[0, \tau]$, $\tau > nh$. Hence $d(H) \neq 0$. But if we take $\delta = \|H - L^0\|_{\mathscr{L}}$, then $\delta > 0$ and $L = L^0 + (\varepsilon/\delta)(H - L^0) \in S_\varepsilon$. Consequently, $d(L^0 + z(\varepsilon/\delta)(H - L^0)) = 0$ for all $z \in R$. Taking $z = \delta/\varepsilon$ we get $d(H) = 0$, a contradiction. Thus $\mathscr{L}_c$ is dense in $\mathscr{L}_n$ and the corollary is proved.

Remark 4.2. It is clear that having closed range and finite deficiency of the solution operator (for system (2.2)), $u \to x_\tau(\cdot, 0, u)$, $u \in U$, is also a generic property of $\mathscr{L}$ even without the assumption that B is $n \times 1$; that is, B can be $n \times m$, $m \geqq 1$.

We can obtain some equivalent criteria for controllability of a system $Q(D, S)x = Bu$ which in some cases may be more readily applied than (4.45). For one of these it is convenient to use the commutator notation

$$[F, H] = FH - HF \tag{4.48}$$

for $n \times n$ matrices F, H. Also we use e_j to denote the jth column of I_n, $j = 1, 2, \cdots, n$. Now introduce the $n^2 \times n^2$ matrix $\mathscr{K}$,

$$\mathscr{K} = \begin{bmatrix} \mathscr{K}_{11} & -\mathscr{K}_{12} \\ \mathscr{K}_{21} & \mathscr{K}_{22} \end{bmatrix}, \tag{4.49}$$

where $\mathscr{K}_{11}$ is $\delta \times \delta$, $\mathscr{K}_{12}$ is $\delta \times n$, $\mathscr{K}_{21}$ is $n \times \delta$ and $\mathscr{K}_{22}$ is $n \times n$ and these are given by

(4.50)
$$\text{(i)}\quad \mathscr{K}_{11} = e^{Gh} - \mathscr{J}^*,$$

$$\text{(ii)}\quad \mathscr{K}_{12} = \begin{bmatrix} e_1 & 0 & \cdots & 0 & 0 \\ 0 & e_1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & e_1 & 0 \end{bmatrix},$$

$$\text{(iii)}\quad \mathscr{K}_{21} = [[\hat{K}_{n-1}, J_n^*], [\hat{K}_{n-2}, J_n^*], \cdots, [\hat{K}_1, J_n^*]],$$

$$\text{(iv)}\quad \mathscr{K}_{22} = [\hat{K}_{n-1}e_1, \hat{K}_{n-2}e_1, \cdots, \hat{K}_1 e_1, -e_1].$$

We then have the following corollary to Theorem 4.2.

COROLLARY 4.4. *Let B be $n \times 1$ and $\tau > nh$. Then $Q(D, S)x = Bu$ is controllable on $[0, \tau]$ if and only if* rank $C_n[A_{-1}, B] = n$ *and* rank $\mathscr{K} = n^2$.

Proof. By Remark 4.1 the system $Q(D, S)x = Bu$ is controllable on $[0, \tau]$ if and only if rank $C_n[A_{-1}, B] = n$ and (4.44) has only the trivial solution satisfying boundary conditions (4.10); that is,

$$\tilde{J}_n D^i \omega(\tau) = \tilde{J}_n J_n^* D^i \omega(\tau - h), \qquad i = 0, 1, \cdots, n-1. \tag{4.51}$$

Now the null space of $\tilde{J}_n$ is spanned by e_1 so (4.51) holds if and only if

$$D^i \omega(\tau) = J_n^* D^i \omega(\tau - h) + \mu_i e_1, \qquad i = 0, 1, \cdots, n-1, \tag{4.52}$$

for some scalars μ_i. Using (4.44) and the representation (4.11) for $K(D)$ we may write (cf. (4.22))

$$D^{n-1}\omega(t) = \sum_{i=1}^{n-1} \hat{K}_i D^{n-1-i}\omega(t).$$

Using this in (4.52) for $i = n-1$ and the rest of (4.52), we can then write the set of boundary conditions in the equivalent form

(4.53)
$$\text{(i)}\quad D^i \omega(\tau) = J_n^* D^i \omega(\tau - h) + \mu_i e_1, \qquad i = 0, 1, \cdots, n-2,$$

$$\text{(ii)}\quad \sum_{i=1}^{n-1} [\hat{K}_i, J_n^*] D^{n-1-i}\omega(\tau - h) = -\sum_{i=1}^{n-1} \mu_{n-1-i}\hat{K}_i e_1 + \mu_{n-1} e_1.$$

Now with Ω as in (4.20) we have $\dot{\Omega} = G\Omega$ so that

$$\Omega(\tau) = e^{Gh}\Omega(\tau - h).$$

With this (4.53(i)) becomes (cf. (4.50(i)))

$$\mathscr{K}_{11}\Omega(\tau - h) = \mathscr{K}_{12}\mu, \tag{4.54}$$

where

$$\mu = \begin{bmatrix} \mu_0 \\ \mu_1 \\ \vdots \\ \mu_{n-1} \end{bmatrix}.$$

Moreover, (4.53(ii)) becomes

$$\mathcal{K}_{21}\Omega(\tau-h) = -\mathcal{K}_{22}\mu. \tag{4.55}$$

Equations (4.54), (4.55) are now equivalent to

$$\mathcal{K}\begin{bmatrix}\Omega(\tau-h)\\ \mu\end{bmatrix} = 0. \tag{4.56}$$

If rank $\mathcal{K} = n^2$, then we must have $\Omega(\tau-h)=0$ and thus only the trivial solution to the homogeneous boundary value problem. Conversely, if we have only the trivial solution to the boundary value problem, then we must have $\mu = 0$ from (4.52) so that only $\Omega(\tau-h)=0$, $\mu=0$ satisfies (4.56) and hence rank $\mathcal{K}=n^2$.

The criterion in Corollary 4.4 is not easily applied if n is larger than 2 or 3. It leads, however, to an easily applied criterion in case $h>0$ is small.

THEOREM 4.3. *Let B be $n\times 1$ and* rank $C_n[A_{-1}, B]=n$ *and define* $\sigma = e_1+e_2+\cdots+e_n=[1,1,\cdots,1]^*$. *If*

$$\det[K_{n-1}\sigma, K_{n-2}\sigma, \cdots, K_1\sigma, K_0\sigma] \neq 0, \tag{4.57}$$

then the system $Q(D,S)x = Bu$ is controllable on $[0,\tau]$ for all $h>0$ sufficiently small.

Proof. Let $\mathcal{K}_{11}(h) = e^{Gh} - \mathcal{J}^*$ and use $\mathcal{K}(h)$ to indicate the explicit dependence of $\mathcal{K}$ in (4.49) on h. Define $E = \mathcal{K}_{11}(0) = I_\delta - \mathcal{J}^*$ and note that E is invertible since $(\mathcal{J}^*)^n = 0$. Now

$$\det \mathcal{K}(0) = \det\left(\begin{bmatrix} I_\delta & 0\\ -\mathcal{K}_{21}E^{-1} & I_n\end{bmatrix}\mathcal{K}(0)\right) = \det(\mathcal{K}_{22}+\mathcal{K}_{21}E^{-1}\mathcal{K}_{12}).$$

It is easy to verify that

$$(\mathcal{J}^*)^{k-1}\mathcal{K}_{12} = \begin{bmatrix} e_k & 0 & \cdots & 0 & 0\\ 0 & e_k & \cdots & 0 & 0\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \cdots & e_k & 0\end{bmatrix}, \qquad k=1,2,\cdots,n$$

and since $E^{-1} = \sum_{k=1}^{n-1}(\mathcal{J}^*)^{k-1}$ we find

$$E^{-1}\mathcal{K}_{12} = \begin{bmatrix} \sigma & 0 & \cdots & 0 & 0\\ 0 & \sigma & \cdots & 0 & 0\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \cdots & \sigma & 0\end{bmatrix}.$$

Thus from (4.50(iii)) we have

$$\mathcal{K}_{21}E^{-1}\mathcal{K}_{12} = [[\hat{K}_{n-1}, J_n^*]\sigma, \cdots, [\hat{K}_1, J_n^*]\sigma, 0].$$

But

$$[\hat{K}_i, J_n^*]\sigma = \hat{K}_i J_n^*\sigma - J_n^*\hat{K}_i\sigma = -K_i e_1 + (I - J_n^*)\hat{K}_i\sigma$$

since $J_n^*\sigma = \sigma - e_1$. Thus from (4.50(iv))

$$\begin{aligned}\mathscr{K}_{22} + \mathscr{K}_{21}E^{-1}\mathscr{K}_{12} &= (I - J_n^*)[\hat{K}_{n-1}\sigma, \hat{K}_{n-2}\sigma, \cdots, \hat{K}_1\sigma, -\sigma] \\ &= -(I - J_n^*)K_0^{-1}[K_{n-1}\sigma, K_{n-2}\sigma, \cdots, K_1\sigma, K_0\sigma].\end{aligned}$$

It follows that

$$\det \mathscr{K}(0) = (-1)^n (\det K_0^{-1}) \det [K_{n-1}\sigma, \cdots, K_1\sigma, K_0\sigma]$$

so $\mathscr{K}(0)$ is nonsingular if (4.57) holds. Thus rank $\mathscr{K}(h) = n^2$ for all sufficiently small $h > 0$ if (4.57) holds.

COROLLARY 4.5. *Let* B *be* $n \times 1$, rank $C_n[A_{-1}, B] = n$ *and* $\sigma = [1, 1, \cdots, 1]^*$. *If* (4.57) *holds, then the system* $Q(D, S)x = Bu$ *is controllable on* $[0, \tau]$ *for all* $h > 0$ *such that* $nh < \tau$ *except possibly for a finite number of such* h.

Proof. Condition (4.57) assures that $\det \mathscr{K}(0) \neq 0$. But $\det \mathscr{K}(h)$ is an entire function of h and since it is not identically zero it can vanish at only a finite number of points in the interval $0 < h < \tau/n$.

In § 6 we give an example of a system which is not controllable for any $h > 0$.

Remark 4.3. In the proof of Corollary 4.4 the columns of e^{Gt} were used, in effect, as a basis for the solution Ω of $\dot{\Omega} = G\Omega$. In checking controllability for concrete examples it is often possible to obtain a more convenient basis and apply Remark 4.1 directly. In this connection we note that the system of differential equations (4.44) is autonomous so we may shift the independent variable by the amount $\tau - h$ and express the boundary value problem on the interval $[0, h]$ as

$$K(D)\omega(t) = 0, \qquad t \in [0, h], \tag{4.58}$$

$$\tilde{J}_n D^i \omega(h) - \tilde{J}_n J_n^* D^i \omega(0) = 0, \qquad i = 0, 1, \cdots, n-1. \tag{4.59}$$

Now suppose B is $n \times 1$, $\tau > nh$, rank $C_n[A_{-1}, B] = n$ and that $\omega^{(j)}$, $j = 1, \cdots, \delta$, form a basis for the solutions of (4.58). Then every solution ω of (4.58) can be written in the form

$$\omega(t) = \sum_{j=1}^{\delta} c_j \omega^{(j)}(t) \tag{4.60}$$

for some scalars c_j, $j = 1, \cdots, \delta$. The boundary conditions (4.59) now give a linear homogeneous system of δ equations in the coefficients c_j and the system $Q(D, S)x = Bu$ is controllable on $[0, \tau]$ if and only if the determinant of the coefficient matrix $\mathscr{W}$ is nonzero. From (4.59) and (4.60) it is seen that the jth column of the $\delta \times \delta$ matrix $\mathscr{W}$ is formed from the elements $D^i\omega_{k+1}^{(j)}(h) - D^i\omega_k^{(j)}(0)$; $k = 1, \cdots, n-1$; $i = 0, 1, \cdots, n-1$.

COROLLARY 4.6. *Let* B *be* $n \times 1$, $\tau > nh$ *and* rank $C_n[A_{-1}, B] = n$. *If* $Q(D, S)x = Bu$ *is controllable on* $[0, \tau]$, *then for every complex* λ

$$K(\lambda)\mathscr{S}_\lambda^n \neq 0, \tag{4.61}$$

where $\mathcal{S}_\lambda^n$ is defined by

$$\mathcal{S}_\lambda^n = \begin{bmatrix} 1 \\ e^{-\lambda h} \\ \vdots \\ e^{-(n-1)\lambda h} \end{bmatrix}.$$

Proof. If $K(\lambda)\mathcal{S}_\lambda^n = 0$ for some complex λ, then $\omega(t) = e^{\lambda t}\mathcal{S}_\lambda^n$ is a nontrivial solution of the homogeneous boundary value problem (4.10), (4.44) (or (4.58), (4.59)). This implies, by Remark 4.1 (or Remark 4.3), that $Q(D, S)x = Bu$ is not controllable on $[0, \tau]$.

COROLLARY 4.7. *Let $\Delta(D) = \det K(D)$. The following two statements are equivalent:*

(i) *If ρ is a scalar solution of $\Delta(D)\rho(t) = 0$ for $t \leqq \tau$ such that $K(D)\mathcal{S}^n\rho(t) = 0$, $t \leqq \tau$, then $\rho(t) \equiv 0$, $t \leqq \tau$. ($\mathcal{S}^n\rho$ is defined in* (3.2).)

(ii) *$K(\lambda)\mathcal{S}_\lambda^n \neq 0$ for every complex λ.*

Proof. If $K(\lambda)\mathcal{S}_\lambda^n = 0$ for some λ, then $\Delta(\lambda) = 0$ and $\rho(t) = e^{\lambda t}$ satisfies the conditions in statement (i) but is not identically zero. Thus (i) implies (ii). To establish the converse we first suppose that ρ is a scalar solution of $\Delta(D)\rho(t) = 0$, $t \leqq \tau$. Let $\lambda_1, \cdots, \lambda_r$ be the distinct roots of $\Delta(\lambda) = 0$ and let the multiplicity of λ_i be m_i, $i = 1, \cdots, r$. Then there are scalars μ_{ij}, $i = 1, \cdots, r$; $j = 0, 1, \cdots, m_i - 1$ such that

$$\rho(t) = \sum_{i=1}^{r} \sum_{j=0}^{m_i-1} \mu_{ij} t^j e^{\lambda_i t}. \tag{4.62}$$

Note that $\sum_{i=1}^r m_i = \delta$. Let $D_\lambda = \partial/\partial\lambda$. We suppose now that also $K(D)\mathcal{S}^n\rho(t) = 0$, $t \leqq \tau$. Interpreting $D = \partial/\partial t$ and noting that D_λ and D commute, we then get

$$K(D)\mathcal{S}^n\rho(t) = \sum_{i=1}^{r} \sum_{j=1}^{m_i-1} \mu_{ij}[D_\lambda^j(e^{\lambda t}K(\lambda)\mathcal{S}_\lambda^n)]_{\lambda=\lambda_i} = 0 \tag{4.63}$$

for $t \leqq \tau$. Using Leibniz's rule for differentiating a product, we see that (4.63) can be written

$$\sum_{i=1}^{r} \sum_{j=0}^{m_i-1} \sum_{k=0}^{j} \mu_{ij}\binom{j}{k} t^k e^{\lambda_i t}[D_\lambda^{j-k}K(\lambda)\mathcal{S}_\lambda^n]_{\lambda=\lambda_i} = 0. \tag{4.64}$$

Interchanging the order of summation of j and k in (4.64), we obtain

$$\sum_{i=1}^{r} \sum_{k=0}^{m_i-1} \left(\sum_{j=k}^{m_i-1} \mu_{ij}\binom{j}{k}[D_\lambda^{j-k}K(\lambda)\mathcal{S}_\lambda^n]_{\lambda=\lambda_i} \right) t^k e^{\lambda_i t} = 0 \tag{4.65}$$

for $t \leqq \tau$. Since the functions $t^k e^{\lambda_i t}$ appearing in (4.65) are linearly independent, it follows that

$$\sum_{j=k}^{m_i-1} \mu_{ij}\binom{j}{k}[D_\lambda^{j-k}K(\lambda)\mathcal{S}_\lambda^n]_{\lambda=\lambda_i} = 0 \tag{4.66}$$

for $i = 1, \cdots, r$ and $k = 0, 1, \cdots, m_i - 1$. If one examines the equations (4.66) in the order $k = m_i - 1$, $m_i - 2$, etc., one finds that when (ii) holds then all $\mu_{ik} = 0$. That is, $\rho(t) \equiv 0$ for $t \leqq \tau$ so (ii) implies (i) and the corollary is proved.

Remark 4.4. In analyzing numerous examples of systems $Q(D, S)x = Bu$ with B being $n \times 1$ and satisfying rank $C_n[A_{-1}, B] = n$ all the cases of systems found to be not controllable on $[0, \tau]$, $\tau > nh$, also satisfied $K(\lambda)\mathcal{S}_\lambda^n = 0$ for some λ. This suggests the possibility that (4.61) and rank $C_n[A_{-1}, B] = n$ are sufficient for controllability of an n-dimensional system $Q(D, S)x = Bu$ on $[0, \tau]$, $\tau > nh$. In case $n = 2$ these conditions are sufficient as was shown in [20]. The proof for 2-dimensional systems depends heavily on the fact that $\delta = n(n-1) = n$ when $n = 2$. For $n \geqq 3$ it is an open question (cf. Example 6.9). We have obtained a partial converse to the necessity of condition (4.61). Suppose the operator $K(D)$ is such that there exists a nontrivial scalar differential operator $q(D)$ with constant coefficients and order $\nu \leqq n$ so that for any solution ω of

$$K(D)\omega(t) = 0, \qquad \tau - h \leqq t \leqq \tau, \tag{4.67}$$

and

$$D^i\omega_{j+1}(\tau) = D^i\omega_j(\tau - h), \qquad i = 0, 1, \cdots, n-1, \quad j = 1, \cdots, n-1 \tag{4.68}$$

it follows that

$$q(D)\omega_i(t) = 0, \quad t \leqq \tau, \quad i = 1, \cdots, n. \tag{4.69}$$

Now a solution of (4.67) on $[\tau - h, \tau]$ must have the form

$$\omega(t) = \sum_{i=1}^{r} \sum_{j=1}^{m_i - 1} c_{ij} t^j e^{\lambda_i t} \tag{4.70}$$

for some constant n-vectors c_{ij}. (The λ_i and m_i are as in the proof of Corollary 4.7.) Then ω as in (4.70) must satisfy $K(D)\omega(t) = 0$ for all $t \leqq \tau$. Hence $I\Delta(D)\omega(t) = [\text{adj } K(D)]K(D)\omega(t) = 0$ and, in particular, $\Delta(D)\omega_1(t) = 0$ for all $t \leqq \tau$. But (4.69) implies

$$q(D)[\omega_{j+1}(t) - \omega_j(t-h)] = 0, \quad t \leqq \tau, \quad j = 1, \cdots, n-1. \tag{4.71}$$

Since the order of $q(D)$ is at most n, the conditions (4.68) along with (4.71) imply

$$\omega_{j+1}(t) - \omega_j(t-h) = 0, \quad t \leqq \tau, \quad j = 1, \cdots, n-1.$$

Hence $\omega(t) = \mathcal{S}^n\omega_1(t)$, $t \leqq \tau$. Thus the hypothesis of (i) of Corollary 4.7 is satisfied by $\rho = \omega_1$. Now if rank $C_n[A_{-1}, B] = n$ and (4.61) holds, then $\omega_1(t) \equiv 0$ by Corollary 4.7. It follows that $\omega(t) \equiv 0$ so $Q(D, S)x = Bu$ (B is $n \times 1$) is controllable on $[0, \tau]$, $\tau > nh$. An application of this approach is given in § 6 (cf. Examples 6.4, 6.7). Note that when $n = 2$ we can take $q(D) = \Delta(D)$ in the above discussion.

In verifying (4.61) it is useful to bear in mind that if $h > 0$ is an algebraic number and $\lambda \neq 0$ is algebraic, then Lindemann's theorem [15] implies that $1, e^{-\lambda h}, \cdots, e^{-(n-1)\lambda h}$ are linearly independent over the field of algebraic numbers. If $h > 0$ is algebraic and the entries in B and in the matrices A_i, $i = \pm 1, 0$, are algebraic, then the distinct roots $\lambda_1, \cdots, \lambda_r$ of $\Delta(\lambda) = \det K(\lambda) = 0$ are algebraic. If rank $C_n[A_{-1}, B] = n$ and $\lambda_i \neq 0$, then one has $K(\lambda_i)\mathcal{S}_{\lambda_i}^n \neq 0$ by Lindemann's theorem providing $K(\lambda_i) \neq 0$. In any case (4.61) need only be checked for those λ

such that $\Delta(\lambda) = 0$ since $K(\lambda)$ is nonsingular otherwise. In particular if $\Delta(0) \neq 0$, if h and the entries of B and the A_i, $i = \pm 1, 0$, are all algebraic, then (4.61) will be fulfilled (provided $K(\lambda) \neq 0$ for any root of $\Delta(\lambda) = 0$).

Remark 4.5. It is also interesting to note that according to Theorem 4.1, if $t \mapsto \rho(t)$, $t \in R$, is any real valued analytic function, then $\psi \in \mathscr{A}(\tau)$, $\tau > nh$, when the corresponding $\tilde{\psi}$ is given by

$$K(D)\mathscr{S}^n \rho(t) = \tilde{\psi}(t), \qquad \tau - h \leqq t \leqq \tau.$$

In view of the computations given in the proof of Corollary 4.7, $\mathscr{A}(\tau)$ for $\tau > nh$ contains all the functions $\theta \mapsto \psi_{\lambda,\nu}(\theta)$, $\theta \in [-h, 0]$, $\lambda \in C$, $\nu = 0, 1, 2, \cdots$, for which the associated $\tilde{\psi}_{\lambda,\nu}$ has the form

$$\tilde{\psi}_{\lambda,\nu}(t) = D_\lambda^\nu e^{\lambda t} K(\lambda) \mathscr{S}_\lambda^n, \qquad \tau - h \leqq t \leqq \tau.$$

(Strictly speaking, $\mathscr{A}(\tau)$ for $\tau > nh$ contains the real and imaginary parts of these $\psi_{\lambda,\nu}$.) Moreover, the set $\{\psi_{\lambda,\nu} | \lambda \in C;\ \nu = 0, 1, 2, \cdots\}$ is linearly independent (cf. the proof that the μ_{ij} are all zero in (4.63)). Using Theorem 3.1, we see that

$$\mathscr{A}(\tau) \supset \text{cl span } \{\psi_{\lambda,\nu} | \lambda \in C;\ \nu = 0, 1, \cdots\}.$$

See Remark 5.1 later for some related comments.

5. Canonical systems. In this section we assume that the system $Q(D, S)x = Bu$ has coefficient matrices A_i, B of a special form. Specifically we assume that n-vector $B = e_n = [0, \cdots, 0, 1]^*$ and the $n \times n$ real matrices A_i are given by

$$A_i = \begin{bmatrix} \alpha_i & \beta_i \\ \gamma_i & \delta_i \end{bmatrix}, \qquad i = \pm 1, 0 \tag{5.1}$$

where α_i is $(n-1) \times (n-1)$, $i = \pm 1, 0$ with $\alpha_{-1} = J_{n-1}$ as in (4.5) and $\beta_{-1} = e_{n-1}$. We call such systems canonical systems. This is no loss of generality for systems which satisfy rank $C_n[A_{-1}, B] = n$ since they can be put in this canonical form by a change of variables $x = Py$ where P is a constant nonsingular $n \times n$ matrix [24]. In fact, if

$$\det (A_{-1} - \lambda I_n) = (-1)^n [\lambda^n - \sum_{i=0}^{n-1} a_i \lambda^i] \tag{5.2}$$

and

$$G_{-1} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_0 & a_1 & a_2 & \cdots & a_{n-1} \end{bmatrix} \tag{5.3}$$

then with $P = FH^{-1}$ where $F = C_n[A_{-1}, B]$, $H = C_n[G_{-1}, e_n]$, the equivalent system in y will have G_{-1} in place of A_{-1} and e_n in place of B. Thus in this section we assume the systems considered are already in the canonical form specified above.

PROPOSITION 5.1. *Let $Q(D, S)x = Bu$ be a canonical system and for $\lambda \in C$ let*

$$\mathcal{S}_\lambda^n = \begin{bmatrix} 1 \\ e^{-\lambda h} \\ \cdot \\ \cdot \\ \cdot \\ e^{-(n-1)\lambda h} \end{bmatrix}.$$

The following conditions are equivalent:

$$K(\lambda)\mathcal{S}_\lambda^n \neq 0, \tag{5.4}$$

$$\operatorname{rank}[\alpha(\lambda), \beta(\lambda)] = n-1, \tag{5.5}$$

where

$$\alpha(\lambda) = \lambda(I_{n-1} - \alpha_{-1}e^{-\lambda h}) - (\alpha_0 + \alpha_1 e^{-\lambda h}), \tag{5.6}$$

$$\beta(\lambda) = \beta_0 + e^{-\lambda h}(\lambda\beta_{-1} + \beta_1). \tag{5.7}$$

Proof. We have

$$Q(D, S) = \begin{bmatrix} I_{n-1}D - \alpha_{-1}DS - \alpha_0 - \alpha_1 S & -\beta_{-1}DS - \beta_0 - \beta_1 S \\ \gamma_{-1}DS - \gamma_0 - \gamma_1 S & D - \delta_{-1}DS - \delta_0 - \delta_1 S \end{bmatrix}.$$

As a polynomial in D and S we get, since $B = e_n$,

$$\sum_{j=0}^{n-1} P_j(D)BS^j = \operatorname{adj} Q(D, S)\, e_n = [g_1(D, S), \cdots, g_n(D, S)]^*, \tag{5.8}$$

where $g_i(D, S)$ is the cofactor of the ith entry in the last row of $Q(D, S)$. We note that these do not depend on the γ_i or δ_i, $i = \pm 1, 0$. Now (cf. (4.4))

$$K(\lambda)\mathcal{S}_\lambda^n = \sum_{j=0}^{n-1} P_j(\lambda)Be^{-j\lambda h}$$

so from (5.8) it is evident that

$$K(\lambda)\mathcal{S}_\lambda^n = \begin{bmatrix} g_1(\lambda, e^{-\lambda h}) \\ \vdots \\ g_n(\lambda, e^{-\lambda h}) \end{bmatrix}.$$

Thus $K(\lambda)\mathcal{S}_\lambda^n = 0$ if and only if $g_i(\lambda, e^{-\lambda h}) = 0$ for $i = 1, \cdots, n$. Since the $g_i(\lambda, e^{-\lambda h})$ are (except possibly for sign) the n determinants of size $(n-1)\times(n-1)$ from the rows of $[\alpha(\lambda), -\beta(\lambda)]$, we see that $K(\lambda)\mathcal{S}_\lambda^n = 0$ if and only if $\operatorname{rank}[\alpha(\lambda), \beta(\lambda)] < n-1$.

From our observation above relative to (5.8) it is clear that for a canonical system the operator $K(D) = [P_0(D)B, \cdots, P_{n-1}(D)B]$ does not depend on γ_i or δ_i, $i = \pm 1, 0$. Hence in any computation concerning controllability of a canonical

system we may assume that γ_i and δ_i, $i = \pm 1, 0$, are all zeros. When this is done we get a quite simple algorithm for constructing $K(D)$. The authors have working computer programs based on this algorithm which analyze the controllability of a canonical system.

We recall from (4.3) that

$$P(D, S) = \sum_{j=0}^{n-1} P_j(D)S^j = \sum_{j=0}^{n-1} \hat{P}_j(S)D^j. \tag{5.9}$$

Taking the last rows of the A_i, $i = \pm 1, 0$, to be zero, we define

$$A(S) = (I_n - A_{-1}S)^{-1}(A_0 + A_1S). \tag{5.10}$$

Now $A_{-1} = J_n$ is nilpotent and we get

$$A(S) = A_0 + \sum_{i=1}^{n-1} A_{-1}^{i-1}[A_{-1}A_0 + A_1]S^i. \tag{5.11}$$

Using (5.10), we have $Q(D, S) = (I_n - A_{-1}S)(I_nD - A(S))$ so

$$P(D, S) = \text{adj } Q(D, S) = \text{adj } (I_nD - A(S)) \text{ adj } (I_n - A_{-1}S). \tag{5.12}$$

The computation in (5.12) can be done iteratively using the Leverrier algorithm (4.12). First we have

$$\text{adj } (I_nD - A(S)) = \sum_{i=1}^{n} D^{n-i}F_i(S), \tag{5.13}$$

where

$$F_1(S) = I_n, \tag{5.14}$$

$$F_{i+1}(S) = A(S)F_i(S) - I_n \text{ tr } (A(S)F_i(S))/i,$$

$i = 1, 2, \cdots, n-1$. Secondly, we find that

$$\text{adj } (I_n - A_{-1}S) = (I_n - A_{-1}S)^{-1} = \sum_{k=0}^{n-1} A_{-1}^k S^k. \tag{5.15}$$

Combining (5.13) and (5.15) in (5.11), we get

$$P(D, S) = \sum_{i=0}^{n-1} \sum_{k=0}^{n-1} F_{n-i}(S)A_{-1}^k D^i S^k.$$

A comparison of this with (5.9) yields

$$\hat{P}_j(S) = \sum_{k=0}^{n-1} F_{n-j}(S)A_{-1}^k S^k, \qquad j = 0, 1, \cdots, n-1. \tag{5.16}$$

Moreover, for $j = 0, 1, \cdots, n-1$,

$$P_j(D) = \frac{1}{j!}\frac{\partial^j}{\partial S^j}P(D, 0) = \frac{1}{j!}\sum_{i=0}^{n-1} \hat{P}_i^{(j)}(0)D^i. \tag{5.17}$$

The coefficient matrices K_i, $i = 0, 1, \cdots, n-1$, in (4.11) can now be calculated; in

fact, the $(j+1)$th column of K_i is given by

$$\sum_{\nu=0}^{j} \frac{1}{(j-\nu)!} F_{i+1}^{(j-\nu)}(0) A_{-1}^{\nu} B, \qquad i = 0, 1, \cdots, n-1.$$

That is,

$$K_i = [F_{i+1}(0)B, \sum_{\nu=0}^{1} \frac{1}{(1-\nu)!} F_{i+1}^{(1-\nu)}(0) A_{-1}^{\nu} B, \cdots, \sum_{\nu=0}^{n-1} \frac{1}{(n-1-\nu)!} F_{i+1}^{(n-1-\nu)}(0) A_{-1}^{\nu} B], \tag{5.18}$$

where the $F_i(S)$ are given recursively by (5.14) and $A(S)$ there is given in (5.11). To recursively compute the $F_{i+1}^{(\nu)}(0)$ we just note that

$$A(0) = A_0, \qquad A^{(k)}(0) = k! A_{-1}^{k-1}(A_{-1}A_0 + A_1) \tag{5.19}$$

for $k = 1, 2, \cdots, n$, and

$$F_{i+1}^{(\nu)}(0) = \sum_{k=0}^{\nu} \binom{\nu}{k} [A^{(k)}(0) F_i^{(\nu-k)}(0) - I_n \operatorname{tr}(A^{(k)}(0) F_i^{(\nu-k)}(0))/i], \tag{5.20}$$

$i = 1, 2, \cdots, n-1$, $\nu = 0, 1, \cdots, n-1$. The authors have implemented (5.18), (5.19) and (5.20) in a computer program which will compute the operators $K(D)$ and analyze the controllability of neutral systems. The program has been used without discernible error in numerous examples where the coefficient matrices were randomly chosen from the ring of integers.

It is useful to note that for any canonical system

$$K_0 = [e_n, e_{n-1}, \cdots, e_1], \tag{5.21}$$

where e_i is the ith column of the $n \times n$ identity matrix, $i = 1, \cdots, n$. (This follows readily from (5.18) since $F_1(s) = I_n$, $B = e_n$ and $A_{-1} = J_n$.) Consequently, K_0 is an involution, i.e., $K_0^2 = I_n$. Thus for canonical systems the matrix G in (4.21) can be readily constructed with no round-off error or additional work due to inverting K_0.

Remark 5.1. The proof of Proposition 5.1 and Remark 4.5 have another aspect of some interest. Let A be the infinitesimal generator of the strongly continuous semigroup $T(t)$, $t \geqq 0$, defined by $T(t)\phi = x_t(\cdot, \phi, 0)$, $\phi \in W_2^{(1)}([-h, 0], R^n)$ (see [17], [18]). Then $\sigma(A)$, the spectrum of A, consists only of eigenvalues and they are the roots of

$$\det Q(z, e^{-zh}) = 0. \tag{5.22}$$

If either of the equivalent conditions (5.4) and (5.5) is satisfied for a canonical system $Q(D, S)x = Bu$ and if z satisfies (5.22), then the vector $K(z)\mathcal{S}_z^n$ is the associated eigenvector. Indeed, $K(z)\mathcal{S}_z^n \neq 0$ and

$$K(z)\mathcal{S}_z^n = \operatorname{adj} Q(z, e^{-zh})B, \tag{5.23}$$

so by (5.22),

$$Q(z, e^{-zh})K(z)\mathscr{S}_z^n = 0.$$

Moreover, since rank $Q(z, e^{-zh}) = n-1$ by (5.5), the null space of $Q(z, e^{-zh})$ is spanned by the single vector $K(z)\mathscr{S}_z^n$. In light of these comments and Remark 4.5 it is evident that rank $C_n[A_{-1}, B] = n$, $K(\lambda)\mathscr{S}_\lambda^n \neq 0$, $\lambda \in C$, and $\tau > nh$ do imply, at least, a weaker kind of controllability mentioned by Banks and Manitius [8, Remark 5.4]. That is, if M_λ denotes the generalized eigenspace corresponding to $\lambda \in \sigma(A)$, then ϕ, $\psi \in \text{cl span}\{M_\lambda | \lambda \in \sigma(A)\}$ implies that there is a $u \in L_2[0, \tau]$ such that $x_\tau(\cdot, \phi, u) = \psi$.

6. Examples. Here we illustrate some of the results in §§ 4 and 5.

Example 6.1. Let $n = 2$, $B = [b_1, b_2]^*$. (b_1 and b_2 are real.) As mentioned in Remark 4.4, when $n = 2$ the conditions

$$\text{rank}\,[B, A_{-1}B] = 2 \tag{6.1}$$

and

$$K(\lambda)\mathscr{S}_\lambda^2 \neq 0, \quad \text{all complex } \lambda, \quad \mathscr{S}_\lambda^2 = \begin{bmatrix} 1 \\ e^{-\lambda h} \end{bmatrix} \tag{6.2}$$

are sufficient as well as necessary that $Q(D, S)x = Bu$ be controllable on $[0, \tau]$, $\tau > 2h$. To compute $K(\lambda)$ we note that for $n = 2$ the mapping $A \mapsto \text{adj } A$ is a linear operator on the collection of all 2×2 matrices. (This is not the case for $n \times n$ matrices if $n \geqq 3$.) Thus we get that $P(D, S) = \text{adj } Q(D, S) = \text{adj}\,(I_2 D - A_{-1}DS - A_0 - A_1 S)$ is given by

$$P(D, S) = P_0(D) + P_1(D)S, \tag{6.3}$$

where

$$\begin{aligned} P_0(D) &= \text{adj}\,(I_2 D - A_0), \\ P_1(D) &= \text{adj}\,(-A_1 - A_{-1}D). \end{aligned} \tag{6.4}$$

Let

$$A_i = \begin{bmatrix} \alpha_i & \beta_i \\ \gamma_i & \delta_i \end{bmatrix}, \qquad i = \pm 1, 0, \tag{6.5}$$

and

$$K(\lambda) = [P_0(\lambda)B, P_1(\lambda)B] = \begin{bmatrix} \kappa_1(\lambda) \\ \kappa_2(\lambda) \end{bmatrix}. \tag{6.6}$$

Using (6.4), we find the 1×2 matrices $\kappa_i(\lambda)$ are given by

$$\begin{aligned} \kappa_1(\lambda) &= [(\lambda - \delta_0)b_1 + \beta_0 b_2, -(\delta_1 + \delta_{-1}\lambda)b_1 + (\beta_1 + \beta_{-1}\lambda)b_2], \\ \kappa_2(\lambda) &= [\gamma_0 b_1 + (\lambda - \alpha_0)b_2, (\gamma_1 + \gamma_{-1}\lambda)b_1 - (\alpha_1 + \alpha_{-1}\lambda)b_2]. \end{aligned} \tag{6.7}$$

Let λ_1, λ_2 be the roots of the quadratic equation $\Delta(\lambda) = \det K(\lambda) = 0$. Then

$Q(D, S)x = Bu$ is controllable on $[0, \tau]$, $\tau > 2h$, if and only if (6.1) holds and

$$(6.8)\quad \begin{bmatrix} (\lambda_i - \delta_0)b_1 + \beta_0 b_2 + e^{-\lambda_i h}[-(\delta_1 + \delta_{-1}\lambda_i)b_1 + (\beta_1 + \beta_{-1}\lambda_i)b_2] \\ \gamma_0 b_1 + (\lambda_i - \alpha_0)b_2 + e^{-\lambda_i h}[(\gamma_1 + \gamma_{-1}\lambda_i)b_1 - (\alpha_1 + \alpha_{-1}\lambda_i)b_2] \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

for $i = 1, 2$. If the system is in canonical form, then $\alpha_{-1} = 0$, $\beta_{-1} = 1$, $b_1 = 0$ and $b_2 = 1$. From (6.8) in this case we deduce that a canonical system $Q(D, S)x = Bu$ is controllable on $[0, \tau]$, $\tau > 2h$, if and only if either

$$(6.9)\quad \beta_0 = 0 \quad \text{and} \quad \alpha_0 + \beta_1 + \alpha_1 e^{\beta_1 h} \neq 0$$

or

$$(6.10)\quad \beta_0 \neq 0 \quad \text{and} \quad \lambda_i + \beta_1 + \beta_0 e^{\lambda_i h} \neq 0, \qquad i = 1, 2,$$

where λ_1, λ_2 are the roots of

$$(6.11)\quad \lambda^2 + (\beta_1 - \alpha_0)\lambda + \alpha_1\beta_0 - \alpha_0\beta_1 = 0.$$

(This was reported earlier in [20].)

Example 6.2. In Corollary 4.3 we proved that the controllable systems $Q(D, S)x = Bu$ correspond to a dense open subset of the space $\mathscr{L}$ of quadrupoles (A_{-1}, A_0, A_1, B). We note here that the system corresponding to the point $L = (A_{-1}, A_0, A_1, B)$ is controllable if and only if the system corresponding to (A_{-1}, A_0, A_1, sB), $s \neq 0$, is controllable. A control function u for the second induces the control function su for the first. With that in mind consider the systems corresponding to points in $\mathscr{L}$ of the form

$$(6.12)\quad L(s) = (A_{-1}, A_0, A_1, se_n)$$

where e_n is the nth column of the $n \times n$ identity matrix and

$$(6.13)\quad \operatorname{rank}[\alpha_0 + \alpha_1, \beta_0 + \beta_1] < n - 1,$$

where α_i, β_i, $i = 0, 1$ are determined from A_i as given in (5.1). The system corresponding to $L(0)$ in (6.12) is obviously not controllable, whereas the controllability of $L(s)$ for $s \neq 0$ is the same as that of $L(1)$. But because of (6.13) the canonical system $Q(D, S)x = Bu$ associated with $L(1)$ is not controllable. This follows from Proposition 5.1 and Corollary 4.6 when one notes that (cf. (5.6) and (5.7))

$$[\alpha(0), \beta(0)] = [-(\alpha_0 + \alpha_1), \beta_0 + \beta_1];$$

that is, (6.13) implies $K(0)\mathscr{S}_0^n = 0$. In particular, the canonical system $Q(D, S)x = Bu$ corresponding to $(A_{-1}, A_0, -A_0, se_n)$ is not controllable on $[0, \tau]$, $\tau > nh$. Thus we have an $(n^2 + 1)$-dimensional linear variety in $\mathscr{L}$ (cf. Corollary 4.3) correspond to canonical neutral systems $Q(D, S)x = Bu$ that are not controllable.

Example 6.3. Consider the point $H = (J_n, J_n, 0, e_n)$ in $\mathscr{L}$ which was used in the proof of Corollary 4.3. It was claimed that the system $Q(D, S)x = Bu$ corresponding to this point is controllable on $[0, \tau]$, $\tau > nh$, for every $h > 0$. For this example $n \geqq 2$ is fixed and we let $J = J_n$ and $I = I_n$ for notational convenience. We have

$$(6.14)\quad Q(D, S) = ID - (DS + 1)J, \qquad B = e_n.$$

One can readily establish that

$$P(D, S) = \text{adj } Q(D, S) = \sum_{j=0}^{n} \left[\sum_{i=j}^{n-1} \binom{i}{j} J^i D^{n-1+j-i} \right] S^j$$

and consequently the $P_j(D)$ in (4.3) are given by

$$P_j(D) = \sum_{i=j}^{n-1} \binom{i}{j} J^i D^{n-1+j-i}, \qquad j = 0, 1, \cdots, n-1. \tag{6.15}$$

Note that $J^i e_n = e_{n-i}$, $i = 0, 1, \cdots, n-1$, so that

$$P_{j-1}(D)B = \sum_{i=1}^{n+1-j} \binom{n-1}{j-1} e_i D^{i+j-2}, \qquad j = 1, \cdots, n. \tag{6.16}$$

Since the matrix operator $K(D)$ is given by

$$K(D) = [P_0(D)B, P_1(D)B, \cdots, P_{n-1}(D)B]$$

we see from (6.16) that the homogeneous equation $K(D)\omega = 0$ can be written out as the following system of differential equations in the components ω_j of ω:

$$\sum_{j=1}^{n+1-i} \binom{n-i}{j-1} D^{i+j-2} \omega_j = 0, \qquad i = 1, 2, \cdots, n. \tag{6.17}$$

We wish to establish that the only solution ω of (6.17) which satisfies the boundary conditions

$$\begin{aligned} D^i \omega_{j+1}(\tau) &= D^i \omega_j(\tau - h), \\ i &= 0, 1, \cdots, n-1, \quad j = 1, 2, \cdots, n-1, \end{aligned} \tag{6.18}$$

is trivial (i.e. $\omega(t) \equiv 0$). We note that the system (6.17) when written in reverse order of i has the following triangular structure:

$$\begin{aligned} &D^{n-1}\omega_1 = 0, \\ &D^{n-2}\omega_1 + D^{n-1}\omega_2 = 0, \\ &\qquad \cdot \\ &\qquad \cdot \\ &\qquad \cdot \\ &\omega_1 + (n-1)D\omega_2 + \cdots + D^{n-1}\omega_2 = 0. \end{aligned} \tag{6.19}$$

This structure together with the boundary conditions enables one to show inductively that

$$D^{n-k}\omega_j = 0, \qquad j = 1, \cdots, k, \tag{6.20}$$

for $1 \leqq k \leqq n$. Indeed (6.20) for $k = 1$ is the first equation of (6.19). Now suppose (6.20) holds for some k satisfying $1 \leqq k \leqq n-1$. Using this in (6.17) with $i = n-k$, we get

$$D^{n-k-1}\omega_1 + D^{n-1}\omega_{k+1} = 0. \tag{6.21}$$

Differentiating this, we obtain $D^n \omega_{k+1} = 0$ by (6.20). Thus $D^{n-1}\omega_{k+1}$ is a constant

and (6.18) and (6.20) imply this constant is zero. Repetition of this argument leads to the conclusion that $D^{n-k}\omega_{k+1}=0$. This with (6.20) gives $D^{n-k-1}\omega_j=c_j$, a constant, $j=1,\cdots,k+1$. But since $D^{n-1}\omega_{k+1}=0$ we see from (6.21) that $c_1=0$. Using the boundary conditions (6.18) with $i=n-k-1$, we conclude successively that $c_j=0$ also for $j=2,\cdots,k+1$. Thus (6.20) holds with k replaced by $k+1$. It follows by induction that (6.20) holds for $k=n$; *that is*, $\omega_j(t)\equiv 0$, $j=1,\cdots,n$. Thus the system determined by $H=(J_n, J_n, 0, e_n)\in\mathscr{L}$ is controllable on $[0,\tau]$, $\tau>nh$, as claimed. Note that in this example $\Delta(\lambda)=\det K(\lambda)=\lambda^{\delta}$.

Example 6.4. Let $n=3$, $B=[0,0,1]^*$ and

$$[A_{-1}, A_0, A_1]=\begin{bmatrix} 0 & 1 & 0 & -1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}. \tag{6.22}$$

Then we find

$$[K_0, K_1, K_2]=\begin{bmatrix} 0 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{6.23}$$

Let $\sigma=[1,1,1]^*$, the matrix in Theorem 4.3. Then the determinant of the matrix $[K_0\sigma, K_1\sigma, K_2\sigma]$ is 4. Hence by Corollary 4.5 the canonical system determined from (6.22) is controllable for all but a finite number of values of h, $0<h<\tau/3$. The roots of $\Delta(\lambda)=\det K(\lambda)=0$ are -1, -1, i, $-i$, 0, 0. One can check that $K(i)\mathscr{S}_i^3=0$ if and only if $h=(4\nu+1)\pi/2$, $\nu=0,1,\cdots(h>0)$. Thus the system is not controllable if $h=(4\nu+1)\pi/2$, $\nu=0,1,\cdots$. However, the system is controllable on $[0,\tau]$, $\tau>3h$, if $h\neq(4\nu+1)\pi/2$. To show this we note that the operator $K(D)$ determined from (6.23) is given by

$$K(D)=\begin{bmatrix} -D & -(D+1) & D^2 \\ -(D+1) & D(D+1) & 0 \\ D(D+1) & D+1 & 0 \end{bmatrix}. \tag{6.24}$$

If we premultiply this by the matrix operator (with constant determinant)

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & -1 & D \\ 0 & 0 & 1 \end{bmatrix},$$

we conclude that the equation $K(D)\omega=0$ is equivalent to the system

$$\begin{aligned} D^2\omega_1 \qquad\qquad + D^2\omega_3 &= 0, \\ (D^2+1)(D+1)\omega_1 \qquad\quad &= 0, \\ D(D+1)\omega_1+(D+1)\omega_2 \qquad &= 0. \end{aligned} \tag{6.25}$$

From the last two equations in (6.25) it follows that

$$\tilde{q}(D)\omega_1=\tilde{q}(D)\omega_2=0, \tag{6.26}$$

where

$$\tilde{q}(D) = (D^2+1)(D+1). \tag{6.27}$$

The boundary conditions (4.68) together with (6.26) imply (as in Remark 4.4) that

$$\omega_2(t) = \omega_1(t-h), \quad \text{all } t. \tag{6.28}$$

From the form of $\tilde{q}(D)$ in (6.26) as given in (6.27) we see that

$$\begin{aligned} \omega_1(t) &= \mu_1 e^{it} + \mu_2 e^{-it} + \mu_3 e^{-t}, \\ \omega_2(t) &= \mu_1 e^{i(t-h)} + \mu_2 e^{-i(t-h)} + \mu_3 e^{-(t-h)} \end{aligned} \tag{6.29}$$

for some constants μ_1, μ_2, μ_3. Substituting these into the last equation of (6.25) we obtain

$$\mu_1(1+i)(i+e^{-ih})\, e^{it} + \mu_2(1-i)(-i+e^{ih})\, e^{-it} \equiv 0.$$

Hence if $e^{ih} \neq i$, then $\mu_1 = \mu_2 = 0$ so $(D+1)\omega_1 = (D+1)\omega_2 = 0$. This with the first equation of (6.25) implies

$$q(D)\omega_i = 0, \qquad i = 1, 2, 3$$

where $q(D) = D^2(D+1)$. Since $q(D)$ is of degree 3 and

$$K(\lambda_j)\mathcal{S}^n_{\lambda_j} \neq 0 \quad \text{if } \lambda_j = 0, -1, \pm i \quad \text{and} \quad e^{ih} \neq i$$

we may apply Remark 4.4. It follows that the canonical system $Q(D, S)x = Bu$ associated with (6.22) is controllable on $[0, \tau]$, $\tau > 3h$, if and only if

$$h \neq (4\nu+1)\pi/2, \qquad \nu = 0, 1, 2, \cdots. \tag{6.30}$$

Example 6.5. Here again $n = 3$ and $B = [0, 0, 1]^*$. For

$$[A_{-1}, A_0, A_1] = \begin{bmatrix} 0 & 1 & 0 & -2 & -2 & -2 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & -1 & -2 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{6.31}$$

we find

$$[K_0, K_1, K_2] = \begin{bmatrix} 0 & 0 & 1 & -2 & -5 & 0 & 2 & -5 & -2 \\ 0 & 1 & 0 & -2 & 2 & 0 & -2 & 1 & 0 \\ 1 & 0 & 0 & 3 & 3 & 0 & 0 & 4 & 0 \end{bmatrix}. \tag{6.32}$$

We compute

$$\Delta(\lambda) = \det K(\lambda) = -(\lambda+1)(\lambda^2-2)(\lambda^3+4\lambda^2+9\lambda+8). \tag{6.33}$$

The equation $\Delta(\lambda) = 0$ has six distinct roots, -1, $\sqrt{2}$, $-\sqrt{2}$ and the roots of the cubic $\lambda^3+4\lambda^2+9\lambda+8 = 0$, which can be expressed exactly in terms of certain radicals. The canonical system $Q(D, S)x = Bu$ corresponding to (6.31) is not controllable on $[0, \tau]$, $\tau > 3h$, if $h = \ln 2$, since in this case $K(-1)\mathcal{S}^3_{-1} = 0$. For $\sigma = [1, 1, 1]^*$ we find $\det[K_0\sigma, K_1\sigma, K_2\sigma] = 11$. Corollary 4.5 then gives that the canonical system determined from (6.31) is controllable for all but a finite number of values of h, $0 < h < \tau/3$. Further analysis enables one to say that the system can

fail to be controllable only if $h>0$ is transcendental. Let λ_i, $i=1, 2, 3$, be the zeros of the cubic factor, $\lambda^3+4\lambda^2+9\lambda+8$, in $\Delta(\lambda)$ and let $\lambda_4=-1$, $\lambda_5=\sqrt{2}$, $\lambda_6=-\sqrt{2}$. For $j=1, \cdots, 6$ let $u(\lambda_j)\neq 0$ be a 3×1 column matrix such that $K(\lambda_j)u(\lambda_j)=0$. Then $\omega^{(j)}(t)=e^{\lambda_j t}u(\lambda_j)$, $j=1, \cdots, 6$, form a basis for the solutions of $K(D)\omega(t)=0$ in this example. One may choose $u(\lambda_4)=[1, 2, 4]^*$, $u(\lambda_5)=u(\lambda_6)=[0, 0, 1]^*$. For $j=1, 2, 3$ the components of $u(\lambda_j)$ are determined as quotients of polynomials in λ_j with integral coefficients so they may be taken to be algebraic numbers. We may now write the coefficient matrix $\mathcal{W}$ referred to in Remark 4.3. The structure of $\mathcal{W}$ is simple enough so that one may show

$$\det \mathcal{W}=2\sqrt{2}(2e^{-h}-1)g(e^{\lambda_1 h}, e^{\lambda_2 h}, e^{\lambda_3 h})$$

where $g(\xi, \eta, \zeta)$ is a nontrivial polynomial in the variables ξ, η, ζ with algebraic coefficients. Now it is not difficult to show that $\lambda_1, \lambda_2, \lambda_3$ are linearly independent over the rational integers so the same is true of $\lambda_1 h, \lambda_2 h, \lambda_3 h$ when $h\neq 0$. Hence we may apply an extended version of the Weierstrass-Lindemann theorem (cf. [37, p. 20]) to conclude that $g(e^{\lambda_1 h}, e^{\lambda_2 h}, e^{\lambda_3 h})\neq 0$ when $h>0$ is algebraic. Since $h=\ln 2$ is transcendental it follows that $\det \mathcal{W}\neq 0$ when h is algebraic. By Remark 4.3 the canonical system $Q(D, S)x=Bu$ corresponding to (6.31) is controllable for all h, $0<h<\tau/3$, except $\ln 2$ and a finite number of other transcendental numbers.

Example 6.6. Let $n=3$, $B=[0, 0, 1]^*$ and

$$[A_{-1}, A_0, A_1]=\begin{bmatrix} 0 & 1 & 0 & -1 & -1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{6.34}$$

The operator $K(D)$ is defined by

$$[K_0, K_1, K_2]=\begin{bmatrix} 0 & 0 & 1 & 0 & -2 & -2 & 1 & 2 & 1 \\ 0 & 1 & 0 & -1 & 0 & 1 & -1 & -2 & -1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & -1 & -1 \end{bmatrix}. \tag{6.35}$$

If $\sigma=[1, 1, 1]^*$, then

$$\det[K_0\sigma, K_1\sigma, K_2\sigma]=32. \tag{6.36}$$

It follows that the canonical system $Q(D, S)x=Bu$ corresponding to (6.34) is controllable for all but a finite number of $h>0$ satisfying $\tau>3h$. We calculate

$$\Delta(\lambda)=\det K(\lambda)=-\lambda^5(\lambda-1). \tag{6.37}$$

The necessary condition (4.61) is satisfied since

$$K(0)\mathcal{S}_0^3\neq 0 \quad \text{and} \quad K(1)\mathcal{S}_1^3\neq 0. \tag{6.38}$$

One can, in fact, prove the system is controllable on $[0, \tau]$, $\tau>3h$, for all $h>0$. Since the roots of $\Delta(\lambda)=0$ are quite simple one can calculate e^{Gt} analytically. One may then determine the matrix coefficients R_k, $k=1, \cdots, 6$, in the relation

$$Me^{Gh}-N=e^h R_1+\sum_{k=0}^{4}\frac{h^k}{k!}R_{k+2}. \tag{6.39}$$

By judicious use of elementary row and column operations one finds that

$$\det(Me^{Gh} - N) = 32e^{h}. \tag{6.40}$$

Example 6.7. Here we consider canonical systems of dimension $n \geqq 2$ with $A_1 = 0$ and

$$A_0 = \begin{bmatrix} \alpha_0 & 0 \\ \gamma_0 & \delta_0 \end{bmatrix}, \tag{6.41}$$

where the $(n-1)\times(n-1)$ matrix $\alpha_0 = \operatorname{diag}(a_1, a_2, \cdots, a_{n-1})$. In this case all elements of $K(D)$ are zero except for the (i,j) elements with $i+j = n+1$. The system $K(D)\omega = 0$ has the form

$$q_j(D)\omega_j = 0, \qquad j = 1, \cdots, n, \tag{6.42}$$

where

$$\begin{aligned} q_j(D) &= D^{j-1}(D-a_1)\cdots(D-a_{n-j}), \qquad j = 1, \cdots, n-1, \\ q_n(D) &= D^{n-1}. \end{aligned} \tag{6.43}$$

We can show that equations (6.42) and the boundary conditions (4.68) imply that

$$\omega = \mathcal{S}^n \omega_1. \tag{6.44}$$

Indeed, we note that for $j = 1, \cdots, n-1$,

$$Dq_j(D) = q_{j+1}(D)(D - a_{n-j}),$$

so from (6.42) we have

$$Dq_j(D)\omega_j = Dq_j(D)\omega_{j+1} = 0, \qquad j = 1, \cdots, n-1. \tag{6.45}$$

Since $Dq_j(D)$ is of degree n in D, equations (6.45) along with the boundary conditions (4.68) imply that

$$\omega_{j+1} = S\omega_j, \qquad j = 1, \cdots, n-1. \tag{6.46}$$

It follows that $\omega_j = S^{j-1}\omega_1$, $j = 1, \cdots, n$; that is, equation (6.44) is valid. Since

$$\begin{aligned} \Delta(\lambda) &= (-1)^n q_1(\lambda) q_2(\lambda) \cdots q_n(\lambda) \\ &= (-1)^n \lambda^{n(n-1)/2} (\lambda - a_1)^{n-1} (\lambda - a_2)^{n-2} \cdots (\lambda - a_{n-1}) \end{aligned} \tag{6.47}$$

it is clear from (6.42) and (6.43) that $\Delta(D)\omega_1 = 0$. As explained in Remark 4.4, it follows that in this example condition (4.61) is necessary and sufficient for controllability of the canonical system with $A_1 = 0$ and A_0 as specified above. We readily find that

$$K(\lambda)\mathcal{S}_\lambda^n = \begin{bmatrix} q_n(\lambda)\, e^{-(n-1)\lambda h} & & \\ & \cdot & \\ & \cdot & \\ q_2(\lambda)\, e^{-\lambda h} & & \\ q_1(\lambda) & & \end{bmatrix}. \tag{6.48}$$

We observe from (6.43) that

$$q_j(0)=0, \qquad j=2,\cdots,n,$$
$$q_1(0)=(-1)^{n-1}a_1a_2\cdots a_{n-1}.$$

Hence if $a_k=0$ for some $k=1,\cdots,n-1$ we have $K(0)\mathscr{S}_0^n=0$, and the system is not controllable. On the other hand, if

$$a_k\neq 0, \qquad k=1,\cdots,n-1, \tag{6.49}$$

then $q_1(0)\neq 0$ and

$$q_n(a_k)=a_k^{n-1}\neq 0, \qquad k=1,\cdots,n-1.$$

From (6.48) and (6.47) we see that if (6.49) holds, then $K(\lambda)\mathscr{S}_\lambda^n\neq 0$ for any root of $\Delta(\lambda)=0$. Thus the system considered here is controllable on $[0,\tau]$, $\tau>nh$, if and only if (6.49) is satisfied.

Example 6.8. Consider the canonical system of dimension 3 for which

$$A_i=\begin{bmatrix}\alpha_i & \beta_i\\ \gamma_i & \delta_i\end{bmatrix}, \qquad i=0,1, \tag{6.50}$$

where

$$\alpha_0=\begin{bmatrix}0 & 0\\ -1 & a\end{bmatrix}, \qquad \beta_0=\begin{bmatrix}0\\ 0\end{bmatrix},$$
$$\alpha_1=\begin{bmatrix}0 & 0\\ 0 & 4\end{bmatrix}, \qquad \beta_1=\begin{bmatrix}b^2\\ 0\end{bmatrix}.$$

In this case

$$K(\lambda)=\begin{bmatrix}0 & b^2(\lambda-a) & \lambda^2-4b^2\\ 0 & \lambda^2-b^2 & 0\\ \lambda(\lambda-a) & -3\lambda & 0\end{bmatrix}$$

so

$$\Delta(\lambda)=-\lambda(\lambda-a)(\lambda^2-b^2)(\lambda^2-4b^2).$$

By computing $K(\lambda_i)\mathscr{S}_{\lambda_i}^n$ and using (4.61) we find that for the system to be controllable it is necessary that

$$b\neq 0 \quad \text{and} \quad a\neq \pm b-3e^{\mp bh}. \tag{6.51}$$

A detailed analysis applying the method in Remark 4.3 shows that condition (6.51) is also sufficient that the system be controllable on $[0,\tau]$, $\tau>3h$.

Example 6.9. As a final example consider the second order homogeneous differential equation

$$K(D)\omega=0, \tag{6.52}$$

where

$$K(D)=\begin{bmatrix}0 & 0 & D^2+1/9\\ 0 & D^2+1/4 & 0\\ D^2+1 & 0 & 0\end{bmatrix}. \tag{6.53}$$

This has all of the immediately apparent features of a $K(D)$ operator associated with a canonical neutral system $Q(D, S)x = Bu$ with $n = 3$ (cf. (5.21) for the form of K_0 in such cases). We have $\Delta(\lambda) = -(\lambda^2+1)(\lambda^2+1/4)(\lambda^2+1/9)$. For any zero λ_1 of $\Delta(\lambda)$ one sees from (6.53) that the null space of $K(\lambda_1)$ has dimension one and any vector in this null space has two of its components equal to zero. Since none of the components of $\mathcal{S}_\lambda^3$ is ever zero it follows that condition (4.61) is satisfied; that is, $K(\lambda)\mathcal{S}_\lambda^3 \neq 0$ for any $\lambda \in C$. Now if we construct the matrices M and N in (4.32) associated with $K(D)$ as in (6.53) we find that

$$\det(M-N) = 0. \tag{6.54}$$

The matrix G in (4.21) has the eigenvalues $\pm i$, $\pm i/2$, $\pm i/3$ so e^{Gh} is a periodic function of h with period 12π. From (6.54) it follows that $\det(Me^{Gh}-N) = 0$ when $h = 12k\pi$ for any integer k. Thus if $K(D)$ were associated with a canonical neutral system $Q(D, S)x = Bu$ that system would not be controllable on $[0, \tau]$ for $\tau > 3h$ if $h = 12\pi$. This example would then demonstrate that (4.61) is not a sufficient condition for controllability of neutral systems $Q(D, S)x = Bu$ (cf. Remark 4.4). However, the $K(D)$ given in (6.53) does not provide a counter-example to the conjecture that if rank $C[A_{-1}, B] = n$, B is $n \times 1$ and $\tau > nh$, then (4.61) implies that $Q(D, S)x = Bu$ is controllable on $[0, \tau]$. Indeed, the operators $K(D)$ which come from canonical neutral systems $Q(D, S)x = Bu$ are determined by the A_i, $i = 0, 1$, in a rather intricate way. One can in fact show that if such a $K(D)$ has the form

$$K(D) = \begin{bmatrix} 0 & 0 & p_3(D) \\ 0 & p_2(D) & 0 \\ p_1(D) & 0 & 0 \end{bmatrix}, \tag{6.55}$$

where $p_i(D)$, $i = 1, 2, 3$, are monic polynomials of degree two, then the first two rows of the coefficient matrices A_0 and A_1 must have the respective forms

$$\begin{bmatrix} a & b & 0 \\ d & e & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & s & -b \\ 0 & -d & w \end{bmatrix}, \tag{6.56}$$

where $b(w+e) = 0$ and $d(a+s) = 0$. In all these cases at least two of the polynomials $p_i(D)$ have a common factor so $K(D)$ in (6.53) cannot arise from a canonical system $Q(D, S)x = Bu$ with $n = 3$. Moreover, one can check that in all cases (6.55) which do arise from such systems, the conjecture mentioned above (regarding the sufficiency of (4.61) for controllability) is valid.

REFERENCES

[1] H. T. BANKS, *Control of functional differential equations with function space boundary conditions*, Delay and Functional Differential Equations and Their Applications, K. Schmitt, ed., Academic Press, New York, 1972, pp. 1–16.

[2] H. T. BANKS AND M. Q. JACOBS, *An attainable sets approach to optimal control of functional differential equations with function space side conditions*, J. Differential Equations, 13 (1973), pp. 127–149.

[3] ———, *The optimization of trajectories of linear functional differential equations*, this Journal, 8 (1970), pp. 461–488.

[4] H. T. BANKS, M. Q. JACOBS AND C. E. LANGENHOP, *Characterization of the controlled states in $W_2^{(1)}$ of linear hereditary systems*, this Journal, 13 (1975), pp. 611–649.

[5] ———, *Applications of alternative methods to controllability of functional differential equations*, Optimal Control Theory and Its Applications, II, B. J. Kirby, ed., Springer-Verlag, New York, 1974, pp. 1–23.

[6] H. T. BANKS, M. Q. JACOBS AND M. R. LATINA, *The synthesis of optimal controls for linear, time optimal problems with retarded controls*, J. Optimization Theory and Appl., 8 (1971), pp. 319–366.

[7] H. T. BANKS AND G. A. KENT, *Control of functional differential equations of retarded and neutral type to target sets in function space*, this Journal, 10 (1972), pp. 567–594.

[8] H. T. BANKS AND A. MANITIUS, *Projection series for retarded functional differential equations with applications to optimal control problems*, J. Differential Equations, 18 (1975), pp. 296–332.

[9] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[10] M. A. CRUZ AND J. K. HALE, *Existence, uniqueness and continuous dependence of hereditary systems*, Ann. Mat. (4), 85 (1970), pp. 63–81.

[11] N. DUNFORD AND J. SCHWARTZ, *Linear Operators. Part I*, Interscience, New York, 1958.

[12] R. A. FRAZER, W. J. DUNCAN AND A. R. COLLAR, *Elementary Matrices*, Cambridge Univ. Press, Cambridge, England, 1938.

[13] R. GABASOV AND F. KIRILLOVA, *Qualitative Theory of Optimal Processes*, Nauka, Moscow, 1971.

[14] F. R. GANTMACHER, *The Theory of Matrices*, vols. 1 and 2, Chelsea, New York, 1960.

[15] A. O. GELFOND, *Transcendental and Algebraic Numbers*, Dover, New York, 1960.

[16] J. K. HALE, *Functional Differential Equations*, Applied Mathematical Sciences, vol. 3, Springer-Verlag, New York, 1971.

[17] J. K. HALE AND K. R. MEYER, *A Class of Functional Differential Equations of Neutral Type*, Mem. Amer. Math. Soc., 76 (1967).

[18] D. HENRY, *Linear autonomous functional differential equations of neutral type in Sobolev space $W_2^{(1)}$*, J. Differential Equations, 15 (1974), pp. 106–128.

[19] M. Q. JACOBS AND T. KAO, *An optimum settling problem for time-lag systems*, 40 (1972), pp. 687–707.

[20] M. Q. JACOBS AND C. E. LANGENHOP, *Controllable two dimensional neutral systems*, Mathematical Control Theory (Proceedings of a Conference, Zakopane, 1974), S. Dolecki, C. Olech, J. Zabczyk, eds., Polish Scientific Publishers, Warsaw, 1976, pp. 107–113.

[21] S. KURCYUSZ, *A local maximum principle for operator constraints and its application to systems with time lags*, to appear.

[22] S. KURCYUSZ AND A. OLBROT, *On the closure in W_1^q of the attainable subspace of time lag systems*, Tech. Rep. 7, Institute of Automatic Control, Technical University of Warsaw, Warsaw, Poland, 1974.

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

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

[25] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[26] A. MANITIUS AND A. OLBROT, *Controllability conditions for linear systems with delayed state and control*, Arch. Avtomat. i Telemenh., 17 (1972), pp. 119–131.

[27] S. A. MINJUK, *On complete controllability of linear systems with delay*, Differencial'nye Uravenija, 8 (1972), pp. 254–259.

[28] S. A. MINJUK AND N. N. STEPANJUK, *The theory of completely controllable linear systems with delay*, Ibid., 10 (1974), pp. 629–634.

[29] A. OLBROT, *Algebraic criteria of controllability to zero function for linear constant time-lag systems*, Control and Cybernetics, 2 (1973), pp. 59–77.

[30] H. POLLARD, *The Theory of Algebraic Numbers*, Carus Math. Monograph, John Wiley, New York, 1950.

[31] W. T. REID, *Some limit theorems for ordinary differential systems*, J. Differential Equations, 14 (1966), pp. 253–262.

[32] D. ROLEWICZ, *Equations with Transformed Argument, An Algebraic Approach*, Elsevier Scientific, Amsterdam, 1973.

[33] D. ROLEWICZ AND S. ROLEWICZ, *Equations in Linear Spaces*, Polish Scientific Publishers, Warsaw, 1968.

[34] D. L. RUSSELL, *A unified controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., 52 (1973), pp. 189–211.

[35] M. SCHECHTER, *Principles of Functional Analysis*, Academic Press, New York, 1971.

[36] D. M. WIBERG, *State Space and Linear Systems*, Schaum's Outline Series, McGraw-Hill, New York, 1971.

[37] C. L. SIEGEL, *Transcendental Numbers*, Ann. of Math. Studies, No. 16, Princeton University Press, Princeton, N.J., 1949.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 6, November 1976

LOCAL CONTROLLABILITY AND SUFFICIENT CONDITIONS IN SINGULAR PROBLEMS. II*

HENRY HERMES†

Abstract. Let $X, Y^2, \cdots, Y^m$ be analytic vector fields on an analytic n-manifold and $\mathscr{D}$ denote the control system $\dot{x}(t) = X(x) + \sum_{i=2}^{m} u_i(t) Y^i(x)$, $x(0) = p$. Our major goal is to give high order, computable, sufficient conditions to assure that the reference solution of $\mathscr{D}$ corresponding to $u \equiv 0$ has value at time t on the boundary of the set of all points attainable at time t for small values $t > 0$. This provides high order sufficient conditions for time optimality when the reference solution is singular. These conditions are phased in terms of elements of the Lie algebra generated by $X, Y^2, \cdots, Y^m$. We also show that quite general nonlinear systems can be approximated by systems of the above form and this approximation retains more information than the standard linearization about the reference solution.

Introduction. Let $X, Y^2, \cdots, Y^m$ be analytic vector fields on an n-dimensional analytic manifold M; I be a real interval of the form $[0, t_1]$, $t_1 > 0$ and $\mathbb{R}^m$ be real m-dimensional space. We denote by $\mathscr{D}$ the control system

$$\dot{x} = X(x) + \sum_{i=2}^{m} u_i(t) Y^i(x), \qquad x(0) = p \in M, \quad (\dot{x} = dx/dt) \tag{1}$$

where, unless stated otherwise, an admissible control is a Lebesgue measurable function $u = (u_2, \cdots, u_m) : I \to \mathbb{R}^m$ with $|u_i(t)| \leqq 1$ for $t \in I$, $i = 2, \cdots, m$. (Note that replacing Y^i with $W^i = \alpha_i Y^i$, $\alpha_i \geqq 0$, shows this formulation includes control bounds of the form $|u_i(t)| \leqq \alpha_i$.) Let $T^X(\cdot)p$ denote the solution of (1) corresponding to all $u_i(t) \equiv 0$ and $\mathscr{A}(t, p, \mathscr{D})$ be the set of points attainable at time t by solutions of $\mathscr{D}$ corresponding to all admissible controls. Our major goal is to obtain computable conditions to determine whether $T^X(t)p$ belongs to the interior of $\mathscr{A}(t, p, \mathscr{D})$ for all $t > 0$, or to the boundary, denoted $\partial\mathscr{A}(t, p, \mathscr{D})$, for sufficiently small $t > 0$.

Control is often introduced into a physical system to assure that a predetermined reference trajectory can be followed, even in the event of small perturbations or a somewhat inaccurate mathematical model. If we assume the control system modeled by (1), and $T^X(\cdot)p$ the reference trajectory, then $T^X(t)p \in$ int. $\mathscr{A}(t, p, \mathscr{D})$ for all $t > 0$ implies the ability to make minor "corrections" at p, i.e., local controllability at p. If, on the other hand, the control is to be used to make the system perform a desired task optimally (e.g., time optimally), one finds a necessary condition for $T^X(\cdot)p$ to be optimal is that $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{D})$. Our results have immediate applications to such problems, when the control system is of the form $\mathscr{D}$. In § 4 we show how very general systems can be approximated by systems of the form $\mathscr{D}$; the approximation given retains more information than a standard linearization.

Consider the real linear space, $V(M)$, of analytic vector fields on M as a real Lie algebra, with Lie product $[X, Y]$. For any collection of vector fields $\mathscr{C} \subset V(M)$

* Received by the editors April 8, 1975, and in final revised form September 15, 1975.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80302. This research was supported by the National Science Foundation under Grant GP27957.

let $L(\mathscr{C})$ denote the subalgebra generated by $\mathscr{C}$; $\mathscr{C}(p)=\{X(p): X\in\mathscr{C}\}$, while for notational convenience $(\operatorname{ad} X, Y)=[X, Y]$ and inductively $(\operatorname{ad}^k X, Y)=[(X, (\operatorname{ad}^{k-1}X, Y)]$. Define

$$\mathscr{S}^1=\{(\operatorname{ad}^j X, Y^i): j\geqq 0, i=2, \cdots, m\}. \tag{2}$$

A necessary and sufficient condition that int. $\mathscr{A}(t, p, \mathscr{D})\neq\varnothing\ \forall t>0$ is $\dim L(\mathscr{S}^1)(p)=n$ while a (first order) sufficient condition that $T^X(t)p\in$ int. $\mathscr{A}(t, p, \mathscr{D})\ \forall t>0$ is rank $\mathscr{S}^1(p)=n$ (see [1, Prop. 1.6], [2, Thm. 3.2] and [3, Prop. 3]). A (first order) necessary condition that $T^X(t)p\in\partial\mathscr{A}(t, p, \mathscr{D})$ for $t\in I$ is given by the Pontryagin maximum principle. Our interest is in deriving higher order conditions. It should be remarked that our results, to be of interest, really depend on the nonlinearity of the system $\mathscr{D}$. Indeed, in the case of a linear system, $\mathscr{F}$, on $\mathbb{R}^n$, i.e., $\dot{x}=Ax+Bu$ with A an $n\times n$ matrix and B an $n\times m$ matrix, we see $\dim L(\mathscr{S}^1)(p)=\operatorname{rank}\mathscr{S}^1(p)=\operatorname{rank}[B, AB, \cdots, A^{n-1}B]$; hence the necessary and sufficient condition ($\dim L(\mathscr{S}^1)(p)=n$) that int. $\mathscr{A}(t, p, \mathscr{F})\neq\varnothing\forall t>0$ is equivalent to the sufficient condition ($\operatorname{rank}\mathscr{S}^1(p)=n$) that $T^X(t)p\in$ int. $\mathscr{A}(t, p, \mathscr{F})\forall t>0$. The interesting case, for $\mathscr{D}$, will always be when $\dim L(\mathscr{S}^1)(p)=n$ but $\operatorname{rank}\mathscr{S}^1(p)<1$. For the linear system $\mathscr{F}$, it readily follows from the above that the property $T^X(t)p\in$ int. $\mathscr{A}(t, p, \mathscr{F})$ is "insensitive" to control bounds, i.e., if true for $|u_i(t)|\leqq 1$ it is true for $|u_i(t)|\leqq\alpha_i$, $\alpha_i>0$. This is *not* the case, when dealing with the nonlinear system $\mathscr{D}$.

In § 1, we review known results both for the case of bounded control (i.e., $|u_i(t)|\leqq 1$) and unbounded control. Specifically, if we let $\mathscr{D}^\infty$ denote the system (1) with u admissible requiring only that each control component u_i belong to the Lebesgue space $\mathscr{L}_1(I)$, Lemma 6.4, [4], or Lemma 3.4, [5], yields.

PROPOSITION 1.1. *$\operatorname{Dim} L\{Y^2, \cdots, Y^m\}(p)=n$ is a sufficient condition to assure that $T^X(t)p\in$ int. cl. $\mathscr{A}(t, p, \mathscr{D}^\infty)\ \forall t>0$.*

This result does not hold if we have finite control bounds since, as shown by Example 1.1, the contribution due to Lie products of the Y^i may, or may not, "override" the influence of X. In this example $T^X(\cdot)p$ is a singular arc (i.e., $\operatorname{rank}\mathscr{S}^1(T^X(\tau)p)<n$ for $\tau\in I$; see [1, Prop. 2.6] for verification that this is equivalent to the usual definition of singular arc) which is time optimal, i.e., $T^X(t)p\in\partial\mathscr{A}(t, p, \mathscr{D})$ if $|u_i(t)|\leqq 1$, but $T^X(t)p\in$ int. $\mathscr{A}(t, p, \mathscr{D}^\infty)\ \forall t>0$ (i.e., is not time optimal) if the control bounds are removed.

In § 2, we assume $\dim L\{Y^2, \cdots, Y^m\}(q)=k<n$ for q in a neighborhood of p. Then $L\{Y^2, \cdots, Y^m\}$ defines an involutive distribution of dimension k. With a further assumption that the vector field X is transverse to the integral manifolds (leaves) of this distribution, we obtain our main result, a high order sufficient condition that $T^X(t)p\in\partial\mathscr{A}(t, p, \mathscr{D})$ for small $t>0$, in Theorem 2.1. This is an extension (and simplification) of results in [1, § 4]. It is of interest to note (as shown in Corollary 2.1) that the assumptions of § 2 make the results of Theorem 2.1 independent of the bounds on the control components.

The most difficult case occurs when no restriction is placed on $\dim L\{Y^2, \cdots, Y^m\}(q)$ for q in a neighborhood of p. In § 3, we obtain some results for this case under the assumption that the algebra $L(\mathscr{S}^1)$ is nilpotent. (The case $L(\mathscr{S}^1)$ solvable is similar, but calculations of the exponentials of matrices becomes more difficult.)

1. Known results. If $X(x)\equiv 0$ in (1) and $\dim L\{Y^2,\cdots,Y^m\}(p)=n$ it follows from a theorem of Chow (see [6, satz B], or [7, Thm. 1] or [8, Thm. 1]) that $T^X(t)p\equiv p\in \text{int. } \mathscr{A}(t,p,\mathscr{D})\ \forall t>0$. In essence, the proof of Proposition 1.1 (as stated in the Introduction) shows that with $u_i\in\mathscr{L}_1(I)$ we can "override" the influence of the vector field X. Furthermore, one may slightly generalize Proposition 1.1 obtaining

COROLLARY 1.1. *If for every $\varepsilon>0$ there exists a $\tau\in[0,\varepsilon)$ such that* $\dim L\{Y^2,\cdots,Y^m\}(T^X(\tau)p)=n$, *then* $T^X(t)p\in\text{int. cl. }\mathscr{A}(t,p,\mathscr{D}^\infty)\ \forall t>0$.

The next example shows these results are not true with finite bounds on the control values. Here, and in other examples, vector functions will be written as row vectors for notational ease.

Example 1.1. Let $M=\mathbb{R}^3$, $X(x)=(1-(x_2^2+x_3^2),0,0)$, $Y^2(x)=(x_3^2,1,0)$, $Y^3(x)=(0,0,1)$ and $p=0$. Then $T^X(t)p=(t,0,0)$ and since $\dot{x}_1(t)\leqq 1$ which implies $x_1(t)\leqq t$ we see $\mathscr{A}(t,p,\mathscr{D})$ lies on one side of the hyperplane $x_1=t$; hence $T^X(t)p\in\partial\mathscr{A}(t,p,\mathscr{D})\ \forall t\leqq 0$.

Computation shows $(\text{ad}\,X,Y^2)(x)=(-2x_2,0,0)$, $(\text{ad}^jX,Y^2)(x)=0$ if $j\geqq 2$, $(\text{ad}\,X,Y^3)(x)=(-2x_3,0,0)$, $(\text{ad}^jX,Y^3(x)=0$ if $j\geqq 2$ so $\mathscr{S}^1=\{Y^2,Y^3,(\text{ad}\,X,Y^2),(\text{ad}\,X,Y^3)\}$ and rank $\mathscr{S}^1(p)=2$. Next, $-[Y^3,[Y^2,Y^3]]=(2,0,0)$; hence $\dim L\{Y^2,Y^3\}(p)=3$ (which shows $\dim L(\mathscr{S}^1)(p)=3$ and int. $\mathscr{A}(t,p,\mathscr{D})\neq\varnothing\ \forall t>0$). If we consider $u_i\in\mathscr{L}_1(I)$, Proposition 1.1 would give $T^X(t)p\in\text{int. cl. }\mathscr{A}(t,p,\mathscr{D}^\infty)\ \forall t>0$.

Returning to the system (1), if u is a fixed admissible control and one attempts to express a solution in the form $T^X(t)\circ y(t;u)$, then (see [9, §§ 3, 4]) y must satisfy

$$\dot{y}(t)=\sum_{i=2}^{m}u_i(t)\sum_{\nu=0}^{\infty}(-t)^\nu/\nu!\,(\text{ad}^\nu X,Y^i)(y),\qquad y(0)=p. \tag{3}$$

Conversely, if we let $y(t;u)(p)$ denote the solution of (3), at time t, corresponding to admissible control u, then $T^X(t)\circ y(t;u)(p)$ is a solution of (1). Letting $\mathscr{B}(t,p)$ be the attainable set at time $t\geqq 0$ for (3) we then have

$$\mathscr{A}(t,p,\mathscr{D})=T^X(t)\mathscr{B}(t,p).$$

Let $I(L(\mathscr{S}^1),p)$ denote the integral manifold of $L(\mathscr{S}^1)$ through p. The above equality immediately yields a part of Theorem 3.9 of [2], which we state as

PROPOSITION 1.2. $\mathscr{A}(t,p,\mathscr{D})\subset T^X(t)I(L(\mathscr{S}^1),p)\ \forall t>0$.

For unbounded control, conditions which insure $\mathscr{A}(t,p,\mathscr{D}^\infty)=T^X(t)I(L(\mathscr{S}^1),p)$ are given in [5].

If $m=n$ and $Y^2,\cdots,Y^n$ are involutive and linearly independent at p, the "behavior" of the vector field X on the $(n-1)$-dimensional integral manifold $I(L\{Y^2,\cdots,Y^n\},T^X(t)p)$ is fundamental. For example, suppose $X(p)=0$ so $T^X(t)p\equiv p$ and X is tangent to ("points to one side" of) $I(L\{Y^2,\cdots,Y^n\},p)$ at all points in a neighborhood of p on this $(n-1)$-dimensional manifold. Geometrically one then expects that, respectively, $\mathscr{A}(t,p,\mathscr{D})\subset I(L\{Y^2,\cdots,Y^n\},p)$ ($\mathscr{A}(t,p,\mathscr{D})$ lies on one side of $I(L\{Y^2,\cdots,Y^n\},p)$) for sufficiently small $t>0$. Let Z be any one form such that $Z(p)\neq 0$, $\langle Z(p),Y^i(p)\rangle=0$, $i=2,\cdots,n$, and let $q(s,p)=T^{Y_2}(s_2)\circ\cdots\circ T^{Y_n}(s_n)p$. Then the behavior of X on $I(L\{Y^2,\cdots,Y^n\},p)$ can be determined by the sign of $\langle Z(q(s,p)),X(q(s,p))\rangle$ for s in a neighborhood of

$0 \in \mathbb{R}^{n-1}$. Notationally, let

$$\nu = (\nu_2, \cdots, \nu_n) \quad \text{with } \nu_i \text{ a nonnegative integer,}$$

$$|\nu| = \sum_2^n \nu_i, \qquad \nu! = \nu_2! \cdots \nu_n!,$$

$$a(\nu) = \langle Z(p), (\operatorname{ad}^{\nu_n} Y^n, (\cdots (\operatorname{ad}^{\nu_2} Y^2, X) \cdots)(p)) \rangle,$$

$$\varphi_r(s) = \sum_{|\nu|=r} (1/\nu!)(-s_2)^{\nu_2} \cdots (-s_n)^{\nu_n} a(\nu).$$

Then, properly restated, we show in [10],

PROPOSITION 1.3 [10, Thm. 2]. *Assume $X(p) = 0$, $Y^2, \cdots, Y^n$ are involutive and linearly independent at p. A necessary and sufficient condition that* int. $\mathscr{A}(t, p, \mathscr{D}) \neq \varnothing \quad \forall t > 0$ *is that there exist an integer $r \geqq 1$ such that $\varphi_r(s) \not\equiv 0$. If this occurs, a necessary and sufficient condition that $p \in$* int. $\mathscr{A}(t, p, \mathscr{D}) \quad \forall t > 0$ *is that $\sum_{r=1}^{\infty} \varphi_r(s)$ change sign in every (sufficiently small) neighborhood of $0 \in \mathbb{R}^{n-1}$.*

In general, one cannot compute $\sum_{r=1}^{\infty} \varphi_r(s)$. However, if r^* is the smallest integer such that $\varphi_{r^*}(s) \not\equiv 0$ and the form $\varphi_{r^*}(s)$ is definite (changes sign in every neighborhood of $0 \in \mathbb{R}^{n-1}$), then $p \in \partial\mathscr{A}(t, p, \mathscr{D})$ for sufficiently small $t > 0$ ($p \in$ int. $\mathscr{A}(t, p, \mathscr{D}) \quad \forall t > 0$). Thus $\varphi_r(s)$ for $r > r^*$ need only be considered if $\varphi_{r^*}(s)$ is semi-definite.

The case $X(p) \neq 0$, $Y^2, \cdots, Y^n$ involutive and $X(p), Y^2(p), \cdots, Y^n(p)$ linearly independent is studied in [1]. Here there exists a unique one form Z such that $\langle Z(x), X(x) \rangle \equiv 1$, $\langle Z(x), Y^i(x) \rangle \equiv 0$ for $i = 2, \cdots, n$ and x in a neighborhood of p. For each integer $j \geqq 0$ we now let

$$a(\nu, j) = \langle Z(p), (\operatorname{ad}^j X, (\operatorname{ad}^{\nu_2} Y^2 (\cdots (\operatorname{ad}^{\nu_n} Y^n, X) \cdots)(p)) \rangle,$$

(4)

$$\varphi_{rj}(s) = \sum_{|\nu|=r} (1/\nu!)(-s_2)^{\nu_2} \cdots (-s_n)^{\nu_n} a(\nu, j).$$

A slight restatement of the main result in [1] is

PROPOSITION 1.4 [1, Thm. 1]. *Assume $Y^2, \cdots, Y^n$ are involutive and $X(p), Y^2(p), \cdots, Y^n(p)$ are linearly independent. A necessary and sufficient condition that $\mathscr{A}(t, p, \mathscr{D})$ have nonempty interior for all $t > 0$ is that some r-form $\varphi_{rj}(s) \not\equiv 0$. If r^* is the smallest integer such that $\varphi_{r^* j}(s) \not\equiv 0$ for some $j \geqq 0$ and j^* the smallest j for which this occurs, a sufficient condition that $T^X(t)p \in$* int. $\mathscr{A}(t, p, \mathscr{D})$ $\forall t > 0$ *is that $\varphi_{r^* j^*}(s)$ assumes both positive and negative values in every neighborhood of $0 \in \mathbb{R}^{n-1}$. If $j^* = 0$, a sufficient condition that $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{D})$ for small $t > 0$ is that $\varphi_{r^* j^*}(s)$ be definite in some neighborhood of $0 \in \mathbb{R}^{n-1}$. A more general sufficient condition that $T^X(t)p \in$* int. $\mathscr{A}(t, p, \mathscr{D})$ $\forall t > 0$ *is that for any $\varepsilon > 0$ $\exists$ a $\tau \in (0, \varepsilon)$ such that $\sum_{j=0}^{\infty} \sum_{r=1}^{\infty} ((-\tau)^j / j!) \varphi_{rj}(s)$ changes sign as a function of s in every neighborhood of $0 \in \mathbb{R}^{n-1}$.*

In the next section we obtain somewhat similar results to Proposition 1.4 with some of the restrictions on the vector fields $Y^2, \cdots, Y^m$ weakened.

2. The case when $L\{Y^2, \cdots, Y^m\}$ defines an involutive distribution of dim $k < n$. If $\dim L\{Y^2, \cdots, Y^m\}(p) = k \leqq n$, Nagano's theorem [8, Thm. 1] yields the existence of a k-dimensional integral manifold for $L\{Y^2, \cdots, Y^m\}$, through p. We denote this manifold $N(p)$. Furthermore, $\dim\{Y^2, \cdots, Y^m\} \cdot (q) = k$ for $q \in N(p)$. If we assume $K < n$ and X is transverse to $N(p)$, $\dim N(T^X(\tau)p)$ can change with τ, creating technical difficulties. To keep matters reasonable, we make the following assumptions throughout this section:

(a-1) $\dim L\{Y^2, \cdots, Y^m\}(q) = k < n$ for all q in a neighborhood of p. (Then $L\{Y^2, \cdots, Y^m\}$ defines an involutive distribution of dim k.)

(a-2) $X(p) \notin L\{Y^2, \cdots, Y^m\}(p)$.

For example, these assumptions are satisfied if $m = 2$, $n \geqq 2$ and $X(p)$, $Y^2(p)$ are linearly independent; or if $m \leqq n$, $X(p)$, $Y^2(p)$, $\cdots$, $Y^m(p)$ are linearly independent and $Y^2, \cdots, Y^m$ are involutive.

PROPOSITION 2.1. *Let $V^2, \cdots, V^{k+1} \in L\{Y^2, \cdots, Y^m\}$ be such that $X(p), V^2(p), \cdots, V^{k+1}(p)$ are linearly independent. There exist $V^{k+2}, \cdots, V^n \in V(M)$ such that $V^2, \cdots, V^n$ are involutive and $X(p), V^2(p), \cdots, V^n(p)$ are linearly independent.*

Proof. Choose any vector fields $W^{k+2}, \cdots, W^n \in V(M)$ such that $X(p), V^2(p), \cdots, V^{k+1}(p), W^{k+2}(p), \cdots, W^n(p)$ are linearly independent. Define

$$r(\sigma_2, \cdots, \sigma_n, \tau, p) = T^{V^2}(\sigma_2) \circ \cdots \circ T^{V^{k+1}}(\sigma_{k+1}) \circ T^{W^{k+2}}(\sigma_{k+2}) \circ \cdots \circ T^{W^n}(\sigma_n) \circ T^X(\tau)p.$$

Then the vector fields $\partial r/\partial\sigma_i$, $i = 2, \cdots, n$, commute. Since $V^2, \cdots, V^{k+1}$ were involutive, for $2 \leqq i \leqq k+1$ and $x = r(\sigma, \tau, p)$, $\partial r(x)/\partial\sigma_i = \sum_{j=2}^{k+1} c_{ij}(x) V^j(x)$, while $\partial r(p)/\partial\sigma_i = V^i(p)$. Thus the matrix $c_{ij}(x)$ is the identity at p hence smoothly invertible in a neighborhood of p; i.e., $V^i(x) = \sum_{j=2}^{k+1} e_{ij}(x)\partial r/\partial\sigma_j$, $i = 2, \cdots, k+1$, with e_{ij} smooth in a neighborhood of p. Define $V^j(x) = \partial r/\partial\sigma_j$ for $j = k+2, \cdots, n$. It follows that $V^2, \cdots, V^n$ are as required. □

With any choice of $V^2, \cdots, V^n$ as in Proposition 2.1 we associate the "extended" system

$$(\mathcal{V}) \qquad \dot{x} = X(x) + \sum_{i=2}^{n} u_i(t) V^i(x), \qquad x(0) = p,$$

with u_i measurable and $|u_i(t)| \leqq 1$ for $i = 2, \cdots, n$. Note that since $Y^2, \cdots, Y^n$ need not be in the collection $V^2, \cdots, V^{k+1}$ it is not necessarily true that the attainable set $\mathcal{A}(t, p, \mathcal{V})$ contains $\mathcal{A}(t, p, \mathcal{D})$ for each $t > 0$.

The following geometric motivation to the main result was conveyed to me by Professor H. J. Sussmann. For each value $t \geqq 0$, let L_t denote the $(n-1)$-dimensional integral manifold (leaf) of $V^2, \cdots, V^n$ through $T^X(t)p$. For each x in a neighborhood of p define $f(x)$ to be the value t such that $x \in L_t$. Since $T^X(t)p$ is transverse to L_t for small $t \geqq 0$, f is well-defined, analytic, constant on a fixed leaf L_t, and $f(T^X(t)p) \equiv t$. Let ψ be any solution of system $\mathcal{D}$ or system $\mathcal{V}$. Then

$$(5) \qquad (d/dt) f(\psi(t)) = (Xf)(\psi(t))$$

since $Y^i f \equiv 0$ and $V^j f \equiv 0$, $i = 2, \cdots, m$; $j = 2, \cdots, n$. Introduce new coordinates for a neighborhood of p via the map

$$(6) \qquad (s_2, \cdots, s_n, \tau) \to q(s, \tau, p) = T^{V^n}(s_n) \circ \cdots \circ T^{V^2}(s_2) \circ T^X(\tau) p.$$

PROPOSITION 2.2. *A sufficient condition that* $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{D})$ *and* $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{V})$ *for small* $t > 0$ *is that there exist an* $\varepsilon > 0$ *such that* $(Xf)(q(s, \tau, p)) \geqq 1$ (or $\leqq 1$) *for* $0 \leqq \tau < \varepsilon$ *and* s *in some neighborhood of* $0 \in \mathbb{R}^{n-1}$.

Proof. Suppose $(Xf)(q(s, \tau, p)) \geqq 1$ for $0 \leqq \tau < \varepsilon$ and s in a neighborhood of $0 \in \mathbb{R}^{n-1}$. Then for sufficiently small $t_1 > 0$, $(Xf)(\psi(t)) \geqq 1$ for $0 \leqq t \leqq t_1$ and (5) yields

$$t_1 \leqq \int_0^{t_1} (Xf)(\psi(t))\, dt = f(\psi(t_1)).$$

Thus $\mathscr{A}(t, p, \mathscr{D})$ and $\mathscr{A}(t, p, \mathscr{V})$ lie on "one side" of the leaf L_{t_1} and $T^X(t_1)p$, which is on this leaf, belongs to the boundaries of $\mathscr{A}(t_1, p, \mathscr{D})$ and $\mathscr{A}(t_1, p, \mathscr{V})$. The case $Xf \leqq 1$ is similar. □

Our main theorem, which we next state, is obtained from Proposition 2.2 by deriving a Taylor series expansion of $(Xf)(q(s, \tau, p))$ to determine the sign of $Xf - 1$. Assume a Riemannian metric on M and let $\langle V, W \rangle$ denote the inner product of two tangent vectors as determined by this metric. Let $W(p)$ be the unique tangent vector such that

$$(7) \qquad \langle W(p), X(p) \rangle = 1, \qquad \langle W(p), V^i(p) \rangle = 0, \qquad i = 2, \cdots, n.$$

Actually, $W(p) = \operatorname{grad} f(p)$. Again, let:

$$\nu = (\nu_2, \cdots, \nu_n) \quad \text{with } \nu_i \text{ a nonnegative integer},$$

$$|\nu| = \sum_2^n \nu_i, \qquad \nu! = \nu_2! \cdots \nu_n!,$$

$$a(\nu, j) = \langle W(p), (\operatorname{ad}^j X, (\operatorname{ad}^{\nu_2} V^2, (\cdots (\operatorname{ad}^{\nu_n} V^n, X) \cdots)(p) \rangle,$$

$$\varphi_{rj}(s) = \sum_{|\nu| = r} (1/\nu!)(-s_2)^{\nu_2} \cdots (-s_n)^{\nu_n} a(\nu, j).$$

THEOREM 2.1. *Let* $\mathscr{D}$ *be as in* (1); *let assumptions* (a-1), (a-2) *hold, and* $V^2, \cdots, V^n$ *be chosen as in Proposition* 2.1. *If* $\varphi_{rj}(s) \equiv 0$ *for all* $r \geqq 1$, $j \geqq 0$, *then* $\mathscr{A}(t, p, \mathscr{D})$ *and* $\mathscr{A}(t, p, \mathscr{V})$ *are contained in the* $(n-1)$*-dimensional leaf* L_t *for each* $t > 0$. *Assume* $\varphi_{rj}(s) \not\equiv 0$ *for some* $r \geqq 1$, $j \geqq 0$; *let* r^* *be the smallest integer* r *such that* $\varphi_{r^* j}(s) \not\equiv 0$ *for some* j *and* j^* *the smallest value of* j *for which this occurs. If* $j^* = 0$ *or if* $\varphi_{rj}(s) = 0$ *when* $r > r^*$, $0 \leqq j < j^*$, *a sufficient condition that* $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{D})$ *and* $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{V})$ *for small* $t > 0$ *is that* $\varphi_{r^* j^*}(s)$ *be definite* (*i.e.*, $\varphi_{r^* j^*}(s) > 0$ *or* < 0 *for* s *in a deleted neighborhood of* $0 \in \mathbb{R}^{n-1}$).

Proof. We proceed to the expansion of $(Xf)(q(s, \tau, p))$ which will be used with Proposition 2.2 for the result.

For fixed small $|s|$, $|\tau|$, $q(s, \tau, \cdot)$ maps a neighborhood of p in M onto M diffeomorphically. The induced tangent space isomorphism will be denoted $q_*(s, \tau, p)$. Letting $DT^X(\tau)$ denote the differential of the map $p \to T^X(\tau)p$ and

using the identity $DT^X(\tau)X(p) = X(T^X(\tau)p)$ one easily shows

$$(\partial/\partial\tau)q(s, \tau, p) = q_*(s, \tau, p)X(p). \tag{8}$$

Introduce the map

$$k(s)(p) = T^{V^n}(s_n) \circ \cdots \circ T^{V^2}(s_2)p.$$

Then $q(s, \tau, p) = k(s) \circ T^X(\tau)p$. Also, since $V^2, \cdots, V^n$ are involutive, for fixed s, τ, $k(s)(\cdot): L_\tau \to L_\tau$; hence the induced tangent space isomorphism, denoted $k_*(s)$, maps tangent vectors of L_τ into tangent vectors of L_τ.

LEMMA 2.1. $(Xf)(q(s, \tau, p)) = \langle (\operatorname{grad} f)(T^X(\tau)p), k_*^{-1}(s)X(k(s) \circ T^X(\tau)p)\rangle$.

Proof. Since $f \equiv \tau$ on the leaf L_τ we have $f(q(s, \tau, p)) \equiv \tau$ so

$$\begin{aligned} 1 = (\partial/\partial\tau)f(q(s, \tau, p)) &= \langle (\operatorname{grad} f)(q(s, \tau, p)), \partial q/\partial\tau\rangle \\ &= \langle (\operatorname{grad} f)(q(s, \tau, p)), q_*(s, \tau, p)X(p)\rangle. \end{aligned} \tag{9}$$

Now suppose $(Xf)(q(s, \tau, p)) \equiv \langle (\operatorname{grad} f)(q(s, \tau, p)), X(q(s, \tau, p))\rangle = \gamma$. From (9) we see

$$\begin{aligned} X(q(s, \tau, p)) &= \gamma q_*(s, \tau, p)X(p) + \sum_{i=2}^{n} c_i V^i(q(s, \tau, p)) \\ &= \gamma k_*(s)X(T^X(\tau)p) + \sum_{i=2}^{n} c_i V^i(q(s, \tau, p)). \end{aligned}$$

Applying $k_*^{-1}(s)$ to both sides and using the fact that this map takes tangent vectors of L_τ into tangent vectors of L_τ while $V^2(T^X(\tau)p), \cdots, V^n(T^X(\tau)p)$ span the tangent space of L_τ at $T^X(\tau)p$, we obtain

$$k_*^{-1}(s)X(k(s) \circ T^X(\tau)p) = \gamma X(T^X(\tau)p) + \sum_{i=2}^{n} c_i' V^i(T^X(\tau)p).$$

Then $\langle (\operatorname{grad} f)(T^X(\tau)p), k_*^{-1}(s)X(k(s) \circ T^X(\tau)p)\rangle = \gamma$ since $\langle (\operatorname{grad} f)(T^X(\tau)p), V^i(T^X(\tau)p)\rangle = 0$, $i = 2, \cdots, n$. The argument is reversible. □

LEMMA 2.2. $(Xf)(q(s, \tau, p)) = 1 + \sum_{|\nu|=1}^{\infty} (1/\nu!)(-s_2)^{\nu_2} \cdots (-s_n)^{\nu_n} \langle (\operatorname{grad} f) \cdot (T^X(\tau)p), (\operatorname{ad}^{\nu_2} V^2, (\cdots (\operatorname{ad}^{\nu_n} V^n, X) \cdots)(T^X(\tau)p)\rangle$.

Proof. Use Lemma 2.1 and the identity $k_*^{-1}(s)X(k(s) \circ T^X(\tau)p) = \sum_{\nu=0}^{\infty} (1/\nu!)(-s_2)^{\nu_2} \cdots (-s_n)^{\nu_n} (\operatorname{ad}^{\nu_2} V^2, (\cdots (\operatorname{ad}^{\nu_n} V^n, X) \cdots)(T^X(\tau)p)$ which follows by repeated use of the Campbell–Hausdorff formula. Apply grad f; note that for $\nu = 0$, $\langle (\operatorname{grad} f)(T^X(\tau)p), X(T^X(\tau)p)\rangle = 1$, and the result follows. □

Our goal is to obtain an expansion about p rather than about $T^X(\tau)p$. Let $\tau > 0$ and V be any vector field. It is not in general true that $\langle (\operatorname{grad} f)(T^X(\tau)p), V(T^X(\tau)p)\rangle = \langle (\operatorname{grad} f)(p), DT^X(-\tau)V(T^X(\tau)p)\rangle$. We shall show that this does hold precisely in the case of interest to us. Let

$$\mathscr{S}^1(\mathscr{V}) = \{(\operatorname{ad}^j X, V^i): i = 2, \cdots, n; j \geqq 0\}.$$

The standard first order test, i.e., rank $\mathscr{S}^1(\mathscr{V})(p) = n$, implies $T^X(t)p \in \operatorname{int}. \mathscr{A}(t, p, \mathscr{V})$ $\forall t > 0$; hence we shall be interested in the case rank $\mathscr{S}^1(\mathscr{V})(p) < n$. This first order condition, together with the definition of $\varphi_{1i}(s)$, yield

LEMMA 2.3. rank $\mathscr{S}^1(\mathscr{V})(p) = n \Leftrightarrow \varphi_{1j}(s) \not\equiv 0$ *for some* $j \geqq 0 \Rightarrow T^X(t)p \in \operatorname{int}. \mathscr{A}(t, p, \mathscr{V})$ $\forall t > 0$.

Remark 2.1. For odd r, $\varphi_{rj}(s) = -\varphi_{rj}(-s)$ for all j. Thus $\varphi_{1j}(s) \not\equiv 0$ implies $\varphi_{1j}(s)$ is not definite, and the case $r^* = 1$ in Theorem 2.1 is taken care of. Indeed, if $\varphi_{r^*j^*}(s)$ is to be definite, we must have r^* even.

LEMMA 2.4. $\varphi_{1j}(s) \equiv 0$ *for all* $j \geqq 0$ *implies*

$$\langle (\operatorname{grad} f)(T^X(\tau)p),\ V(T^X(\tau)p)\rangle = \langle (\operatorname{grad} f)(p),\ DT^X(-\tau)V(T^X(\tau)p)\rangle$$

for any $V \in V(M)$ *and* $t \geqq 0$.

Proof.

$$\varphi_{1j}(s) \equiv 0\ \forall j \Rightarrow \langle W(p),\ (\operatorname{ad}^j X,\ V^i)(p)\rangle = 0$$

$$\Rightarrow \left\langle W(p),\ \sum_{j=0}^{\infty} \left(\frac{(-\tau)^j}{j!}\right)(\operatorname{ad}^j X,\ V^i(p)\right\rangle = 0$$

$$\Rightarrow \langle W(p),\ DT^X(-\tau)V^i(T^X(\tau)p)\rangle = 0, \qquad i = 2, \cdots, n,\ \tau \geqq 0.$$

Now $X(T^X(\tau)p)$, $V^2(T^X(\tau)p)$, $\cdots$, $V^n(T^X(\tau)p)$ span the tangent space to M at $T^X(\tau)p$ for small $\tau \geqq 0$; hence we may write

$$V(T^X(\tau)p) = \gamma X(T^X(\tau)p) + \sum_{i=2}^{n} c_i V^i(T^X(\tau)p).$$

Then $\langle (\operatorname{grad} f)(T^X(\tau)p),\ V(T^X(\tau)p)\rangle = \gamma = \langle W(p),\ \gamma X(p)\rangle = \langle (\operatorname{grad} f)(p),\ \gamma DT^X \cdot (-\tau)X(T^X(\tau)p)\rangle = \langle (\operatorname{grad} f)(p),\ DT^X(-\tau)V(T^X(\tau)p)\rangle$. □

LEMMA 2.5. $\varphi_{1j}(s) \equiv 0$ *for all* $j \geqq 0$ *implies*

$$(Xf)(q(s, \tau, p)) = 1 + \sum_{r=2}^{\infty} \sum_{j=0}^{\infty} ((-\tau)^j/j!)\varphi_{rj}(s). \tag{10}$$

Proof. Apply Lemma 2.4 to each term in the expansion of Xf given in Lemma 2.2. □

To complete the proof of Theorem 2.1, we first note that if $\varphi_{rj}(s) \equiv 0$ for all $r \geqq 1$, $j \geqq 0$, then $(Xf)(q(s, \tau, p)) \equiv 1$ and if ψ is any comparison solution of $\mathscr{V}$ (or of $\mathscr{D}$), it follows from (5) that $f(\psi(t_1)) = t_1$, i.e., $\psi(t_1) \in L_{t_1}$, for any $t_1 > 0$. On the other hand, if r^*, j^* are chosen as stated and $\varphi_{r^*j^*}(s)$ is definite, say $\varphi_{r^*j^*}(s) > 0$ if $s \neq 0$, with $j^* = 0$, then from (10) we see $(Xf)(q(s, \tau, p)) \geqq 1$ for τ in some interval $[0, \varepsilon)$, $\varepsilon > 0$ and s in some neighborhood of $0 \in \mathbb{R}^{n-1}$. This also holds if $\varphi_{rj}(s) = 0$ when $r > r^*$ and $0 \leqq j < j^*$. Proposition 2.2 now yields the desired result. □

Remark 2.2. For the relation between $\varphi_{1j}(s) \equiv 0 \quad \forall j \geqq 0$ and $T^X(\,\cdot\,)p$ being a singular arc, see [1, Props. 2.6 and 2.7].

Let $\mathscr{D}$ be as in (1); assumptions (a-1), (a-2) hold; $V^2, \cdots, V^n$ chosen as in Proposition 2.1, and $\varphi_{r^*j^*}(s)$, defined relative to this choice, as in Theorem 2.1. Assume $\varphi_{r^*j^*}(s)$ is definite so $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{D})$ for $0 \leqq t \leqq \varepsilon$, some $\varepsilon > 0$. If $\tilde{\mathscr{D}}$ denotes a system $\dot{x} = X(x) + \sum_{i=2}^{l} u_i \tilde{Y}^i(x)$, with $|u_i(t)| \leqq 1$ and $L\{\tilde{Y}^2, \cdots, \tilde{Y}^l\}(q) = L\{Y^2, \cdots, Y^m\}(q)$ for q in a neighborhood of p, then $T^X(t)p \in \partial\mathscr{A}(t, p, \tilde{\mathscr{D}})$ for $0 \leqq t \leqq \varepsilon$. This follows since $V^2, \cdots, V^n$ is a choice for $\tilde{\mathscr{D}}$ which satisfies Proposition 2.1 and $\varphi_{rj}(s)$ depends only on $V^2, \cdots, V^n$. In particular, if $\tilde{Y}^i = \alpha_i Y^i$, $\alpha_i > 0$; then $\tilde{\mathscr{D}}$ with $|u_i(t)| \leqq 1$ is equivalent to $\mathscr{D}$ with $|u_i(t)| \leqq \alpha_i$. Clearly $L\{Y^2, \cdots, Y^m\}(q) = L\{\alpha_2 Y^2, \cdots, \alpha_m Y^m\}(q)$ for all q; hence we have

COROLLARY 2.1. *The sufficient condition in Theorem* 2.1 *is independent of the value* $\alpha_i > 0$ *in the bounds* $|u_i(t)| \leqq \alpha_i$ *on control components in* $\mathscr{D}$.

Example 2.1. (For notational ease, we continue to write all vectors as row vectors.) Let $X(x) = (1, 0, x_3, 0)$, $Y^2(x) = (x_3, 0, 1, 1)$, $Y^3(x) = (0, x_3, 0, 0)$ be vector fields on $M^4 = \mathbb{R}^4$. We consider the control system $\mathscr{D}: X(x) + \sum_{i=2}^{3} u_i(t) Y^i(x)$, $x(0) = p = 0$, $|u_i(t)| \leqq 1$. Computation shows $[Y^2, Y^3](x) = (0, -1, 0, 0)$; higher order products of the Y^i vanish and $\dim L\{Y^2, Y^3\}(x) = 2$ for all x in a neighborhood of p. Here $Y^2(p)$, $Y^3(p)$ are linearly dependent while $Y^2(p)$, $[Y^2, Y^3](p)$ do span $L\{Y^2, Y^3\}(p)$, and $X(p) \notin L\{Y^2, Y^3\}(p)$. Thus (a-1) and (a-2) hold. We choose $V^2 = Y^2$, $V^3 = [Y^2, Y^3]$ and $V^4 = (0, 0, 0, 1)$ so $X(p)$, $V^2(p)$, $V^3(p)$, $V^4(p)$ are linearly independent while $[V^4, V^2] = 0$, $[V^4, V^3] = 0$ and V^2, V^3, V^4 are involutive. By inspection, $W(p) = (1, 0, 0, 0)$.

Computation shows $(\mathrm{ad}^j X, V^2)(x) = ((-1)^j x_3, 0, 1, 0)$; $(\mathrm{ad}^j X, V^3)(x) = 0$; $(\mathrm{ad}^j X, V^4)(x) = 0$ for $j \geqq 1$; hence $a(1, 0, 0, j) = a(0, 1, 0, j) = a(0, 0, 1, j) = 0$ for all $j \geqq 0$ or $\varphi_{1j}(s) \equiv 0 \quad \forall j \geqq 0$. The linear test fails to show whether or not $T^X(t)p = (t, 0, 0, 0)$ belongs to the interior of $\mathscr{A}(t, p, \mathscr{D})$ or $\mathscr{A}(t, p, \mathscr{V})$ for $t > 0$.

Next, for $r = 2$, $(\mathrm{ad}^2 V^4, X)(x) = 0$; $(\mathrm{ad}\, V^3, (\mathrm{ad}\, V^4, X))(x) = 0$; $(\mathrm{ad}\, V^2, (\mathrm{ad}\, V^4, X))(x) = 0$, $(\mathrm{ad}^2 V^3, X)(x) = 0$; $(\mathrm{ad}\, V^2, (\mathrm{ad}\, V^3, X))(x) = 0$; $(\mathrm{ad}^2 V^2, X)(x) = (-2, 0, 0, 0)$. This shows $a(0, 0, 2, 0) = 0$, $a(0, 1, 1, 0) = 0$, $a(1, 0, 1, 0) = 0$, $a(0, 2, 0, 0) = 0$, $a(1, 1, 0, 0) = 0$, $a(2, 0, 0, 0) = -2$ and $r^* = 2$, $j^* = 0$ while $\varphi_{r^* j^*}(s) = -s_2^2$ which is only semi-definite; hence Theorem 2.1, as stated, does not apply. On the other hand, we may compute that $\varphi_{2j}(s) \equiv 0$ if $j \geqq 1$ while $\varphi_{rj}(s) \equiv 0$ if $r \geqq 3$, $j \geqq 0$. Thus from (10), $(Xf)(q(s, \tau, p)) = 1 - s_2^2 \leqq 1$ and Proposition 2.2 shows $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{D})$ and $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{V})$ for sufficiently small $t > 0$.

3. The case $L(\mathscr{P}^1)$ nilpotent. We shall be very brief; the ideas involved go back to Sophus Lie and an exposition can be found in [9]. We consider the auxiliary system (3), with u_i measurable, $|u_i(t)| \leqq 1$ and again denoting the attainable set of this system as $\mathscr{B}(t, p)$, recall that $\mathscr{A}(t, p, \mathscr{D}) = T^X(t)\mathscr{B}(t, p)$. Thus $T^X(t)p \in \mathrm{int}.\, \mathscr{A}(t, p, \mathscr{D})$ if and only if $p \in \mathrm{int}.\, \mathscr{B}(t, p)$.

The assumption $L(\mathscr{P}^1)$ nilpotent, and say of dimension k, implies this algebra admits a base $W^1, \cdots, W^k$ such that if $\mathscr{G}_i = L\{W^{i+1}, \cdots, W^k\}$, then we have the ideal decomposition

$$L(\mathscr{P}^1) = \mathscr{G}_0 \supset \mathscr{G}_1 \supset \cdots \supset \mathscr{G}_{k-1} \supset \mathscr{G}_k = \{0\}, \tag{11}$$

$$\text{where } [\mathscr{G}_i, \mathscr{G}_j] \subset \mathscr{G}_{j+1}, \quad 0 \leqq i \leqq j \leqq k.$$

Furthermore $\mathrm{ad}\, W^i: \mathscr{G}_i \to \mathscr{G}_i$ has a nilpotent matrix, M_i, as its representation relative to this basis.

Returning to (3), each term $(\mathrm{ad}^r X, Y^i)$ is a linear combination (with real coefficients) of the $W^1, \cdots, W^k$. Using this we can rewrite (3) as

$$\dot{y}(t) = f_{11}(t; u) W^1(y) + \cdots + f_{1k}(t; u) W^k(y), \qquad y(0) = p. \tag{12}$$

Now let $F_1(t; u) = \int_0^t f_{11}(\tau; u)\, d\tau$ where u is admissible and fixed; let $\exp(F_1(t; u) W^1)(p)$ denote the solution of $\dot{x} = f_{11}(t; u) W^1(x)$, $x(0) = p$ and we find (as in obtaining (3) from (1)) that a solution of (12) admits the representation $y(t; u) = \exp(F_1(t; u) W^1) \circ v(t; u)$ where v satisfies $\dot{v}(t) =$

$\exp(-F_1(t;u)\operatorname{ad} W^1)\{f_{12}(t,u)W^2(v)+\cdots+f_{1k}(t;u)W^k(v)\}$ with $v(0)=p$. Let M_1 be the matrix representation of $\operatorname{ad} W^1: \mathcal{G}_1 \to \mathcal{G}_1$. Then, in components, we let

$$\begin{pmatrix} f_{22}(t;u) \\ f_{23}(t;u) \\ \cdot \\ \cdot \\ \cdot \\ f_{2k}(t;u) \end{pmatrix} = e^{-F_1(t,u)M_1} \begin{pmatrix} f_{12}(t;u) \\ \cdot \\ \cdot \\ \cdot \\ f_{1k}(t;u) \end{pmatrix};$$

hence v satisfies $\dot{v}(t)=f_{22}(t;u)W^2(v)+\cdots+f_{2k}(t;u)W^k(v)$, $v(0)=p$. Continuing, inductively, with $F_2(t;u)=\int_0^t f_{22}(\tau,u)\,dt$, etc., the solution of (12) is responsible as

$$y(t;u)=(\exp F_1(t;u)W^1)\circ\cdots\circ(\exp F_k(t;u)W^k)(p). \tag{13}$$

As we show next, the $F_i(t;u)$ are computable and the representation (13) reduces the original problem to a new problem, which may be solvable. A similar representation to (13) can be obtained if $L(\mathcal{S}^1)$ is solvable; however, computation of the exponentials is more difficult since the M_i need no longer be nilpotent.

Example 3.1. Let $M=\mathbb{R}^3$, $X(x)=(1-x_2,0,0)$, $Y^2(x)=Y(x)=(0,1,x_1)$, $p=0$. Then $[X,Y](x)=(-1,0,x_2-1)$, $(\operatorname{ad}^j X,Y)(x)=0$ if $j\geqq 2$, $[Y,[X,Y]](x)=(0,0,-2)$, while higher order products vanish. We see rank $\mathcal{S}^1(p)=2$; hence the linear test fails; dim $L(\mathcal{S}^1)(p)=3$ showing int. $\mathcal{A}(t,p,\mathcal{D})\neq\varnothing$ $\forall t>0$. If $W^1=Y$, $W^2=[X,Y]$ and $W^3=[Y,[X,Y]]$, then $\{W^1,W^2,W^3\}$ is an ordered basis for $L(\mathcal{S}^1)$ such that the ideal decomposition holds. Relative to this basis

$$\operatorname{ad} W^1: \mathcal{G}_1 \to \mathcal{G}_1 \quad \text{has matrix representation} \quad M_1=\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

$$\operatorname{ad} W^2: \mathcal{G}_2 \to \mathcal{G}_2 \quad \text{has matrix representation} \quad M_2=(0).$$

The equation (3), for example, is $\dot{y}(t)=u(t)Y(y)-tu(t)[X,Y](y)$, $y(0)=p$ or, with the right side in terms of the W^i,

$$\dot{y}(t)=f_{11}(t;u)W^1(y)+f_{12}(t;u)W^2(y)+f_{13}(t,u)W^3(y), \qquad y(0)=p, \tag{14}$$

where $f_{11}(t;u)=u(t)$, $f_{12}(t;u)=-tu(t)$, $f_{13}(t;u)=0$. Let $F_1(t;u)=\int_0^t f_{11}(\tau;u)\,dt=\int_0^t u(\tau)\,dt$; attempt a solution of (14) of the form $y(t;u)=\exp(F_1(t,u)W^1)\circ v(t;u)$; compute

$$e^{-F_1(t;u)M_1}\begin{pmatrix} f_{12}(t;u) \\ f_{13}(t;u) \end{pmatrix} = \begin{pmatrix} -tu(t) \\ tu(t)\displaystyle\int_0^t u(\tau)\,d\tau \end{pmatrix} = \begin{pmatrix} f_{21}(t;u) \\ f_{22}(t;u) \end{pmatrix}$$

so v satisfies $\dot{v}(t)=-tu(t)W^2(v)+(tu(t)\int_0^t u(\tau)\,d\tau)W^3(v)$, $v(0)=p$. Continuing, we find that $F_2(t;u)=-\int_0^t \tau u(\tau)\,d\tau$; $F_3(t;u)=\int_0^t \tau u(\tau)(\int_0^\tau u(\sigma)\,d\sigma)\,d\tau$, while the solution of (14) can be written as

$$y(t;u)=\exp(F_1(t;u)W^1)\circ\exp(F_2(t;u)W^2)\circ\exp(F_3(t;u)W^3)(p). \tag{15}$$

Now $W^1(p)$, $W^2(p)$ and $W^3(p)$ are linearly independent. Thus to check whether, for any $t_1>0$, the points $y(t_1;u)$ cover a neighborhood of p as u varies,

equation (15) shows we should consider the map $u \to F(t_1, u) = (F_1(t_1, u), F_2(t_1, u), F_3(t_1, u)) \in \mathbb{R}^3$. If this map covers a neighborhood of $0 \in \mathbb{R}^3$, $\{y(t_1; u) : u \text{ admissible}\}$ covers a neighborhood of p. The "reduced controllability problem" becomes

$$\dot{F}_1(t) = u(t), \quad \dot{F}_2(t) = -tu(t), \quad \dot{F}_3(t) = tF_1(t)u(t) \tag{16}$$

with $F_1(0) = F_2(0) = F_3(0) = 0$. (In general, for nilpotent $L(\mathscr{S}^1)$ of dimension k, we will have a nonautonomous system of k equations, such as (16), with $k \neq n$.) For the particular system, above, we may calculate that for any $t_1 > 0$, $F_3(t_1) = \int_0^{t_1} \tau F_1(\tau)\dot{F}_1(\tau)\, d\tau = t_1 F_1^2(t_1)/2 - \int_0^{t_1} (F_1^2(\tau)/2)\, d\tau$. It is clear, therefore, that we cannot have $F(t_1; u) = (0, \alpha_2, \alpha_3)$ with $\alpha_3 = F_3(t_1; u) > 0$; hence the map $u \to F(t_1; u)$ does not cover a neighborhood of $0 \in \mathbb{R}^3$.

Example 3.2. Let the system be that of Example 3.1; we shall show $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{D})$ for small $t > 0$ by the methods of § 2.

dim $L\{Y^2\}(x) = 1$ for all x in a neighborhood of p. Choose $V^2 = Y^2$ and $V^3(x) = (-1, 0, -1 - x_2)$. Then $X(p)$, $V^2(p)$, $V^3(p)$ are linearly independent; assumptions (a-1), (a-2) are satisfied and our choice of V^3 satisfies Proposition 2.1. Here $W(p) = (0, 1, -1)$. Computing $(\operatorname{ad} X, V^2)(p) = (-1, 0, -1)$, $(\operatorname{ad}^j X, V^2)(p) = 0$ if $j \geqq 2$, $(\operatorname{ad}^j X, V^3) = 0$ if $j \geqq 1$; hence rank $\{(\operatorname{ad}^j X, V^i)(p) : j \geqq 0, i = 1, 2\} = 2$ and $a(1, 0, j) = a(0, 1, j) = 0 \ \forall j$ so $\varphi_{1j}(s) = 0 \ \forall j$ and the "linear test" fails. Next, $a(2, 0, 0) = -2$, $a(1, 1, 0) = 0$, $a(0, 2, 0) = 0$ so $r^* = 2$, $j^* = 0$, $\varphi_{r^*j^*}(s) = -s_2^2/2$ which is only semi-definite. Further computation, however, shows $\varphi_{2j}(s) = 0$ if $j \geqq 1$ and $\varphi_{rj}(s) = 0$ if $r \geqq 3$. Thus $(Xf)(q(s, \tau, p)) = 1 - s_2^2/2$ and Proposition 2.2 shows $T^X(t)p \in \partial\mathscr{A}(t, p, \mathscr{D})$ for small $t > 0$.

Further results related to the decomposition technique of this section can be found in papers of Brocket [11] and Krener [12].

4. Systems of the form (1) as semi-linearizations of nonlinear systems. Many physical systems can be modeled by mathematical systems of the form (1). Examples in which the reference solution is "singular" (i.e., rank $\mathscr{S}^1(T^X(\tau)p) < n$) include the problem of maximum height for a variable thrust rocket through the atmosphere; maximum range of a reentry vehicle with lift capability; minimum heating reentry trajectories and minimum fuel transfer between coplanar, elliptic, orbits (the Lawden spiral). Our goal, here, will be to show the use of system (1) as a semi-linearization of a nonlinear system.

Let $U \subset \mathbb{R}^m$ have nonempty interior relative to $\mathbb{R}^m$; $0 \in \operatorname{int} U$, and for each $v \in U$ assume $f(\cdot, v)$ is an analytic vector field on the analytic n-manifold M. We consider the control system

$$\dot{x}(t) = f(x(t), u(t)), \qquad x(0) = p, \tag{17}$$

where u is an admissible control implies u is Lebesgue measurable and $u(t) \in U \ \forall t$. Let $u(t) \equiv 0$ generate the reference solution of (17), denoted $\varphi(\cdot)$; f_x, f_u denote matrices of partial derivatives with respect to local coordinates, and $\mathscr{F}(t, p)$ denote the set of points attainable at time t by solutions initiating from p at time 0.

It is classically known that if the linear (variational) equation

$$\dot{y} = f_x(\varphi(t), 0)y + f_u(\varphi(t), 0)v(t), \qquad y(0) = 0, \tag{18}$$

is controllable on $[0, t_1]$ for all $t_1 > 0$ (see [13, § 19]), then $\varphi(t) \in \text{int. } \mathscr{F}(t, p) \quad \forall t > 0$. In addition to (18), we introduce the "semi-linearized" system

$$\dot{x} = f(x, 0) + f_u(x, 0)v(t), \qquad x(0) = p, \tag{19}$$

where we require the values of v to be in an origin centered disc, $D(r)$, of radius $r > 0$ in $\mathbb{R}^m$. Geometrically, (19) "linearizes" the contingent cone $F(x) = \{f(x, \sigma) \in TM_x : \sigma \in U\}$, where TM_x denotes the tangent space to M at x, by replacing it by $G(x, r) = \{f(x, 0) + f_u(x, 0)\sigma : \sigma \in D(r)\}$. System (19) has the form of system (1); we denote its attainable set at time t by $\mathscr{A}(t, p, r)$. The reference solution, φ, of (17) is also a solution of (19) obtained for $v \equiv 0$. Our goal is to be able to obtain information about the attainable set $\mathscr{F}(t, p)$ by studying $\mathscr{A}(t, p, r)$ via the theory developed in the previous sections.

THEOREM 4.1. *Let $t_1 > 0$ be given and $m = n$, so we have $n - 1$ components of control. Assume the vectors $f(p, 0), f_{u_2}(p, 0), \cdots, f_{u_n}(p, 0)$ are linearly independent. Then given any $\varepsilon > 0$ there exists $r > 0$ such that $\mathscr{A}(t_1, p, r) \subset \bigcup_{|\gamma| < \varepsilon} \mathscr{F}(t_1 + \gamma, p)$.*

Proof. For notational ease, let $f_{u_i}(x, 0) = Y^i(x)$, $i = 2, \cdots, n$. The map $\sigma \to f(x, \sigma)$ defines (locally) a smooth $(n-1)$-manifold with tangent space at $f(x, 0)$ spanned by $Y^2(x), \cdots, Y^n(x)$. For x in a sufficiently small neighborhood of p, $f(x, 0)$, $Y^2(x), \cdots, Y^n(x)$ are linearly independent; hence for small $r > 0$ and $|v| \leqq r$, the line $\{\alpha[f(x, 0) + \sum_{i=2}^n v_i Y^i(x)] : \alpha \in \mathbb{R}^1\}$ is transverse to $F(x)$. By the implicit function theorem, there exists a smooth scalar valued function, $\alpha(x, v)$, such that $\alpha(x, v)[f(x, 0) + \sum_{i=2}^n v_i Y^i(x)] \in F(x)$; $\alpha(x, 0) = 1$, while if given $\delta > 0$ and a fixed (compact) neighborhood of p, we may choose $r > 0$ such that $|\alpha(x, v) - 1| < \delta$ if $|v| \leqq r$ and x is in this neighborhood of p.

Now let ψ be a solution of (19) corresponding to admissible v with $|v(t)| \leqq r$. Then $\psi(t_1) \in \mathscr{A}(t_1, p, r)$. Rescale time by letting $\tau = \tau_v(t) = [\int_0^t \alpha(\psi(\sigma), v(\sigma))\, d\sigma]^{-1}$, so $\tau_v'(t)$ is near one implying the inverse function $t(\tau)$ is well-defined. Let $\varphi(\tau) = \psi(t(\tau))$. Then $\varphi'(\tau) = \psi'(t(\tau))t'(\tau) = [f(\psi(t(\tau)), 0) + \sum_{i=2}^n v_i(t(\tau)) Y^i(t(\tau))]\alpha(\psi(t(\tau), v(t(\tau))) \in F(\varphi(\tau))$ showing φ is a solution of (17), i.e., $\varphi(\tau_v(t_1)) \in \mathscr{F}(\tau_v(t_1), p)$. Now given any $\varepsilon > 0$, by choosing $r > 0$ sufficiently small we can assure α remains near enough to one for all admissible v with $|v(t)| \leqq r$ so that $|t_1 - \tau_v(t_1)| < \varepsilon$. This means $\psi(t_1) = \varphi(\tau_v(t_1)) \in \bigcup_{|\gamma| < \varepsilon} \mathscr{F}(t_1 + \gamma, p)$. □

For applications, the main use of Theorem 4.1 would seemingly be as follows. Suppose $\varphi(\cdot, q)$, $\varphi(0, q) = q$, is the (desired reference) solution, corresponding to control $u \equiv 0$, of the n-dimensional nonlinear system

$$\dot{x} = -f(x, u), \tag{20}$$

where $u(t) \in U \subset \mathbb{R}^{n-1}$ and $0 \in \text{int. } U$. Let $\Gamma = \{\varphi(t, q) : t \geqq 0\}$ denote the orbit of φ and $\mathscr{F}(t, p)$ be the attainable set, at time t, of the system (20) with time reversed, i.e.,

$$\dot{x} = f(x, u), \qquad x(0) = p. \tag{21}$$

At time $t_0 > 0$ we desire the system to be in the state $p = \varphi(t_0, q)$. For some small $\varepsilon > 0$ we may measure the state at time $t_0 - \varepsilon$ obtaining a value $\tilde{p}$ which, due to perturbations or inaccuracies in the model, is near but not equal to $\varphi(-\varepsilon, p)$; i.e., $\varphi(\varepsilon, \tilde{p}) \neq p$. If $\tilde{p} \in \mathscr{F}(\varepsilon, p)$ we would expect to be able to "correct the error" in the

sense that there would be an admissible control u such that the corresponding solution $\psi^u(\cdot, \tilde{p})$ of (21) satisfies $\psi^u(0, \tilde{p}) = \tilde{p}$ and $\psi^u(\varepsilon, \tilde{p}) = p$. Thus $\varphi(-\varepsilon, p) \in$ int. $\mathscr{F}(\varepsilon, p)$ implies arbitrary small errors could be corrected in this manner.

Let $\mathscr{A}(t, p, r)$ be the attainable set at time t, of the semi-linearized system (19) associated with (21). Theorem 4.1 shows if we can establish that $\varphi(-\varepsilon, p) \in$ int. $\mathscr{A}(\varepsilon, p, r)$, then if $\tilde{p}$ is sufficiently near $\varphi(-\varepsilon, p)$, there will be an admissible control v such that the corresponding solution $\psi^v(\cdot, \tilde{p})$ of (20) satisfies $\psi^v(0, \tilde{p}) = \tilde{p}$ and $\psi^v(\varepsilon + \gamma, \tilde{p}) = p$ where $|\gamma|$ is small (but need not be zero). This means we can return to the *orbit* of the reference solution φ. Proposition 1.4 provides a computable sufficient condition to determine if $\varphi(-\varepsilon, p) \in$ int. $\mathscr{A}(\varepsilon, p, r)$.

Example 4.1. For the nonlinear system (17), we consider

$$\begin{aligned} \dot{x}_1 &= 4 + x_2 \sin u_2 + \sin(x_3 u_3), \\ \dot{x}_2 &= x_2 + u_2 + 3u_2^2 + u_3, \\ \dot{x}_3 &= u_2 + u_3^3, \end{aligned} \tag{22}$$

with $x(0) = p = (0, 0, 0)$. The associated semi-linearized system is $(\mathscr{D})$: $\dot{x} = X(x) + \sum_{i=2}^{3} v_i(t) Y^i(x)$, $x(0) = p$, where $X(x) = f(x, 0) = (4, x_2, 0)$, $Y^2(x) = f_{u_2}(x, 0) = (x_2, 1, 1)$, $Y^3(x) = f_{u_3}(x, 0) = (x_3, 1, 0)$. Then $X(p)$, $Y^2(p)$, $Y^3(p)$ are linearly independent and Theorem 4.1 applies.

We first consider the "linearized system (18)" associated with (22) and the reference solution $T^X(t)p = (4t, 0, 0)$. This has the form $\dot{y} = Ay + Bv$ where

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

Clearly rank $[B, AB, A^2B] = 2$; hence the variational system is not controllable and we cannot conclude, on this basis, that $T^X(t)p$ is in the interior of $\mathscr{F}(t, p)$, the attainable set of (22).

We next consider the semi-linearized system $(\mathscr{D})$, as above. Computing, we get $[Y^2, Y^3](x) = 0$; hence dim $L\{Y^2, Y^3\}(q) = 2$ for all q in a neighborhood of p; i.e., Y^2, Y^3 are involutive. Also (as expected from the lack of controllability of the variational equation) $(\mathrm{ad}^j X, Y^2)(x) = ((-1)^j x_2, 1, 0)$; $(\mathrm{ad}^j X, Y^3)(x) = (0, 1, 0)$ for $j \geqq 1$ so rank $\mathscr{S}^1(p) = 2 < 3$. Thus, with $Z(p) = (\frac{1}{4}, 0, 0)$ so $\langle Z(x), X(x) \rangle \equiv 1$, $\langle Z(x), Y^i(x) \rangle \equiv 0$, $i = 2, 3$, we have $a(1, 0, j) = a(0, 1, j) = 0 \quad \forall j \geqq 0$ and $\varphi_{1j}(s) \equiv 0$. We next compute higher order terms, specifically, $(\mathrm{ad}^2 Y^2, X)(x) = (-2, 0, 0)$ so $a(2, 0, 0) = -\frac{1}{2}$; $(\mathrm{ad}\, Y^3, (\mathrm{ad}\, Y^2, X))(x) = (-1, 0, 0)$ so $a(1, 1, 0) = -\frac{1}{4}$ while $(\mathrm{ad}^2 Y^3 X)(x) = 0$ and $a(0, 2, 0) = 0$. Then $r^* = 2$, $j^* = 0$, $\varphi_{r^* j^*}(s) = -\frac{1}{4}(s_2^2 - s_2 s_3)$ which changes sign in every neighborhood of $0 \in \mathbb{R}^2$. Proposition 1.4 shows that this is sufficient that $T^X(t)p \in$ int. $\mathscr{A}(t, p, \mathscr{D})$ $\forall t > 0$, and this holds if $|v_i(t)| \leqq r$ with $r > 0$. Thus Theorem 4.1 yields that for any $\varepsilon > 0$, $t_1 > 0$, $T^X(t_1)p \in$ int. $\bigcup_{|\gamma| \leqq \varepsilon} \mathscr{F}(t_1 + \gamma, p)$.

Acknowledgment. I would like to express my thanks to professor H. J. Sussmann for pointing out needed corrections and geometrical motivations for several results in the first version of this manuscript.

REFERENCES

[1] H. HERMES, *Local controllability and sufficient conditions in singular problems*, J. Differential Equations, 20, (1976), pp. 213–232.
[2] J. H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, Ibid., 12 (1972), pp. 95–116.
[3] H. HERMES, *On local and global controllability*, this Journal, 12 (1974), pp. 252–261.
[4] H. SUSSMANN AND V. JURDJEVIC, *Control systems on Lie groups*, J. Differential Equations, 12 (1972), pp. 313–329.
[5] R. M. HIRSCHORN, *Global controllability of nonlinear systems*, this Journal, to appear.
[6] L. W. CHOW, *Über Systeme von linearen partiellen Differential-gleichungen erster Ordnung*, Math. Ann., 117 (1940–41), pp. 98–105.
[7] G. W. HAYNES AND H. HERMES, *Nonlinear controllability via Lie theory*, this Journal, 8 (1970), pp. 450–460.
[8] T. NAGANO, *Linear differential systems with singularities and an application of transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.
[9] K. T. CHEN, *Decomposition of differential equations*, Math. Ann., 146 (1962), pp. 263–278.
[10] H. HERMES, *High order algebraic conditions for controllability*, Proceedings, Udine, Italy, to appear.
[11] R. W. BROCKETT, *System theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265–284.
[12] A. J. KRENER, *A decomposition theory for differentiable systems*, Proceedings IFAC, 1975.
[13] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 6, November 1976

A MORE DIRECT SOLUTION OF THE NEARLY SINGULAR LINEAR REGULATOR PROBLEM*

R. E. O'MALLEY, JR.†

Abstract. The asymptotic solution to the linear state regulator problem with cheap control is obtained for the situation where the limiting solution is a singular arc of first order and the initial impulse is that of a delta function. The presentation simplifies previous studies and allows generalization to further cases.

1. Introduction.

Let us consider the linear regulator problem

$$\dot{x} = Ax + Bu, \qquad 0 \leqq t \leqq T < \infty, \quad x(0) \text{ prescribed}, \tag{1}$$

with the scalar quadratic performance index

$$J(\varepsilon) = \frac{1}{2}\int_0^T (x'Qx + \varepsilon^2 u'Ru)\, dt \tag{2}$$

to be minimized for symmetric matrices $Q \geqq 0$ and $R > 0$. When $\varepsilon > 0$, we have a classical control problem with a unique easily computed solution (Kalman [16]), while we have a totally singular control problem when $\varepsilon = 0$ (cf. Jacobson [12], Gabasov and Kirillova [7] and Bell [2]). Recently, singular perturbation methods have enabled us to obtain the asymptotic solution of the nearly singular problem when the small positive parameter ε tends to zero (O'Malley and Jameson [31], Jameson and O'Malley [15] and O'Malley [29], [30]). These asymptotic studies not only allow us to investigate the impulsive nature of singular controls, but the results are of independent interest in several important control contexts (cf., e.g., Friedland [6], Jacobson, Gershwin, and Lele [13], Jacobson and Speyer [14], Kwakernaak and Sivan [19], Lions [21], Moylan and Anderson [24], Moylan [23], Wonham [38] and Womble [37]).

Just as the singular arc problem has been treated by various methods, several analogous techniques have been used for the nearly singular problem in the frequent situation that $B'QB > 0$. Preliminary changes of variables were used in O'Malley and Jameson [31] and O'Malley [29], while a singularly perturbed Riccati equation was integrated in O'Malley [30]. Case 1, where $B'QB > 0$, is the first of a sequence of interesting cases featuring successively more impulsive limiting controls and lower dimensional singular arcs. The cases correspond to generalized Legendre–Clebsch conditions of increasing order in singular control theory (cf. Goh [8] and Robbins [32]). This more direct approach in Case 1 allows simpler generalization to further cases and simpler determination of the asymptotic solution. The complicated generalizations possible are evident in the singular arc literature (cf., e.g., Moore [22]). A relatively straightforward generalization will be reported as a sequel to O'Malley and Jameson [31].

* Received by the editors July 22, 1975.

† Department of Mathematics, University of Arizona, Tucson, Arizona 85721. This work was supported by the Office of Naval Research under Contract No. N00014-67A-0209-0022.

For each $\varepsilon > 0$, classical results (cf., e.g., Anderson and Moore [1]) imply that the optimal control is given by

$$u = -\frac{1}{\varepsilon^2} R^{-1} B' p \tag{3}$$

where the costate vector p satisfies

$$\dot{p} = -Qx - A'p, \qquad p(T) = 0. \tag{4}$$

Eliminating the control from the state equation (1) results in the singularly perturbed two-point boundary value problem

$$\begin{aligned} \varepsilon^2 \dot{x} &= \varepsilon^2 Ax - BR^{-1}B'p, \qquad x(0) \text{ prescribed}, \\ \dot{p} &= -Qx - A'p, \qquad p(T) = 0, \end{aligned} \tag{5}$$

which, together with (3), is equivalent to the original control problem (1)–(2). Because $\varepsilon^2 A$ is nearly singular, this problem is beyond the scope of the most familiar singular perturbation techniques for two-point problems (cf. Harris [10], O'Malley [27], and Vasil'eva and Esipova [34]). We shall solve the problem by projecting into appropriate subspaces, noting that the technique has potentially wider applicability in other singular perturbation problems.

Having assumed that $B'QB > 0$ throughout $0 \leqq t \leqq T$, we will find the asymptotic solution of (5) in the form

$$\begin{aligned} x(t, \varepsilon) &= X(t, \varepsilon) + m(\tau, \varepsilon) + \varepsilon n(\sigma, \varepsilon), \\ p(t, \varepsilon) &= P(t, \varepsilon) + \varepsilon r(\tau, \varepsilon) + \varepsilon^2 s(\sigma, \varepsilon), \end{aligned} \tag{6}$$

where

$$\tau = \frac{t}{\varepsilon} \quad \text{and} \quad \sigma = (T - t)/\varepsilon \tag{7}$$

and m and $r \to 0$ as $\tau \to \infty$ while n and $s \to 0$ as $\sigma \to \infty$. (Although we shall find p, we note that only $B'p$ is needed in the solution of the control problem.) Within $(0, T)$, the solution will be asymptotically determined by the outer solution (X, P) since τ and σ are asymptotically infinite there. Near $t = 0$, the initial boundary layer correction $(m, \varepsilon r)$ becomes significant, while the terminal boundary layer correction $(\varepsilon n, \varepsilon^2 s)$ becomes important near $t = T$. The nonuniform endpoint convergence as $\varepsilon \to 0$ exhibited by the suggested representation (6) is, of course, typical of singular perturbation phenomena and of the transient endpoint behavior observed in singular arc problems. (The reader should note that previous applications of singular perturbation techniques in optimal control theory are surveyed by Kokotović, O'Malley, and Sannuti [18].) As usual, we shall separately obtain the outer solution and the boundary layer corrections as asymptotic series in ε. Then we shall determine all unspecified boundary values for the separate expansions and interpret the composite result asymptotically in its control context.

We observe that computation of these asymptotic expansions will require infinite differentiability of the coefficients A, B, Q and R in (1) and (2). Less

differentiability would require termination with the appropriate asymptotic approximations.

2. The outer solution. Since the solution (x, p) of (5) is asymptotic to (X, P) within $(0, T)$, it follows that the outer solution $(X(t, \varepsilon), P(t, \varepsilon))$ must satisfy the linear system (5) there. We shall solve (5) by seeking a solution in the feedback form

$$P(t, \varepsilon) = K(t, \varepsilon)X(t, \varepsilon) \tag{8}$$

(cf., e.g., Anderson and Moore). It follows, then, that the gain $K(t, \varepsilon)$ must satisfy the matrix Riccati equation

$$\varepsilon^2(\dot{K} + KA + A'K + Q) = KBR^{-1}B'K \tag{9}$$

for a symmetric matrix $K \geqq 0$, while X will satisfy the linear equation

$$\varepsilon^2\dot{X} = (\varepsilon^2 A - BR^{-1}B'K)X. \tag{10}$$

We note that boundary values for K and X are not yet obvious, though the representation (6) suggests that

$$K(T, 0) = 0 \tag{11}$$

since $P(T, 0) = 0$ and we cannot expect $X(T, 0) = 0$. Our development of the outer solution closely follows O'Malley [30]. It is included here because it is basic to the remainder of this paper and further work.

We shall first seek an asymptotic (regular perturbation) solution

$$K(t, \varepsilon) \sim \sum_{j=0}^{\infty} K_j(t)\varepsilon^j \tag{12}$$

to (9) and (11). When $\varepsilon = 0$, (9) implies the familiar singular arc condition

$$B'K_0 = 0 \tag{13}$$

(cf. Bryson and Ho [4]). This further implies that $BR^{-1}B'K_0$ is singular, so the usual singular perturbation theory (cf. O'Malley [27]) does not immediately apply to the initial value problem for (10) or the terminal value problem for (9). In particular, setting $\varepsilon = 0$ in (9) fails to determine K_0 unless B is nonsingular. (This special (but important) circumstance will be considered highly unusual in the presentation which follows.)

Postmultiplying (9) by B and, then, premultiplying by B' implies the two equations

$$\varepsilon^2[(KB)^{\cdot} + KB_1 + A'KB + QB] = KBR^{-1}B'KB$$

and

$$\varepsilon^2[B'(KB)^{\cdot} + B'KB_1 + B'A'KB + B'QB] = B'KBR^{-1}B'KB,$$

where

$$B_1 = AB - \dot{B}.$$

Lowest order terms in the second equation imply

$$(B'K_1B)R^{-1}(B'K_1B) = B'QB > 0.$$

Thus there is a unique solution

$$B'KB \sim \varepsilon B'K_1B = \varepsilon\sqrt{R^{1/2}(B'QB)R^{1/2}} > 0.$$

Knowing that $B'KB$ is nonsingular allows us to solve the first equation for KB and the second for $(B'KB)^{-1}$. Thus,

$$KB = \varepsilon^2[KB_1 + QB + (KB)^{\cdot} + A'KB](B'KB)^{-1}R. \tag{14}$$

We can thereby find the K_jB's successively, e.g.,

$$K_1B = (K_0B_1 + QB)(B'K_1B)^{-1}R.$$

Substituting back into (9) implies the regularly perturbed nonlinear equation

$$\begin{aligned} \dot{K} + KA + A'K + Q = {} & [KB_1 + QB + (KB)^{\cdot} + A'KB] \\ & \cdot [B'QB + B'(KB)^{\cdot} + B'KB_1 + B'A'KB]^{-1} \\ & \cdot [B_1'K + B'Q + (B'K)^{\cdot} + B'KA]. \end{aligned} \tag{15}$$

Differential equations for each term of the expansion (12) now follow by equating coefficients successively in (15). In particular, when $\varepsilon = 0$, we have the parameter-free Riccati equation

$$\dot{K}_0 + K_0A_1 + A_1'K_0 + Q_1 = K_0S_1K_0, \tag{16}$$

where

$$A_1 = A - B_1(B'QB)^{-1}B'Q,$$
$$Q_1 = Q - QB(B'QB)^{-1}B'Q,$$

and

$$S_1 = B_1(B'QB)^{-1}B_1'.$$

We note that standard results (cf., e.g., Anderson and Moore) imply the existence of a unique solution $K_0 \geqq 0$ to (16) satisfying $K_0(T) = 0$ since

$$Q_1 = (I - B(B'QB)^{-1}B'Q)'Q(I - B(B'QB)^{-1}B'Q) \geqq 0$$

and $S_1 \geqq 0$ because $B'QB > 0$. Indeed, any solution of the terminal value problem for K_0 automatically satisfies $B'K_0 = 0$ throughout $0 \leqq t \leqq T$ since $A_1B = \dot{B}$ and $Q_1B = 0$. Knowing K_0, we note that K_1B follows from (14). Higher order terms in (15) imply that, for $j \geqq 1$, K_j will satisfy a linear equation of the form

$$\dot{K}_j + K_j(A_1 - S_1K_0) + (A_1 - S_1K_0)'K_j = \alpha_j, \tag{17}$$

where α_j is determined by the K_i's with $i < j$. Thus, the Fredholm alternative for the linearized version of (16) implies the existence of a unique solution to the terminal value problem for (17).

One might suspect that our solution K is overdetermined since the vector $B'K$ follows termwise from both (14) and (17). That this is not so is clarified by

introducing the projection

$$E = I - B(B'K_1B)^{-1}B'K_1. \tag{18}$$

We note that

$$B'K_1E = 0, \quad EB = 0 \quad \text{and } E^2 = E. \tag{19}$$

Moreover, for $\dim x \geqq \dim u$, B and $B'K$ have rank equal to $\dim u$, so that the rank of E is $\dim x - \dim u$ (cf. Friedland [6] and Kwatny [20]). Since $B'K_0 = 0$, we have

$$K_0 = K_0' = K_0E = E'K_0.$$

For higher order terms K_j in (12), K_jB follows from equating coefficients termwise in (14). Then (18) implies that

$$K_j = K_jE + K_jB(B'K_1B)^{-1}B'K_1, \qquad j \geqq 1,$$

will be uniquely determined termwise through a linear terminal value problem of the form

$$\begin{gathered}(K_jE)^{\cdot} + K_jE(A_1 - S_1K_0) + (A_1 - S_1K_0)'E'K_j = \tilde{\alpha}_j, \\ K_j(T)E(T) \text{ specified.}\end{gathered} \tag{20}$$

Thus the determination of K is not overspecified and it will be uniquely determined by the terminal value $K(T, \varepsilon)E(T)$. (That this Riccati equation can be reduced from $\dim x$ to rank E is also shown by Kwatny.) Instead of using the projection E, one might instead use a pseudoinverse of B. We merely note that $K_1B(B'K_1B)^{-1}$ and $QB(B'QB)^{-1}$ are both generalized right inverses of B. We recall that $(B'K_1B)^{-1}B'K_1 = (B'QB)^{-1}(B'Q + B_1'K_0)$ is the generalized inverse used by Friedland.

To complete the outer solution, we must still solve the singularly perturbed linear state equation

$$\varepsilon\dot{X} = \varepsilon AX - \frac{1}{\varepsilon}BR^{-1}B'KX. \tag{21}$$

Again standard techniques do not apply because $BR^{-1}B'K_1$ is generally singular. We shall proceed by separately calculating $B'K_1X$ and EX termwise, thereby obtaining

$$X = B(B'K_1B)^{-1}B'K_1X + EX.$$

Multiplying (21) by $B'K_1$ and rearranging, we have

$$B'K_1X = \varepsilon\left[R(B'K_1B)^{-1}B'K_1(AX - \dot{X}) - \frac{B'}{\varepsilon^2}(K - \varepsilon K_1)X\right]. \tag{22}$$

Likewise, multiplying (21) by E and using $EB = 0$ implies

$$(EX)^{\cdot} = (\dot{E} + EA)[EX + B(B'K_1B)^{-1}B'K_1X]. \tag{23}$$

Let us now take

$$X(t, \varepsilon) \sim \sum_{j=0}^{\infty} X_j(t)\varepsilon^j. \tag{24}$$

Setting $\varepsilon = 0$ in (22) then implies the singular arc condition

$$B'K_1X_0 = 0 \tag{25}$$

(cf. Moylan and Moore [25]). Likewise setting $\varepsilon = 0$ in (23) implies that EX_0 will follow from

$$(EX_0) = (\dot{E} + EA)(EX_0) \tag{26}$$

once the initial value $E(0)X_0(0)$ is specified. (This confirms Ho's [11] observation that the singular arc behavior in $(0, T)$ is determined by a dynamical system of order dim x −dim u when $B'QB > 0$.) Higher order coefficients in (22) and (23) show that $B'K_1X_j$ is determined successively and that EX_j satisfies a nonhomogeneous version of the linear equation of (26). Thus, the X_j are all determined in terms of K and the unspecified initial value $E(0)X(0, \varepsilon)$.

Since the limiting state X_0 within $(0, T)$ satisfies the restriction $B'K_1X_0 = 0$, the singular arc is restricted to a lower dimensional trajectory. Such lowering of dimensionality is, of course, the characteristic feature of singular perturbation problems. Generally, the constructed outer solution cannot be uniformly valid throughout $0 \leqq t \leqq T$ because it will not satisfy the boundary conditions of (5). At $t = 0$, the condition $B'(0)K_1(0)X_0(0) = 0$ may be incompatible with the prescribed initial value $x(0)$. Thus, an initial boundary layer correction is needed to drive the limiting state into the null space of $B'K_1$. Higher order terms in the expansion $X(0, \varepsilon)$ may likewise be nonzero. At the terminal time, $B'(T)P(T, \varepsilon) = B'(T)K(T, \varepsilon)X(T, \varepsilon) \sim \varepsilon^2 B'(T)[K_1(T)X_1(T) + K_2(T)X_0(T)]$ will generally be nonzero and thereby inconsistent with the terminal condition $p(T, 0) = 0$. Thus, a terminal boundary layer correction is also needed.

We note that the nonuniform convergence at terminal time is not directly influenced by selection of the terminal value for KE. Thus, we might anticipate our later result that $K(T, \varepsilon)E(T) = 0$ in the belief that the asymptotic solution will be as simple as possible.

3. The initial boundary layer correction. The linearity of the equations (5) and the representation (6) imply that the initial boundary layer correction $(m, \varepsilon r)$ must satisfy the linear system

$$\begin{aligned} \frac{dm}{d\tau} &= \varepsilon Am - BR^{-1}B'r \\ \frac{dr}{d\tau} &= -Qm - \varepsilon A'r \end{aligned} \tag{27}$$

and tend to zero as the stretched variable τ tends to infinity. Setting

$$r(\tau, \varepsilon) = \frac{1}{\varepsilon}K(\varepsilon\tau, \varepsilon)m(\tau, \varepsilon) \tag{28}$$

will satisfy the differential equation for r since

$$\frac{dr}{d\tau} = \frac{dK}{dt}m + \frac{K}{\varepsilon}\left(\varepsilon A - \frac{1}{\varepsilon}BR^{-1}B'K\right)m = -Qm - A'Km$$

follows from the differential equation (9) for K. (The unmotivated Ansatz (28) is not as arbitrary as it may seem. It would result naturally if a Riccati solution of the full system (5) were attempted as in O'Malley [30]. Wilde and Kokotović [36] use the analogous procedure for another singularly perturbed regulator problem, observing that $K_0(0)$ satisfies an algebraic Riccati equation so that the limiting boundary layer problem can be interpreted as an "initial layer regulator" on the semi-infinite τ interval.) Using (28), there remains the differential equation

$$\frac{dm}{d\tau} = \varepsilon A m - \frac{1}{\varepsilon} B R^{-1} B' K m. \tag{29}$$

We now write

$$m = E(0)m + B(0)(B'(0)K_1(0)B(0))^{-1}B'(0)K_1(0)m,$$

and proceed to find $E(0)m$ and $B'(0)K_1(0)m$. Multiplying (29) by $E(0)$ and using $E(0)B(0) = 0$, we obtain

$$\frac{d}{d\tau}(E(0)m) = \varepsilon E(0)\left[A - \frac{1}{\varepsilon^2}(B - B(0))R^{-1}B'K\right]m. \tag{30}$$

(The right side is $O(\varepsilon)$ since $B'K = O(\varepsilon)$.) Likewise, multiplying (29) by $B'(0)K_1(0)$ and rearranging, we obtain

$$\begin{aligned} &\frac{d}{d\tau}(B'(0)K_1(0)m) + (B'(0)K_1(0)B(0))R^{-1}(0)B'(0)K_1(0)m \\ &\quad = \varepsilon B'(0)K_1(0)\left[A - \frac{1}{\varepsilon^2}BR^{-1}B'K - \varepsilon B(0)R^{-1}(0)B'(0)K_1(0))\right]m \end{aligned} \tag{31}$$

where the right side is again $O(\varepsilon)$. Together, (30) and (31) will allow the termwise determination of the decaying vector $m(\tau, \varepsilon)$ up to selection of the initial value $B'(0)K_1(0)m(0, \varepsilon)$. Thus, we set

$$m(\tau, \varepsilon) \sim \sum_{j=0}^{\infty} m_j(\tau)\varepsilon^j \tag{32}$$

and proceed.

Setting $\varepsilon = 0$ implies that

$$\frac{d}{d\tau}(E(0)m_0(\tau)) = 0$$

and

$$\frac{d}{d\tau}(B'(0)K_1(0)m_0(\tau)) + B'(0)K_1(0)B(0)R^{-1}(0)B'(0)K_1(0)m_0(\tau) = 0.$$

Since $m_0 \to 0$ as $\tau \to \infty$, we must have

$$\begin{aligned} &E(0)m_0(\tau) = 0 \\ &\text{and} \\ &B'(0)K_1(0)m_0(\tau) = R^{1/2}(0)\, e^{-C_0(0)\tau}R^{-1/2}(0)B'(0)K_1(0)m_0(0), \end{aligned} \tag{33}$$

where

$$C_0(0) = R^{-1/2}(0)B'(0)K_1(0)B(0)R^{-1/2}(0) > 0.$$

We note that the initial values for X_0 and m_0 are now determined since the representation (6) implies that

$$x(0) = X_0(0) + m_0(0).$$

Since $E(0)m_0(0) = 0$, however,

$$E(0)X_0(0) = E(0)x(0) \tag{34}$$

and this completely determines $X_0 = EX_0$(cf. (26)). Further,

$$B'(0)K_1(0)m_0(0) = B'(0)K_1(0)(x(0) - X_0(0)) = B'(0)K_1(0)x(0)$$

by (25). Thus,

$$\begin{aligned} m_0(\tau) = B(0)(B'(0)K_1(0)B(0))^{-1}R^{1/2}(0)\, e^{-C_0(0)\tau} \\ \cdot R^{-1/2}(0)B'(0)K_1(0)x(0) \end{aligned} \tag{35}$$

is completely specified. We note that this initial state transfer occurs in the range of $B(0)$. Moreover, since $K_0(0)B(0) = 0$, (28) implies that

$$r(\tau, 0) = K_1(0)m_0(\tau)$$

is also determined and $r = O(1)$ in spite of its representation (28).

Higher order coefficients m_j in the expansion (32) follow in succession. Thus, further terms in (30) imply that

$$E(0)m_j(\tau) = \int_\tau^\infty \beta_j(s)\, ds, \tag{36}$$

where β_j is known in terms of the m_l's with $l < j$ as an exponentially decaying vector. This also determines X_j since we now have the previously unspecified initial value

$$E(0)X_j(0) = -E(0)m_j(0). \tag{37}$$

Integrating (31), we find

$$\begin{aligned} B'(0)K_1(0)m_j(\tau) = -R^{1/2}(0)\Big[e^{-C_0(0)\tau}B'(0)K_1(0)X_j(0) \\ + \int_0^\tau e^{-C_0(0)(\tau - s)}\gamma_j(s)\, ds\Big], \end{aligned} \tag{38}$$

where $B'(0)K_1(0)X_j(0)$ and $\gamma_j(\tau)$ are already known. From (36) and (38), then, we have the exponentially decaying vector

$$m_j(\tau) = E(0)m_j + B(0)(B'(0)K_1(0)B(0))^{-1}B'(0)K_1(0)m_j.$$

Thus, the initial boundary layer correction has been constructed.

4. The terminal boundary layer correction. Linearity implies that the terminal boundary layer correction $(\varepsilon n(\sigma, \varepsilon), \varepsilon^2 s(\sigma, \varepsilon))$ must satisfy the linear

system

$$\begin{aligned} \frac{dn}{d\sigma} &= -\varepsilon An + BR^{-1}B's, \\ \frac{ds}{d\sigma} &= Qn - \varepsilon A's \end{aligned} \tag{39}$$

and tend to zero as $\sigma \to \infty$. This boundary layer correction should be determined in an analogous manner to the initial boundary layer correction. (This would be completely obvious had we considered the fixed endpoint problem.) As for (28), let us seek a solution to (39) in the form

$$s(\sigma, \varepsilon) = \frac{1}{\varepsilon} L(T - \varepsilon\sigma, \varepsilon) n(\sigma, \varepsilon), \tag{40}$$

where

$$\varepsilon^2(\dot{L} + LA + A'L + Q) = LBR^{-1}B'L. \tag{41}$$

Then, there remains

$$\frac{dn}{d\sigma} = -\varepsilon An + \frac{1}{\varepsilon} BR^{-1}B'Ln. \tag{42}$$

We recall that K is a symmetric, positive semidefinite solution of (41) satisfying $K(T, \varepsilon) = 0$. Taking $L = K$ will not, however, allow a nontrivial decaying solution to (42), as is generally needed. We instead select a symmetric negative semidefinite solution

$$L(t, \varepsilon) \sim \sum_{j=0}^{\infty} L_j(t)\varepsilon^j \tag{43}$$

of (41). It can be found termwise in a manner similar to K, with the terminal value $L(T, \varepsilon)E(T)$ yet to be determined. We note, in particular, that

$$L_0(T) = 0, \quad B'L_0 = 0, \quad B'L_1B = -B'K_1B$$

and (14) and (15) hold for L as well as K. (One could alternatively specify $-L$ as the positive semidefinite solution of the appropriate Riccati equation (cf. Wilde and Kokotović [36]).

Continuing, we multiply (42) by $E(T)$ to get

$$\frac{d}{d\sigma}(E(T)n) = \varepsilon\left[-E(T)An + \frac{1}{\varepsilon^2}E(T)(B - B(T))R^{-1}B'Ln\right] \tag{44}$$

since $E(T)B(T) = 0$. Upon setting $\varepsilon = 0$ and integrating, we have

$$E(T)n_0(\sigma) = 0. \tag{45}$$

Likewise, for $j \geqq 1$,

$$E(T)n_j(\sigma) = \int_{\sigma}^{\infty} \delta_j(s)\, ds \tag{46}$$

for successively known, exponentially decaying δ_j's. Since

$$n_j = E(T)n_j + B(T)(B'(T)K_1(T)B(T))^{-1}B'(T)K_1(T)n_j,$$

we must also find $B'(T)K_1(T)n_j$ successively. Multiplying (42) by $B'(T)K_1(T)$, however, yields

$$\begin{aligned}(47)\quad &\frac{d}{d\sigma}[B'(T)K_1(T)n] + B'(T)K_1(T)B(T)R^{-1}(T)[B'(T)K_1(T)n] \\ &= \varepsilon B'(T)K_1(T)\left[-A + \frac{1}{\varepsilon^2}(BR^{-1}B'L + \varepsilon B(T)R^{-1}(T)B'(T)K_1(T))\right]n,\end{aligned}$$

where $B'L = -\varepsilon B'(T)K_1(T) + O(\varepsilon^2)$. Integrating when $\varepsilon = 0$ implies the decaying solution

$$(48)\qquad B'(T)K_1(T)n_0(\sigma) = R^{1/2}(T)\, e^{-C_0(T)\sigma}R^{-1/2}(T)B'(T)K_1(T)n_0(0),$$

where

$$C_0(T) = R^{-1/2}(T)B'(T)K_1(T)B(T)R^{-1/2}(T) > 0.$$

From higher order terms, we obtain

$$\begin{aligned}(49)\quad B'(T)K_1(T)n_j(\sigma) = R^{1/2}(T)\Big[&e^{-C_0(T)\sigma}R^{-1/2}(T)B'(T)K_1(T)n_j(0) \\ &+ \int_0^\sigma e^{-C_0(T)(\sigma-s)}\zeta_j(s)\,ds\Big],\end{aligned}$$

where ζ_j is known successively. Thus, $B'(T)K_1(T)n(0,\varepsilon)$ is still unspecified, but the terminal boundary layer correction is otherwise completely determined and exponentially decaying.

We observe that this solution for the terminal boundary layer is much simpler than that of O'Malley [30] which first obtained a boundary layer correction for a Riccati gain for the full system (5) and used it to obtain terminal boundary layer corrections for the optimal state and control vectors.

5. The remaining boundary values. Thus far we have determined the asymptotic solution to (1)–(2) up to specifying the boundary values

$$K(T,\varepsilon)E(T), \quad L(T,\varepsilon)E(T), \quad \text{and} \quad B'(T)K_1(T)n(0,\varepsilon).$$

We shall now obtain all these values by imposing the terminal condition $p(T,\varepsilon) = 0$. Asymptotically

$$\begin{aligned}(50)\quad p(T,\varepsilon) &\sim P(T,\varepsilon) + \varepsilon^2 s(0,\varepsilon) \\ &= K(T,\varepsilon)X(T,\varepsilon) + \varepsilon L(T,\varepsilon)n(0,\varepsilon).\end{aligned}$$

Setting

$$K(T,\varepsilon) = K(T,\varepsilon)E(T) + K(T,\varepsilon)B(T)(B'(T)K_1(T)B(T))^{-1}B'(T)K_1(T),$$

$$X(T,\varepsilon) = E(T)X(T,\varepsilon) + B(T)(B'(T)K_1(T)B(T))^{-1}B'(T)K_1(T)X(T,\varepsilon),$$

and doing likewise for $L(T,\varepsilon)$ and $n(0,\varepsilon)$, (50) and the orthogonality properties

of E (cf. (19)) imply that

$$
\begin{aligned}
0 \sim\ & K(T,\varepsilon)E^2(T)X(T,\varepsilon) \\
& + K(T,\varepsilon)B(T)(B'(T)K_1(T)B(T))^{-1}B'(T)K_1(T)X(T,\varepsilon) \\
& + \varepsilon L(T,\varepsilon)E^2(T)n(0,\varepsilon) \\
& + \varepsilon L(T,\varepsilon)B(T)(B'(T)K_1(T)B(T))^{-1}B'(T)K_1(T)n(0,\varepsilon).
\end{aligned} \tag{51}
$$

Multiplying by $E'(T)$ and proceeding termwise, our knowledge of $E(T)X(T,\varepsilon)$ and $E(T)n(0,\varepsilon)$ imply that we must have

$$
K(T,\varepsilon)E(T) = 0 = L(T,\varepsilon)E(T). \tag{52}
$$

Multiplying (51) by $B'(T)$ then implies that

$$
\begin{aligned}
B'(T)K_1(T)n(0,\varepsilon) = & -\frac{1}{\varepsilon}(B'(T)K_1(T)B(T))(B'(T)L(T,\varepsilon)B(T))^{-1} \\
& (B'(T)K(T,\varepsilon)B(T))(B'(T)K_1(T)B(T))^{-1}B'(T)K_1(T)X(T,\varepsilon).
\end{aligned} \tag{53}
$$

Thus, all the needed boundary values are known termwise.

6. Conclusion. We have uniquely completely determined the uniform asymptotic expansions

$$
\begin{aligned}
x(t,\varepsilon) &= X(t,\varepsilon) + m(\tau,\varepsilon) + \varepsilon n(\sigma,\varepsilon), \\
p(t,\varepsilon) &= P(t,\varepsilon) + \varepsilon r(\tau,\varepsilon) + \varepsilon^2 s(\sigma,\varepsilon)
\end{aligned}
$$

for the state and costate vectors. The control law (3), then, implies that the optimal control has the expansion

$$
u(t,\varepsilon) = U(t,\varepsilon) + \frac{1}{\varepsilon}v(\tau,\varepsilon) + w(\sigma,\varepsilon), \tag{54}
$$

where

$$
U(t,\varepsilon) = -\frac{1}{\varepsilon^2}R^{-1}(t)B'(t)P(t,\varepsilon) = -\frac{1}{\varepsilon^2}R^{-1}(t)B'(t)K(t,\varepsilon)X(t,\varepsilon),
$$

$$
v(\tau,\varepsilon) = R^{-1}(\varepsilon\tau)B'(\varepsilon\tau)r(\tau,\varepsilon),
$$

and

$$
w(\sigma,\varepsilon) = R^{-1}(T-\varepsilon\sigma)B'(T-\varepsilon\sigma)s(\sigma,\varepsilon).
$$

Substituting the expansions for x and u into the performance index (2) implies that the optimal cost $J^*(\varepsilon)$ will have a power series expansion in ε with leading term determined by the limiting outer solution. Alternatively, usual Riccati gain considerations (cf. O'Malley [30]) imply that the optimal cost is given by

$$
J^*(\varepsilon) = \frac{1}{2}x'(0)K(0,\varepsilon)x(0) \sim \frac{1}{2}\sum_{l=0}^{\infty}(x'(0)K_l(0)x(0))\varepsilon^l \tag{55}
$$

since K has the expansion (12). Thus, we have the following

THEOREM. *Consider the state regulator problem*

$$\dot{x} = Ax + Bu, \qquad x(0) \text{ prescribed},$$

where

$$J(\varepsilon) = \frac{1}{2}\int_0^T (x'Qx + \varepsilon^2 u'Ru)\, dt$$

is to be minimized for $B'QB > 0$. *For each sufficiently small* $\varepsilon > 0$ *and each integer* $N \geqq 1$, *the optimal control, corresponding trajectory, and optimal cost are uniquely determined and satisfy*

$$u(t, \varepsilon) = \frac{v_0(\tau)}{\varepsilon} + \sum_{j=0}^{N} (U_j(t) + v_{j+1}(\tau) + w_j(\sigma))\varepsilon^j + O(\varepsilon^{N+1}),$$

$$x(t, \varepsilon) = X_0(t) + m_0(\tau) + \sum_{j=1}^{N} (X_j(t) + m_j(\tau) + n_{j-1}(\sigma))\varepsilon^j + O(\varepsilon^{N+1})$$

and

$$J^*(\varepsilon) = \frac{1}{2}\sum_{l=0}^{N} (x'(0)K_l(0)x(0))\varepsilon^l + O(\varepsilon^{N+1})$$

uniformly throughout $0 \leqq t \leqq 1$. *Here, the terms which are functions of* $\tau = t/\varepsilon$ *(or* $\sigma = (T-t)/\varepsilon$*) decay to zero as* $\tau \to \infty$ *(or* $\sigma \to \infty$*).*

Although the usual singular perturbation theorems don't apply to the preceding formal expansion technique, they do apply to the transformation approach of O'Malley and Jameson [31] and O'Malley [29] and to the Riccati method of O'Malley [30]. Those valid (unique) solutions coincide with our results here, and thereby justify them as well as the expansions of Friedland for a corresponding stochastic problem. We note that our previous transformation

$$u_1 = \int_0^t u(s)\, ds, \qquad x_1 = x - Bu_1$$

relates to our current use of the projection E and is of independent interest in control.

7. Final comments. (a) The optimal control obtained will be unbounded at $t = 0$ as $\varepsilon \to 0$, but bounded for each fixed $t > 0$. This impulsive behavior comes from the term $(1/\varepsilon)v_0(\tau)$ which is a multiple of the matrix

$$\frac{C_0(0)}{\varepsilon} e^{-C_0(0)t/\varepsilon},$$

and behaves like a matrix delta function (peaked at $t = 0$) in the limit $\varepsilon \to 0^+$.

(b) For $0 < t < T$, we have the limiting control

$$U_0 = -R^{-1}B'(K_1X_1 + K_2X_0).$$

Then (22), (26), and the relations $(B'K_1)\dot{E} + (B'K_1)\dot{}E = 0$ and $X_0 = EX_0$ imply the limiting feedback control law

$$U_0 = -(B'K_1B)^{-1}(B'K_1A + (B'K_1)\dot{})X_0. \tag{56}$$

We note that this result also follows from the limiting outer problem

$$\dot{X}_0 = AX_0 + BU_0, \qquad B'K_1X_0 = 0$$

(see also Moylan and Moore).

(c) We note that the optimal cost is determined by the Riccati gain $K(t, \varepsilon)$ of the outer solution, but it is not equal to the cost

$$\tfrac{1}{2}X'(0, \varepsilon)K(0, \varepsilon)X(0, \varepsilon)$$

of the outer solution. The two costs have the same limit $\frac{1}{2}x'(0)K_0(0)x(0)$ as $\varepsilon \to 0$ since

$$\begin{aligned} X_0'(0)K_0(0)X_0(0) &= X_0'(0)(E'(0)K_0(0)E(0))X_0(0) \\ &= (E(0)X_0(0))'K_0(0)(E(0)X_0(0)) = x'(0)K_0(0)x(0) \end{aligned}$$

by (34). Thus, the cost of the boundary layer corrections tends to zero with ε in spite of the unbounded initial impulse as $\varepsilon \to 0$.

(d) In the special controllable and observable situation when B and Q are invertible (as for a scalar control), the outer solution (U, X), the terminal correction (w, n), the projection E, and the limiting cost $J^*(0)$ will all be negligible. The initial boundary layer correction then allows transfer from the given state $x(0)$ to zero in the null space of $B'K_1$. It acts asymptotically like an infinite interval regulator (on $\tau \geqq 0$) with cost $\frac{1}{2}x'(0)K(0, \varepsilon)x(0) = O(\varepsilon)$.

(e) The semi-infinite interval problem with time invariant coefficients can be solved in obvious fashion under appropriate stabilizability-detectability hypotheses (cf. Jameson and O'Malley). The formal calculation of the initial boundary layer correction and the outer solution follow as in the finite interval case (except that the Riccati differential equation is replaced by an algebraic Riccati equation), while the outer solution decays to zero as $t \to \infty$ and the terminal boundary layer correction disappears.

(f) Several generalizations of this work should be pursued. Among them are:

(i) Asymptotic solutions when $B'QB > 0$ is not satisfied. If $B'QB$ is zero at isolated points of the interval, our solution technique breaks down and we encounter a turning point problem (cf. O'Malley [27]). If $B'QB$ is singular, but not zero, certain transformation techniques can be applied (cf. Moylan and Moore [25] and Hutton [39]). Problems where $B'QB = 0$ and $B_1'QB_1 > 0$ (and further cases) have been solved by Jameson and O'Malley and will be published elsewhere.

(ii) Related problems with bounded controls have been considered elsewhere (cf. Jacobson, Gershwin and Lele [13], Collins [5], Kokotović and Haddad [17] and Binding [3]). We mention only one example (cf. O'Malley [26]). The scalar problem

$$\dot{x} = u, \qquad x(0) = \frac{1}{2},$$

$$J(\varepsilon) = \frac{1}{2}\int_0^1 (x^2 + \varepsilon^2 u^2)\, dt, \qquad |u| \leqq k, \quad k > \frac{1}{2},$$

has the limiting solution

$$X_0 = \frac{1}{2} - kt, \qquad U_0 = -k \quad \text{for } 0 \leqq t < \frac{1}{2k}$$

$$X_0 = U_0 = 0, \qquad \frac{1}{2k} < t \leqq 1.$$

It is interesting to consider the limiting solution as $k \to \infty$. It converges to the optimal solution of the unconstrained problem $u = -\frac{1}{2}\delta(0)$.

(iii) Previous singular perturbation results (cf. Kokotović, O'Malley and Sannuti) could be combined with the preceding to provide the asymptotic solution to problems of the form

$$\dot{x} = A_{11}x + A_{12}z + B_1 u, \qquad x(0) \text{ given},$$

$$\mu \dot{z} = A_{21}x + A_{22}z + B_2 u, \qquad z(0) \text{ given},$$

with performance index

$$J(\varepsilon) = \frac{1}{2}\int_0^T \left[\begin{pmatrix} x \\ z \end{pmatrix}' Q \begin{pmatrix} x \\ z \end{pmatrix} + u'Ru \right] dt,$$

where $R = \operatorname{diag}(R_1, \varepsilon^2 R_2, \delta^2 R_3)$ and μ, ε and δ simultaneously tend to zero in appropriately related ratios. Here difficulties implementing an open loop control would be encountered if A_{22} had eigenvalues with positive real parts. As Wilde and Kokotović [36] explained in an analogous situation, a partially-closed loop stabilizing control would then be preferable. No such problem occurs for the original problem (1)–(2), since small parameters don't multiply derivatives in the state equation (1).

(iv) Nonlinear generalizations could be considered, as was done in other control contexts by Hadlock [9], O'Malley [28], and Sannuti [33].

REFERENCES

[1] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[2] D. J. BELL, *Singular problems in optimal control—A survey*, Internat. J. Control, 21 (1975), pp. 319–331.

[3] P. BINDING, *Singularly perturbed optimal control problems, I. Convergence*, this Journal, 14 (1976), pp. 591–612.

[4] A. E. BRYSON, JR AND Y.-C. HO, *Applied Optimal Control*, Blaisdell, Waltham, Mass., 1969.

[5] W. D. COLLINS, *Singular perturbations in linear time-optimal control*, Recent Mathematical Developments in Control, D. J. Bell, ed., Academic Press, New York, 1973, pp. 123–136.

[6] B. FRIEDLAND, *Limiting forms of optimal stochastic linear regulators*, Trans. ASME Ser. G. J. Dynamic Systems, Measurement and Control, 93 (1971), pp. 134–141.

[7] G. R. GABASOV AND F. M. KIRILLOVA, *High order necessary conditions for optimality*, this Journal, 10 (1972), pp. 127–168.

[8] B. S. GOH, *Necessary conditions for singular extremals involving multiple control variables*, this Journal, 4 (1966), pp. 716–731.

[9] C. R. HADLOCK, *Existence and dependence on a parameter of solutions of a nonlinear two point boundary value problem*, J. Differential Equations, 14 (1973), pp. 498–517.

[10] W. A. HARRIS, JR., *Singularity Perturbed Boundary Value Problems Revisited*, Lecture Notes in Mathematics 312, Springer-Verlag, Berlin, 1973, pp. 54–64.

[11] Y.-C. HO, *Linear stochastic singular control problems*, J. Optimization Theory Appl., 9 (1972), pp. 24–31.
[12] D. H. JACOBSON, *Totally singular quadratic minimization problems*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 651–658.
[13] D. H. JACOBSON, S. B. GERSHWIN AND M. M. LELÈ, *Computation of optimal singular controls*, Ibid., AC-15 (1970), pp. 67–73.
[14] D. H. JACOBSON AND J. L. SPEYER, *Necessary and sufficient conditions for optimality for singular control problems: A limit approach*, J. Math. Anal. Appl., 34 (1971), pp. 239–266.
[15] A. JAMESON AND R. E. O'MALLEY, JR., *Cheap control of the time-invariant regulator*, Appl. Math. Optimization, 1 (1975), pp. 337–354.
[16] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.
[17] P. V. KOKOTOVIĆ AND A. H. HADDAD, *Controllability and time-optimal control of systems with slow and fast modes*, IEEE Trans. Automatic Control, AC-20 (1975), pp. 111–113.
[18] P. V. KOKOTOVIĆ, R. E. O'MALLEY, JR., AND P. SANNUTI, *Singular perturbations and order reduction in control theory—An overview*, Automatica—J. IFAC, 12 (1976), pp. 123–132.
[19] H. KWAKERNAAK AND R. SIVAN, *The maximally achievable accuracy of linear optimal regulators and linear optimal filters*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 79–86.
[20] H. G. KWATNY, *Minimal order observers and certain singular problems of optimal estimation and control*, Ibid., AC-19 (1974), pp. 274–276.
[21] J. L. LIONS, *Perturbations Singulières dans les Problèmes aux Limites et en Contrôle Optimal*, Lecture Notes in Mathematics 323, Springer-Verlag, Berlin, 1973.
[22] J. B. MOORE, *The singular solution to a singular quadratic minimization problem*, Internat. J. Control, 20 (1974), pp. 383–393.
[23] P. J. MOYLAN, *A note on Kalman–Bucy filters with zero measurement noise*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 263–264.
[24] P. J. MOYLAN AND B. D. O. ANDERSON, *Nonlinear regulator theory on an inverse optimal control problem*, Ibid., 18 (1973), pp. 460–465.
[25] P. J. MOYLAN AND J. B. MOORE, *Generalizations of singular optimal control theory*, Automatica—J. IFAC, 7 (1971), pp. 591–598.
[26] R. E. O'MALLEY, JR., *Examples illustrating the connection between singular perturbations and singular arcs*, Proceedings, Eleventh Annual Allerton Conference on Circuit and System Theory, Univ. of Illinois, Urbana, 1973, pp. 678–685.
[27] ———, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.
[28] ———, *Boundary layer methods for certain nonlinear singularly perturbed optimal control problems*, J. Math. Anal. Appl., 45 (1974), pp. 468–484.
[29] ———, *The singular perturbation approach to singular arcs*, Int. Conf. on Differential Equations, H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 595–611.
[30] ———, *The nearly singular linear regulator problem*, Bol. Soc. Mat. Mexicana, 1976.
[31] R. E. O'MALLEY, JR. AND ANTONY JAMESON, *Singular perturbations and singular arcs* I, IEEE Trans. Automatic Control, AC-20 (1975), pp. 218–226.
[32] H. M. ROBBINS, *A generalized Legendre–Clebsch condition for the singular cases of optimal control*, IBM J. Res. Develop., 3 (1967), pp. 361–372.
[33] P. SANNUTI, *Asymptotic expansions of singularly perturbed quasi-linear optimal systems*, this Journal, 13 (1975), pp. 572–592.
[34] A. B. VASIL'EVA AND V. A. ESIPOVA, *Conditionally stable singularly perturbed systems*, Soviet Math. Dokl., 15 (1974), pp. 720–724.
[35] R. R. WILDE AND P. V. KOKOTOVIĆ, *Stability of singularly perturbed systems and networks with parasitics*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 245–246.
[36] ———, *Optimal open and closed loop control of singularly perturbed linear systems*, Ibid., AC-18 (1973), pp. 616–626.
[37] M. E. WOMBLE, *The linear-quadratic Gaussian problem with ill-conditioned Riccati matrices*, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Mass., 1972.
[38] W. M. WONHAM, *Linear Multivariable Control*, Lecture Notes in Economic and Mathematical Systems 101, Springer-Verlag, Berlin, 1974.
[39] M. F. HUTTON, *Solutions of the singular stochastic regulator problem*, Trans, ASME Ser. G, J. Dynamic Systems, Measurement, and Control, 95 (1973), pp. 414–417.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 6, November 1976

THE MAXIMUM PRINCIPLE UNDER MINIMAL HYPOTHESES*

FRANK H. CLARKE†

Abstract. We consider the optimal control system

$$\dot{x}(t) = f(t, x(t), u(t)), \qquad u(t) \in U(t) \quad \text{a.e.}$$

with given initial and terminal constraints and a cost functional. We derive necessary conditions for optimality in a form similar to Pontryagin's maximum principle under hypotheses which are in a certain sense minimal in order that the problem be meaningful. In particular we do not assume $f(t, s, u)$ continuous in u or differentiable in s, nor do we require $U(t)$ or $f(t, s, U(t))$ to be bounded or closed. These necessary conditions, which are expressed in terms of certain "generalized Jacobians," reduce to the usual ones when classical hypotheses are imposed.

1. Introduction.

In considering a control problem governed by the equation

$$\dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e.}, \qquad u(t) \in U(t) \quad \text{a.e.}, \tag{1}$$

we may ask for the minimal hypotheses on f so that the problem "makes sense." The measurable control function u having been chosen, the possible solutions x to the differential equation (1) are considered. Consequently, to begin with, we would want to be sure that the function $t \to f(t, s, u(t))$ is Lebesgue measurable for each s. Next, to avoid ambiguity, we would like to be sure that at most one solution x exists for a given initial condition. The purpose of this paper is to obtain necessary conditions for an optimal control under essentially the hypotheses that assure these properties. This amounts to a weak measurability hypothesis on $f(t, s, u)$ with respect to (t, u) and a Lipschitz condition in s. In particular we assume neither the continuity of f in u, nor the closedness or boundedness of $U(t)$. The results are in the form of Pontryagin's maximum principle, but do not require the existence of derivatives.

As one might expect from the above, the paper employs methods quite different from any previous work on the subject. We use a theorem of I. Ekeland [6] to obtain slightly perturbed problems that admit solutions, these perturbations involving discontinuous cost functionals. We then show that certain of the author's previous results [4] apply to these nonstandard problems, and make use of a limiting process.

The plan of the paper is as follows: § 2 presents and discusses the results, § 3 consists of preliminary lemmas, while the proofs of the main results appear in § 4.

2. Main results.

We shall be dealing with a function $f: [0, 1] \times R^n \times R^m \to R^n$ and a *multifunction* $U: [0, 1] \to R^m$ (i.e., $U(t)$ is a subset of R^m for each t in $[0, 1]$). The choice of the time interval $[0, 1]$ is a normalization for notational convenience; "a.e." will mean "for almost all t in $[0, 1]$" (Lebesgue measure).

* Received by the editors August 11, 1975.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5, and Mathématiques de la Décision, Université de Paris IX (Dauphine), Paris, France 75116. This work was supported in part by the National Research Council of Canada under Grant A 9082.

A Lebesgue measurable function $u\colon [0, 1]\to R^m$ satisfying $u(t)\in U(t)$ a.e. will be called a *control.* Let X be a given closed subset of R^n. An *admissible trajectory* x corresponding to a control u is an absolutely continuous function $x\colon [0, 1]\to R^n$ satisfying (1) and also $x(t)\in X$ for all t. We do not assume that every control yields an admissible trajectory, or indeed any trajectory at all. We call x *interior* if in addition $x(t)$ belongs to the interior of X for each t.

We let L be the collection of Lebesgue measurable subsets of $[0, 1]$ and B^k the Borel subsets of R^k. We denote by $\boldsymbol{L\times B^k}$ the σ-algebra of subsets of $[0, 1]\times R^k$ generated by products of sets in L and B^k. The usual Euclidean norm is denoted $|s|$, where s belongs to some R^k.

The following are the hypotheses that will be made:

(H_1) For each s in a neighborhood of X, the function $(t, u)\to f(t, s, u)$ is $L\times B^m$-measurable.

(H_2) There is a function k in $L^1(0, 1)$ such that for t in $[0, 1]$, u in $U(t)$, and s_1, s_2 in a neighborhood of X,

$$|f(t, s_1, u)-f(t, s_2, u)|\leqq k(t)|s_1-s_2|.$$

(H_3) The *graph* of U (i.e., the set $\{(t, u)\in[0, 1]\times R^m : u\in U(t)\}$) is $L\times B^m$-measurable.

Remark 1. (H_1) is the least measurability hypothesis assuring that $t\to f(t, s, u(t))$ is L-measurable for every L-measurable $u(t)$, and is satisfied if f is L-measurable in t and continuous in u [8, Prop. 3]. (H_1) is also satisfied if every coordinate function f^i of f is (upper or lower) semicontinuous in (t, u).

If U is closed-valued, (H_3) is equivalent (see Proposition 1, § 3) to the following: For every closed set S, the set $\{t : U(t)\cap S\neq\phi\}$ is L-measurable. (H_3) is satisfied in particular if $U(t)$ is for every t a given Borel set U_0.

DEFINITION 1. Let $g : R^n\to R^k$ be locally Lipschitz. We define the *generalized Jacobian* $\partial g(s)$ of g at s as follows:

$$\partial g(s) = \text{co}\{\lim Dg(s_i)\},$$

where we consider all sequences $\{s_i\}$ converging to s such that the usual Jacobian matrix $Dg(s_i)$ exists, as well as the limit of the sequence $\{Dg(s_i)\}$ (co denotes convex hull).

When $k = 1$, we obtain the "generalized gradient" as defined in [1]; it follows as in that case that Definition 1 yields a nonempty convex compact set of $k\times n$ matrices. Note that if g is C^1, $\partial g(s) = \{Dg(s)\}$ (or more generally if g is "strongly differentiable" at s). We also defined in [1] an extension of the generalized gradient to functions $g : R^n\to(-\infty, \infty]$ that are lower semicontinuous; this concept enters into Corollary 1 below. Finally we mention the concept of the *normal cone* $N_C(s)$ to an arbitrary closed set C at s in C, also defined in [1]. We limit ourselves here to recalling that $N_C(s)$ reduces to the usual normal space if C is a C^1 manifold, and to the normal cone in the sense of convex analysis if C is a convex set. The statement that a vector ζ is normal to C at s means, of course, that ζ belongs to $N_C(s)$.

For a given closed set C in R^n, we define $A(C)$, the *attainable set* starting from C, as the collection of all points $x(1)$ where x is an admissible trajectory such that $x(0) \in C$. The basic result is the following:

THEOREM 1. *We posit* (H_1)–(H_3). *If the control* v *generates an interior admissible trajectory* z *with* $z(0)$ *in* C *such that* $z(1)$ *lies on the boundary of* $A(C)$, *then there exists a nonvanishing absolutely continuous functton* $p:[0, 1] \to R^n$ *such that*:

$$-\dot{p}(t) \in p(t)\partial_s f(t, z(t), v(t)) \quad a.e., \tag{2}$$

$$p(t) \cdot f(t, z(t), v(t)) = \sup\{p(t) \cdot f(t, z(t), u) : u \in U(t)\} \quad a.e., \tag{3}$$

$$p(0) \text{ is normal to } C \text{ at } z(0). \tag{4}$$

(The notation $\partial_s f$ refers of course to the generalized Jacobian of the function $s \to f(t, s, v(t))$, and $p\partial_s f$ is matrix multiplication.)

Remark 2. The hypothesis that z is an interior trajectory serves to eliminate state constraints, which are not considered here.

We may use Theorem 1 (and a variant, Theorem 1′ in § 4) to derive necessary conditions for optimality. We shall prove

COROLLARY 1. *We posit* (H_1)–(H_3), *and we suppose given a lower semicontinuous function* $l: R^n \times R^n \to (-\infty, \infty]$. *If the interior admissible trajectory* z *corresponding to the control* v *minimizes* $l(x(0), x(1))$ *over all admissible trajectories* x, *then there exist an absolutely continuous function* $p:[0, 1] \to R^n$ *and a number* λ *equal to* 0 *or* -1 *such that* $|p(t)| + |\lambda|$ *is nonvanishing,* p *satisfies* (2) *and* (3), *and*

$$(p(0), -p(1), \lambda) \text{ is normal to epi } l \text{ at the point } (z(0), z(1), l(z(0), z(1))). \tag{5}$$

Remark 3. epi l is the set $\{(s_0, s_1, r) \in R^n \times R^n \times R : l(s_0, s_1) \leqq r\}$. Relation (5) is a "transversality condition" of a very general nature. If $\lambda = -1$ (the conditions are then said to be "normal") (5) is equivalent to $(p(0), -p(1)) \in \partial l(z(0), z(1))$, the generalized gradient of l. Note that constaints such as $x(1) \in C_1$ are implicitly accounted for by letting l be $+\infty$ when $x(1) \notin C_1$. Corollary 1, for example, implies the following:

COROLLARY 2. *Let the control* v *and corresponding interior admissible trajectory* z *minimize*

$$\int_0^1 f^0(t, x(t), u(t))\, dt$$

over the admissible trajectories x *and corresponding control* u *such that* $x(0) \in C_0$, $x(1) \in C_1$ *and the integral above is defined, where* C_0 *and* C_1 *are closed. If* (H_1)–(H_3) *are satisfied for* f *replaced by* $\bar{f} = (f^0, f)$, *then there exist an absolutely continuous function* $p:[0, 1] \to R^n$ *and a number* λ *equal to* 0 *or* -1 *such that* $|p(t)| + |\lambda|$ *is nonvanishing,*

$$-\dot{p}(t) \in p(t)\partial_s f(t, z(t), v(t)) + \lambda \partial_s f^0(t, z(t), v(t)) \quad a.e., \tag{6}$$

$$\begin{aligned} &p(t) \cdot f(t, z(t), v(t)) + \lambda f^0(t, z(t), v(t)) \\ &\quad = \sup\{p(t) \cdot f(t, z(t), u) + \lambda f^0(t, z(t), u) : u \in U(t)\} \quad a.e., \end{aligned} \tag{7}$$

$$\begin{aligned} &p(0) \text{ is normal to } C_0 \text{ at } z(0), \\ &-p(1) \text{ is normal to } C_1 \text{ at } z(1). \end{aligned} \tag{8}$$

Remark 4. As will be shown in the proof, the set X may be allowed to depend on t, as long as for some positive ε, the ε-neighborhood of $z(t)$ is contained in $X(t)$ for each t. (H_2) may also be weakened as follows: If $U_j(t)$ is the set

$$\{u \in U(t) : |u - \nu(t)| \leqq j, |f(t, z(t), u) - f(t, z(t), \nu(t))| \leqq j\}$$

$j = 1, 2, \cdots$, then there exists for each j some k_j in $L^1(0, 1)$ such that (H_2) is satisfied for u in $U_j(t)$.

One may show by counterexample that (H_3) cannot be dispensed with; a counterexample exists with $f(t, s, u) = u$.

None of the well-known general theories of necessary conditions seem capable of treating the optimal control problem in the above generality. However, J. Warga [9]–[11] has recently derived necessary conditions without differentiability, stated in terms of "derivative containers" conceptually related to the generalized Jacobian. He requires f to be continuous in u and U compact-valued, but treats state constraints.

3. Preliminary lemmas. Let K be a multifunction from a measure space T to R^n (t will for the moment denote a point of T). We say K is *measurable* if for every closed set C the following set is measurable in T:

$$\{t \in T : K(t) \cap C \neq \varnothing\}.$$

We shall say K is *G-measurable* if its graph (see (H_3)) is measurable in $T \times R^n$ with respect to the product σ-algebra of T with the Borel sets in R^n. The following is an extension by Rockafellar [7, Thm. 2] of a result of Debreu.

PROPOSITION 1. *If K is closed-valued and measurable, then K is G-measurable. If K is G-measurable and T is complete and σ-finite, then K is measurable.*

We now show that we may (and henceforth do) suppose that for $u \notin U(t)$, $f(t, s, u) = 0$ (say). The function $\tilde{f}$ so redefined continues to satisfy (H_1): if $0 \notin V$, for a closed set V, then

$$\tilde{f}^{-1}(\cdot, s, \cdot)(V) = f^{-1}(\cdot, s, \cdot)(V),$$

while if 0 belongs to V we have

$$\tilde{f}^{-1}(\cdot, s, \cdot)(V) = \{f^{-1}(\cdot, s, \cdot)(V) \cap \operatorname{graph}(U)\} \cup \{\operatorname{graph}(U)\}^c,$$

and this lies in $L \times B^m$ by (H_3). Consequently neither the hypotheses nor the conclusions of the theorem are affected.

LEMMA 1. *Let f satisfy (H_1) and (H_2), and let $x_0(t)$ be a continuous function taking values in X. Then the mapping $(t, u) \to f(t, x_0(t), u)$ is $L \times B^m$-measurable.*

Proof. It is easy to verify the above when x_0 is a step function. Now let $x_0 = \lim x_i$ where each x_i is a step function and the limit is uniform. Then for any closed set V in R^n, letting V_j be the closed $1/j$ neighborhood of V, we have

$$h_0^{-1}(V) = \bigcap_j \bigcup_{i \geqq j} h_i^{-1}(V_j),$$

where h_i is the map $(t, u) \to f(t, x_i(t), u)$, $i = 0, 1, \cdots$. This completes the proof. (Note that the result is generally false if x_0 is merely L-measurable, because $L \times B^n$ is not complete.) Q.E.D.

LEMMA 2. *Under* (H_3) *and the hypotheses of Lemma* 1, *the multifunction* $t \to \operatorname{cl} f(t, x_0(t), U(t))$ *is measurable.*

Proof. By [7, Cor. 1.2] it suffices to prove that the multifunction $K: t \to f(t, x_0(t), U(t))$ is measurable. Consider the sets S_1 and S_2 defined by

$$S_1 = \{(t, u, f(t, x_0(t), u)) : t \in [0, 1], u \in R^m\},$$

$$S_2 = \{(t, u, s) : t \in [0, 1], u \in U(t), s \in R^n\}.$$

S_1 is $L \times B^m \times B^n$-measurable by Lemma 1 and Proposition 1, while S_2 is $L \times B^m \times B^n$-measurable by (H_3). It follows that the following set is $L \times B^{m+n}$-measurable:

$$S_1 \cap S_2 = \{(t, u, f(t, x_0(t), u)) : t \in [0, 1], u \in U(t)\}.$$

Invoking Proposition 1 once again, we derive the measurability of $t \to \{(u, f(t, x_0(t), u)) : u \in U(t)\}$. But K, as the projection on R^n of this multifunction, is then measurable (this is easily seen directly from the definition of measurability). Q.E.D.

Suppose now the theorem to be true under the following added hypothesis:

(H_4) $\qquad f(t, z(t), U(t))$ is bounded by $\alpha(t)$, for some α in $L^1(0, 1)$.

For every positive integer j let us define

$$U_j(t) = \{u \in U(t) : |f(t, z(t), u) - f(t, z(t), v(t))| \leqq j\}.$$

The graph of U_j is seen to be the intersection of graph (U) with the set

$$S = \{(t, u) \in [0, 1] \times R^m : |f(t, z(t), u) - f(t, z(t), v(t))| \leqq j\}.$$

Since $(t, u) \to |f(t, z(t), u) - f(t, z(t), v(t))|$ is $L \times B^m$-measurable by Lemma 1, S is $L \times B^m$-measurable and it follows that graph(U_j) is $L \times B^m$-measurable. We now note that all the hypotheses of the theorem remain in force if U_j replaces U, and of course U_j satisfies (H_4). We could then apply the theorem to deduce the existence of an absolutely continuous p_j satisfying (2)–(4). We may suppose $|p_j(0)| = 1$. It follows from [1, Prop. 1.11] that $\partial_s(p \cdot f) = p\partial_s f$, and this fact combined with [1, 1.4] and [7, Thm. 3] can be used to show that the multifunction $t \to \partial_s f(t, z(t), v(t))$ is measurable. We have $\partial_s f(t, z(t), v(t))$ bounded by $k(t)$, by (H_2). These facts enable us to apply [4, Lemma 8] to conclude that a subsequence of $\{p_j\}$ converges uniformly to an absolutely continuous p satisfying (2). It follows that p is nonvanishing, satisfies (4), and satisfies (3) for U. We have just proved:

PROPOSITION 2. *Without loss of generality we may posit* (H_4).

Suppose now the theorem proved under the following added hypothesis:

(H_5) $\qquad U(t)$ is bounded for each t.

We now define U_j as follows:

$$U_j(t) = \{u \in U(t) : |u - v(t)| \leqq j\}.$$

Once again the graph of U_j is $L \times B^m$-measurable, and the same argument as above yields:

PROPOSITION 3. *Without loss of generality we may posit* (H_5).

The two preceding results explain the assertion in Remark 4. The following lemma makes use of some recent work of Dauer and Van Vleck [5].

LEMMA 3. *Let* $v(t)$ *be* L*-measurable, and suppose*

$$v(t) \in \operatorname{cl} f(t, x_0(t), U(t)) \quad a.e.$$

Then, under the hypotheses of Lemma 2 *and under* (H_5), *for any positive* δ *there is a control* u_0 *such that*

$$|v(t) - f(t, x_0(t), u_0(t))| \leqq \delta \quad a.e.$$

Proof. Let $\Gamma(t) = \{(u, f(t, x_0(t), u)) : u \in U(t)\}$. We showed in Lemma 2 that $S_1 \cap S_2$, the graph of Γ, is $L \times B^{m+n}$-measurable. Hence Γ is measurable (Proposition 1) and consequently so is $\operatorname{cl} \Gamma$ (by [7, Cor. 1.2]). By Proposition 1, $\operatorname{cl} \Gamma$ is G-measurable. Now let

$$K(t) = \{(u, v(t)) : u \in R^m, (u, v(t)) \in \operatorname{cl} \Gamma(t)\}.$$

Then K is closed-valued and nonempty (because U is bounded). Since graph(K) is the intersection of graph$(\operatorname{cl} \Gamma)$ and the set $\{(t, u, v(t)) : t \in [0, 1], u \in R^m\}$, we deduce that K is G-measurable and hence measurable. By [7, Cor. 1.1] K has a measurable selector $(u(t), v(t))$. Thus

$$(u(t), v(t)) \in \operatorname{cl} \Gamma(t) \quad \text{a.e.}$$

We now apply [5, Thm. 2] to deduce the existence of a measurable selector $(u_0(t), f(t, x_0(t), u_0(t)))$ for Γ such that

$$|(u_0(t) - u(t), f(t, x_0(t), u_0(t)) - v(t))| \leqq \delta \quad \text{a.e.} \qquad \text{Q.E.D.}$$

LEMMA 4. *Let the interior admissible trajectory* z *minimize* $g(x)$ *over the admissible trajectories* x *satisfying* $x(0) \in C$, *where* g *is continuous in the sup norm and where* (H_1)–(H_5) *hold. Then* z *minimizes* $g(x)$ *over the absolutely continuous functions* $x : [0, 1] \to R^n$ *satisfying* $x(t) \in \operatorname{int} X$, $x(0) \in C$, *and*

$$\dot{x}(t) \in E(t, x(t)) \quad a.e.,$$

where $E(t, s) = \operatorname{cl} f(t, s, U(t))$.

Proof. Note that z is certainly feasible for the new problem. Suppose for some x_0 as described we had $g(x_0) < g(z)$. We set $K = \exp\left(\int_0^1 k(t)\, dt\right)$ and choose $\varepsilon > 0$ so that the ε-neighborhood of $x_0(t)$ lies in X for each t. Now let δ be any positive number less than ε / K. We let u_0 be as in Lemma 3, for $v = \dot{x}_0$. Then

$$\int_0^1 |\dot{x}_0(t) - f(t, x_0(t), u_0(t))| dt \leqq \delta,$$

and by [2, Prop. 2] there is an absolutely continuous solution x to the equation

$$\dot{x}(t) = f(t, x(t), u_0(t)) \quad \text{a.e.}, \qquad x(0) = x_0(0),$$

such that the sup norm of $x - x_0$ is no greater than $K\delta$. But then for δ sufficiently small we obtain an admissible trajectory x with $x(0)$ in C and $g(x) < g(z)$, a contradiction. Q.E.D.

LEMMA 5. *Let a multifunction $F:[0,1]\times R^n \to R^n$ have nonempty convex compact values, where $F(t,s)$ is G-measurable, and upper semicontinuous in s. Suppose there exist a subset S of R^n and α in $L^1(0,1)$ such that for all s in S, for all v in $F(t,s)$, $|v| \leqq \alpha(t)$. Let $\{x_j\}$ be a sequence of absolutely continuous functions and A_j a sequence of L-measurable subsets of $[0,1]$ such that*

(i) $\dot{x}_j(t) \in F(t, x_j(t))$ *for t in A_j*

(ii) L-*measure*$(A_j) \to 1$,

(iii) $x_j(t) \in S$ *for all t,*

(iv) $|\dot{x}_j(t)| \leqq \alpha(t)$ *a.e.*

(v) $\{x_j(0)\}$ *is bounded.*

Then there is a subsequence of $\{x_j\}$ that converges uniformly to an absolutely continuous function x satisfying $\dot{x}(t) \in F(t, x(t))$ a.e.

Proof. This is a standard result when each $A_j = [0,1]$ (cf. [4, Lemma 8]). We sketch what remains essentially a standard proof. We first apply the theorem of Arzelà–Ascoli to get a uniformly convergent subsequence, converging to x, say, and then we invoke the Dunford–Pettis criterion to extract a subsequence of $\{\dot{x}_j\}$ converging weakly to v (say) in $L^1(0,1)$ (we do not bother relabeling subsequences). From $x_j(t) = x_j(0) + \int_0^t \dot{x}_j$ we deduce $x(t) = x(0) + \int_0^t v$ and hence x is absolutely continuous and $\dot{x} = v$ a.e. Now let $h(t,s,p)$ be the support function of $F(t,s)$; i.e., $h(t,s,p) = \max\{p \cdot \zeta : \zeta \in F(t,s)\}$. Fix p in R^n and any L-measurable set V in $[0,1]$. We can prove the integrability of $t \to h(t, x_j(t), p)$ with the help of [7, Thm. 3]. Then, letting χ_j be the characteristic function of A_j,

$$
\begin{aligned}
0 &\leqq \limsup \int_{V \cap A_j} (h(t, x_j, p) - p \cdot \dot{x}_j)\, dt \\
&\leqq \int_V \limsup \chi_j h(t, x_j, p)\, dt + \limsup \int_{V \cap A_j} (-p \cdot \dot{x}_j)\, dt \\
&\leqq \int_V h(t, x, p)\, dt + \limsup \int_V (-p \cdot \dot{x}_j)\, dt + \limsup \int_{V \cap A_j^c} p \cdot \dot{x}_j\, dt \\
&= \int_V (h(t, x, p) - p \cdot \dot{x})\, dt.
\end{aligned}
$$

Since V is arbitrary, it follows that $h(t, x(t), p) \geqq p \cdot \dot{x}(t)$ a.e., and we arrive at $\dot{x}(t) \in F(t, x(t))$ a.e. by obtaining this for a dense set of vectors p. Q.E.D.

LEMMA 6. *Let U be an abstract set, s_0 a point in R^n and $g: R^n \times U \to R$ a function such that $g(s, U)$ is bounded for all s near s_0. We suppose that for some constant K, for all u in U, and for all s_1, s_2 near s_0, we have $|g(s_1, u) - g(s_2, u)| \leqq K|s_1 - s_2|$. We set $f(s) = \sup\{g(s,u) : u \in U\}$, and we let S be a given subset of R^n of L-measure 0. Then f is Lipschitz near s_0, and $\partial f(s_0) \subset \Gamma(s_0)$, where*

$$\Gamma(s) = \operatorname{co}\{\lim \nabla_s g(s_i, u_i)\}$$

where we consider all such limits such that $\lim s_i = s$, $s_i \notin S$, $u_i \in U$, $\nabla_s g(s_i, u_i)$ exists, and $g(s, u_i) \to f(s)$ (i.e., $\{u_i\}$ is a maximizing sequence for $g(s, \cdot)$).

Proof. That f is Lipschitz is easily verified (see [1, Thm. 2.1]), as well as the fact that Γ assumes nonempty convex compact values and is upper semicontinuous. By the way in which ∂f is defined, it therefore suffices to show that $\Gamma(s)$

contains $\nabla f(s)$ whenever the latter exists (which is a.e.). We first prove the following: for any vector α in R^n,

$$\limsup \{g(s+h+\lambda\alpha, u)-g(s+h, u)\}/\lambda \leqq \max \alpha \cdot \Gamma(s), \tag{9}$$

where the lim sup is taken as $h \to 0$ in R^n, λ decreases to 0 and $g(s, u) \to f(s)$ with $u \in U$. We may suppose α different from 0. Let ε be any positive number, and set $m = \max \alpha \cdot \Gamma(s)$. Then for some positive δ, whenever $|s-s'| < \delta$, $s' \notin S$, $u \in U$, and $|g(s, u)-f(s)| < \delta$, then

$$\alpha \cdot \nabla_s g(s', u) \leqq m+\varepsilon$$

(if the derivative exists).

Fix any such u and let N be the set of measure 0 where $g(\cdot, u)$ fails to be differentiable. For almost all h, the ray $s+h+\lambda\alpha$ meets N in a set of 0 (one-dimensional) measure. Thus if we take any such h with $|h| < \delta/2$ and any positive $\lambda < \delta/(2|\alpha|)$, we have

$$\begin{aligned} g(s+h+\lambda\alpha, u)-g(s+h, u) &= \int_0^\lambda \alpha \cdot \nabla_s g(s+h+t\alpha, u)\, dt \\ &\leqq \lambda(m+\varepsilon). \end{aligned}$$

Since $g(\cdot, u)$ is continuous, this must in fact be true for all h with $|h| < \delta/2$ and $\lambda < \delta/(2|\alpha|)$. Thus we obtain (9).

Now suppose $\nabla f(s)$ exists, and let λ_i decrease to 0. We have

$$\begin{aligned} \alpha \cdot \nabla f(s) &= \lim \{f(s+\lambda_i\alpha)-f(s)\}/\lambda_i \\ &\leqq \limsup \{g(s+\lambda_i\alpha, u_i)-g(s, u_i)\}/\lambda_i \\ &\quad \text{(where } u_i \text{ in } U \text{ satisfies } g(s+\lambda_i\alpha, u_i) > f(s+\lambda_i\alpha)-\lambda_i^2\text{)} \\ &\leqq \max \alpha \cdot \Gamma(s) \end{aligned}$$

by (9), since $g(s, u_i) \to f(s)$. Since α is arbitrary, we deduce $\nabla f(s) \in \Gamma(s)$ by [1, Cor. 1.10]. Q.E.D.

The following result is distilled from [4, Thm. 1] (compare [4, Cor. 2]). A trajectory for a multifunction $E(t, s)$ is an absolutely continuous function $x : [0, 1] \to R^n$ satisfying $\dot{x}(t) \in E(t, x(t))$ a.e.

LEMMA 7. *Let E be compact-valued, integrably bounded, measurable in t and Lipschitz in s (in the Hausdorff metric) on an open region containing the trajectory z, and suppose z minimizes (locally) $g_0(x(0))+g_1(x(1))$ over the trajectories x for E satisfying $x(0) \in C$ (we suppose $z(0)$ lies in C) where C is closed. If g_0 and g_1 are locally Lipschitz, there is an absolutely continuous function p from $[0, 1]$ to R^n and a vector v such that*

$$(-\dot{p}(t), \dot{z}(t)) \in \partial H(t, z(t), p(t)) \quad \text{a.e.}, \tag{10}$$

$$p(0)-v \text{ is normal to } C \text{ at } z(0), \text{ and } v \in \partial g_0(z(0)), \tag{11}$$

$$-p(1) \in \partial g_1(z(1)), \tag{12}$$

where $H(t, s, p) = \max p \cdot E(t, s)$ and the generalized gradient in (10) is taken with respect to (s, p).

4. Proof of the theorem. We remind the reader that we have shown (Propositions 2 and 3) that we lose nothing in assuming (H_4) and (H_5). The following definitions and the next two lemmas will set the stage for applying the main theorem of I. Ekeland's elegant paper [6, Thm. 1.1].

Let V be the set of pairs (u, s) such that s lies in C and u is a control yielding an admissible trajectory $x_{u,s}$ such that $x_{u,s}(0) = s$. (H_2) guarantees the uniqueness of $x_{u,s}$. For controls u_1 and u_2 we set

$$\delta(u_1, u_2) = L\text{-measure}\{t \in [0, 1] : u_1(t) \neq u_2(t)\}.$$

For points (u_1, s_1) and (u_2, s_2) in V we define

$$\Delta((u_1, s_1), (u_2, s_2)) = \delta(u_1, u_2) + |s_1 - s_2|.$$

It is easily verified that Δ is a metric on V. We now show completeness (cf. [6, Lemma 7.2]).

LEMMA 8. *Let $\{(u_i, s_i)\}$ be a Cauchy sequence in V (relative to Δ). Then there is an element (u_0, s_0) in V such that (u_i, s_i) converges to (u_0, s_0).*

Proof. Since the sequence is Cauchy, it suffices to show that a subsequence converges to some (u_0, s_0) in V. We may extract a subsequence (we do not relabel) so that

$$\delta(u_i, u_{i+1}) + |s_i - s_{i+1}| < 2^{-i}.$$

It follows as in [6, Lemma 7.2] that a control u_0 exists such that $\delta(u_i, u_0) \to 0$. The completeness of C in the Euclidean norm allows us to suppose that s_i converges to some s_0 in C. It remains to prove that (u_0, s_0) lies in V. We let x_i be the admissible trajectory x_{u_i, s_i}, and we set

$$A_i = \{t \in [0, 1] : u_i(t) = u_0(t)\}.$$

Then L-measure $(A_i) \to 1$, and

$$\dot{x}_i(t) = f(t, x_i(t), u_0(t)), \qquad t \in A_i.$$

If we let $F(t, s) = f(t, s, u_0(t))$, $S = X$, then (H_1)–(H_4) allow us to apply Lemma 5 and obtain a subsequence of $\{x_i\}$ converging uniformly to an absolutely continuous x_0 such that $x_0(0) = s_0$ and $\dot{x}_0(t) = f(t, x_0(t), u_0(t))$ a.e. Necessarily $x_0(t)$ lies in X, since X is closed; thus x_0 is an admissible trajectory. Q.E.D.

We retain the notation of the previous lemma.

LEMMA 9. *If (u_i, s_i) converges to (u_0, s_0) in V, then x_i converges uniformly to x_0.*

Proof. We have $\dot{x}_i(t) - f(t, x_i(t), u_0(t)) = 0$ on A_i, and off A_i we have $|\dot{x}_i(t) - f(t, x_i(t), u_0(t))|$ bounded by an integrable function (independent of i) by (H_4). Since L-measure$(A_i) \to 1$, it follows that for any positive δ, for i sufficiently large,

$$\int_0^1 |\dot{x}_i(t) - f(t, x_i(t), u_0(t))| dt < \delta.$$

We apply [2, Prop. 2] as in Lemma 4 to deduce that an absolutely continuous y_i exists satisfying $y_i(0) = x_i(0)$, $|x_i(t) - y_i(t)| \leqq K\delta$ for each t, and $\dot{y}_i(t) =$

$f(t, y_i(t), u_0(t))$ a.e. From

$$|\dot{x}_0(t)-\dot{y}_i(t)| = |f(t, x_0(t), u_0(t))-f(t, y_i(t), u_0(t))|$$
$$\leqq k(t)|x_0(t)-y_i(t)|,$$

we derive, for each t,

$$|x_0(t)-y_i(t)| \leqq K_1|x_0(0)-y_i(0)| = K_1|s_0-s_i|.$$

Then, for i large, for each t,

$$|x_0(t)-x_i(t)| \leqq |x_0(t)-y_i(t)|+|y_i(t)-x_i(t)|$$
$$\leqq K_1|s_0-s_i|+K\delta. \qquad \text{Q.E.D.}$$

Now for a positive integer j we choose $\zeta_j \notin A(C)$ such that $|\zeta_j - z(1)| < 1/j$ (this is possible since $z(1)$ lies on the boundary of $A(C)$) and we define F on V by

$$F(u, s) = |x_{u,s}(1)-\zeta_j|.$$

Then F is continuous by Lemma 9, and clearly $(\nu, z(0))$ in V satisfies

$$F(\nu, z(0)) \leqq \inf_V F + 1/j.$$

We now apply [6, Thm. 1.1] (with $\lambda = j^{-1/2}$) to deduce the existence of an element (u_j, s_j) in V satisfying (we set $\varepsilon_j = j^{-1/2}$)

$$\Delta((u_j, s_j), (\nu, z(0))) \leqq \varepsilon_j, \tag{13}$$

$$F(u, s)+\varepsilon_j\Delta((u, s), (u_j, s_j)) \geqq F(u_j, s_j) \tag{14}$$

for all (u, s) in V.

Let us now establish a notational convention: $\bar{s}$ will denote a point of R^{n+1}, where we have added a zeroth coordinate to s. Thus $\bar{s} = (s^0, s)$. Similarly a trajectory $\bar{x} = (x^0, x)$, etc. We define $\bar{X} = R \times X$, $\bar{C} = \{0\} \times C$. If $x_j = x_{u_j,s_j}$, we let $\bar{x}_j$ equal $(0, x_j)$.

Define $\alpha : R^m \times R^m \to R$ by

$$\alpha(u, v) = \begin{cases} 1 & \text{if } u \neq v, \\ 0 & \text{if } u = v, \end{cases}$$

and define $\bar{f}: [0, 1] \times R^{n+1} \times R^{n+1} \to R^{n+1}$ by

$$\bar{f}(t, \bar{s}, u) = (\varepsilon_j\alpha(u, u_j(t)), f(t, s, u)).$$

We note that $(t, u) \to \alpha(u, u_j(t))$ is $L \times B^m$-measurable, so that all the hypotheses on f remain in force for $\bar{f}$. We may interpret relation (14) as follows: for every control u and point $\bar{s}$ of $\bar{C}$, if $\bar{x}$ is an absolutely continuous solution to $\bar{x}(t) \in \bar{X}$,

$$\dot{\bar{x}}(t) = \bar{f}(t, \bar{x}(t), u(t)) \quad \text{a.e.}, \qquad \bar{x}(0) = \bar{s},$$

then

$$|x(1)-\zeta_j|+\varepsilon_j|s-s_j|+x^0(1) \geqq |x_j(1)-\zeta_j|.$$

Thus $\bar{x}_j$ is optimal for a certain control problem. In view of (13) and Lemma 9, for j sufficiently large $\bar{x}_j$ is interior to $\bar{X}$. We apply Lemma 4 to deduce that $\bar{x}_j$ minimizes (locally) $g_0(\bar{x}(0))+g_1(\bar{x}(1))$ over the trajectories for $E(t, \bar{s})$ satisfying $\bar{x}(0)\in\bar{C}$, where

$$g_0(\bar{s}) = \varepsilon_j|s-s_j|,$$
$$g_1(\bar{s}) = |s-\zeta_j|+s^0,$$
$$E(t, \bar{s}) = \operatorname{cl} \bar{f}(t, \bar{s}, U(t)).$$

E is compact-valued by (H_4), measurable by Lemma 2 and Lipschitz in $\bar{s}$ by (H_2). We may thus apply Lemma 7 to obtain an absolutely continuous $\bar{p}_j = (p_j^0, p_j)$ satisfying (10)–(12) for some $\bar{v}_j$. Relation (11) yields

$$|v_j| \leqq \varepsilon_j, \tag{15}$$

$$p_j(0) - v_j \text{ is normal to } C \text{ at } x_j(0). \tag{16}$$

Note that $x_j(1)\neq\zeta_j$ since $\zeta_j\notin A(C)$; thus (12) yields

$$|p_j(1)| = 1, \tag{17}$$

$$p_j^0(1) = -1. \tag{18}$$

The function $H(t, \bar{s}, \bar{p})$ is in this case

$$\sup\{p^0\varepsilon_j\alpha(u, u_j(t))+p\cdot f(t, s, u): u\in U(t)\},$$

which is independent of s^0. Thus, from (10), $\dot{p}_j^0(t)=0$ a.e., and hence p_j^0 is identically -1.

Let us define $h:[0, 1]\times R^n\times R^n\times R\to R$ by

$$h(t, s, p, q) = \sup\{q\alpha(u, \nu(t))+p\cdot f(t, s, u): u\in U(t)\}. \tag{19}$$

Note that $H(t, \bar{s}, \bar{p}) = h(t, s, p, \varepsilon_j p^0)$ whenever t is such that $u_j(t)=\nu(t)$. Relation (10), which may be written

$$(-\dot{p}_j, 0, \dot{x}_j, 0)\in\partial H(t, x_j, 0, p_j, -1),$$

is thus seen to imply

$$(-\dot{p}_j, \dot{x}_j, 0)\in\partial h(t, x_j, p_j, -\varepsilon_j) \quad \text{when } u_j(t)=\nu(t). \tag{20}$$

We set $A_j=\{t\in[0, 1]: u_j(t)=\nu(t)\}$ and we note (by (13)) that L-measure $(A_j)\to 1$. Relation (10) yields

$$|\dot{p}_j(t)| \leqq k(t)|(p_j(t), -\varepsilon_j)|$$

and this along with (17) implies that p_j is uniformly bounded (independently of j). It then follows that the set in (20) is integrably bounded (see [4, Lemma 7 and Remark]). We apply Lemma 5 to obtain a subsequence of $\{(x_j, p_j, -\varepsilon_j)\}$ converging uniformly to $(z, p, 0)$ (in view of (13), Lemma 9 and the fact that $\varepsilon_j\to 0$) where p is absolutely continuous, $|p(1)|=1$, and

$$(-\dot{p}(t), \dot{z}(t), 0)\in\partial h(t, z(t), p(t), 0) \quad \text{a.e.} \tag{21}$$

In light of (15) and (16), $p(0)$ is normal to C at $z(0)$.

Fix a value of t such that (21) holds; we now interpret (21) by means of Lemma 6. We let S be the subset of X where $f(t, \cdot, \nu(t))$ fails to be differentiable. We obtain that $(-\dot{p}(t), \dot{z}(t), 0)$ is a convex combination of points of the form (see (19))

$$\lim (p_i D_s f(t, s_i, u_i), f(t, s_i, u_i), \alpha(u_i, \nu(t))) \tag{22}$$

where $s_i \notin S$, $s_i \to z(t)$, $p_i \to p(t)$, $q_i \to 0$ and

$$p \cdot f(t, z(t), u_i) \to h(t, z(t), p(t), 0). \tag{23}$$

Since the last coordinates of these points must be 0, we see that $u_i = \nu(t)$ for all i sufficiently large. Now (22) is seen to yield a point of the form

$$(p(t)\zeta, f(t, z(t), \nu(t)), 0)$$

where ζ belongs to $\partial_s f(t, z(t), \nu(t))$. Since $\partial_s f$ is convex, we obtain (2). We already have (4), and (23) yields (3). Any p satisfying (2) and vanishing once is identically 0, hence p is never 0. Q.E.D.

A variant of Theorem 1. Let a locally Lipschitz function $g: R^n \to R^k$ be given, and let us define

$$A_g(C) = \{g(s) : s \in A(C)\}.$$

Let an interior admissible trajectory z be such that $z(0)$ lies in C and $g(z(1))$ lies on the boundary of $A_g(C)$. We may proceed exactly as in the proof of Theorem 1, the term $|x(1) - \zeta_j|$ in the definition of F becoming $|g(x(1)) - \zeta_j|$. Relation (12), which had given (17), would now yield instead

$$-p_j(1) \in w_j \partial g(x_j(1))$$

for some unit vector $w_j = (g(x_j(1)) - \zeta_j)/|g(x_j(1)) - \zeta_j|$ in R^k. After converging subsequences are taken, we obtain

$$-p(1) \in w \partial g(z(1)). \tag{24}$$

We summarize:

THEOREM 1′. *Let the hypotheses of Theorem* 1 *stand, with* $A(C)$ *replaced by* $A_g(C)$ *for* $g: R^n \to R^k$ *locally Lipschitz. Then there exist an absolutely continuous* p *and a unit vector* w *in* R^k *such that* (2)–(4) *hold, and also* (24).

Remark 5. Theorem 1 is the case $g(s) = s$ of the above; in the more general setting of Theorem 1′, we cannot assert that p is nonzero, since (24) might allow $p(1) = 0$.

Proof of Corollary 1. We shall denote by $\tilde{s}$ points (s^1, s^2, s^3) in $R^n \times R^n \times R$. We define $C = \text{epi } l$ (a closed set), $\tilde{f}(t, \tilde{s}, u) = (f(t, s^1, u), 0, 0)$, $g(\tilde{s}) = (s^3, s^2 - s^1) \in R \times R^n$, $\tilde{z}(t) = (z(t), z(1), l(z(0), z(1)))$. We claim $\tilde{z}(1)$ lies on the boundary of the attainable set $A_g(C)$ corresponding to $\tilde{f}$ and $\tilde{X} = X \times R^n \times R$. If not, there is a control u and a corresponding admissible trajectory $\tilde{x}$ with $\tilde{x}(0) \in C$, $x^2(0) = x^2(1) = x^1(1)$ and

$$l(x^1(0), x^1(1)) = l(x^1(0), x^2(0)) \leqq x^3(0) = x^3(1) < l(z(0), z(1)).$$

But then x^1 is an admissible trajectory (for f and u) better than z, a contradiction.

We apply Theorem 1′ to obtain $\tilde{p}$ satisfying (2)–(4) for $\tilde{f}$, and (24). We deduce

$$-\dot{\tilde{p}} \in (p^1 \partial_s f(t, z, v), 0, 0),$$
$$\tilde{p} \cdot \tilde{f}(t, \tilde{z}, v) = \sup\{\tilde{p} \cdot \tilde{f}(t, \tilde{z}, u) : u \in U(t)\},$$
$$-\tilde{p}(1) = (-w^2, w^2, w^1),$$

where $|(w^1, w^2)| = 1$,

$$\tilde{p}(0) = (p^1(0), -w^2, -w^1) \quad \text{is normal to } C \text{ at } \tilde{z}(0) = (z(0), z(1), l(z(0), z(1))). \tag{25}$$

We deduce $w^1 \geqq 0$ (from (25)) and $p^1(1) = -p^2(1) = w^2$. Thus (25) implies (setting $p^1 = p$)

$$(p(0), -p(1), -w^1) \text{ is normal to epi } l \text{ at } (z(0), z(1), l(z(0), z(1))). \tag{26}$$

If $w^1 = 0$, then w^2 (and hence p) is nonvanishing. If $w^1 > 0$, we replace p by p/w^1 and obtain (2), (3) and (5) with $\lambda = -1$. Q.E.D.

Proof of Corollary 2. We denote as in the proof of Theorem 1 points (s^0, s) of $R \times R^n$ by $\bar{s}$. We define

$$\bar{f}(t, \bar{s}, u) = (f^0(t, s, u), f(t, s, u)),$$
$$l(\bar{s}_0, \bar{s}_1) = \begin{cases} s_1^0 & \text{if } s_0^0 = 0, \quad s_0 \in C_0, \quad s_1 \in C_1, \\ +\infty & \text{otherwise.} \end{cases}$$

We apply Corollary 1 to this problem with $\bar{z}(t)$ equal to $(\int_0^t f^0(\tau, z(\tau), v(\tau))\, d\tau, z(t))$, and a straightforward translation of terms yields the result. Q.E.D.

REFERENCES

[1] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[2] ———, *Admissible relaxation in variational and control problems*, J. Math. Anal. Appl., 51 (1975), pp. 557–576.

[3] ———, *The generalized problem of Bolza*, this Journal, 14 (1976), pp. 682–699.

[4] ———, *Extremal arcs and extended Hamiltonian systems*, Trans. Amer. Math. Soc., to appear.

[5] J. P. DAUER AND F. S. VAN VLECK, *Measurable selectors of multifunctions and applications*, Math. Systems Theory, 7 (1974), pp. 367–376.

[6] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl., 47 pp. pl. 324–353.

[7] R. T. ROCKAFELLAR, *Measurable dependence of convex sets and functions on parameters*, Ibid., 28 (1969), pp. 4–25.

[8] ———, *Existence theorems for general control problems of Bolza and Lagrange*, Advances in Math., 15 (1975), pp. 312–333.

[9] J. WARGA, *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 18 (1975), pp. 41–62.

[10] ———, *Controllability and necessary conditions in unilateral problems without differentiability assumptions*, this Journal, 14 (1976), pp. 546–573.

[11] ———, *Derivative containers, inverse functions, and controllability*, Proc. International Symposium on the Calculus of Variations and Optimal Control, D. L. Russell, ed., Academic Press, New York, 1976.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 6, November 1976

OPTIMAL CONTROL OF AUTONOMOUS LINEAR PROCESSES WITH SINGULAR MATRICES IN THE QUADRATIC COST FUNCTIONAL*

STEPHEN L. CAMPBELL†

Abstract. The optimal control of the autonomous linear process $\dot{x} = Ax + Bu$ with quadratic cost functional is studied. The initial and terminal times and positions are fixed. The matrices in the cost functional are allowed to be singular. An assumption, weaker than invertibility, is placed on the coefficient matrices. Under this assumption, necessary and sufficient conditions are given for the existence of an optimal control in terms of the initial and final position of the process. A closed form for the optimal control is given.

1. Introduction. Closed forms for all solutions of the linear system

$$C\dot{x} + Ex = f \tag{1}$$

were derived in [5] for the case when $(\mu C + E)^{-1}$ existed for some scalar μ. Neither C nor E were required to be invertible. This paper will show how the results of [5] can be used to analyze the optimal control of certain autonomous linear processes with singular matrices in the quadratic cost functional. A particular problem will be presented and handled in detail. Additional problems that may be analyzed by the same techniques are described.

Optimal control problems with singular matrices in the quadratic cost functional have received much attention. They occur naturally as a first order approximation to more general optimal control problems. Reference [11] surveys the known results on one such problem with singular matrices in the cost.

Our work has the advantage that it leads to explicit solutions for the problem studied, as well as a procedure for solution. These explicit, closed form solutions, also simplify the proof and development of the mathematical theory for the problem studied.

2. The control problem. The following notation and terminology is fixed throughout this paper.

Let A, B be $n \times n$ and $n \times m$ matrices, respectively. We allow all matrices and scalars to be complex though, of course, in many applications they are real. An asterisk denotes the conjugate transpose. The usual inner product for complex (or real) vectors is denoted $\langle \cdot, \cdot \rangle$. Let Q, H be positive semi-definite $m \times m$ and $n \times n$ matrices. Finally, let x, u denote vector-valued functions of the real variable t. x is $n \times 1$ while u is $m \times 1$.

Now consider the autonomous control process

$$\dot{x} = Ax + Bu \tag{2}$$

on the time interval $[t_0, t_1]$ with quadratic cost functional

$$J[x, u] = \frac{1}{2}\int_{t_0}^{t_1} \langle Hx, x\rangle + \langle Qu, u\rangle \, dt. \tag{3}$$

* Received by the editors March 17, 1975, and in final revised form January 19, 1976.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27607.

This work was supported in part by a grant from the North Carolina Engineering Foundation.

If one has a fixed pair of vectors x_0, x_1 such that there exists controls u so that the process x is at x_0 at time t_0 and x_1 at time t_1, then one can ask for a control that minimizes the cost (3) subject to the restraint that $x(t_0)=x_0$, $x(t_1)=x_1$.

Using the theory of Lagrange multipliers one gets the system of equations

$$\begin{aligned} \lambda + A^*\lambda + Hx &= 0, \\ \dot{x} - Ax - Bu &= 0, \\ B^*\lambda + Qu &= 0, \end{aligned} \tag{4}$$

as necessary conditions for optimization [1].

If Q is invertible, then u can be eliminated from the second equation and the resulting system formed by the first two equations solved directly. We shall be most interested then in the case when Q is not invertible, though our results will include the case when Q is invertible.

The system (4) can be rewritten as

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} + \begin{bmatrix} A^* & H & 0 \\ 0 & -A & -B \\ B^* & 0 & Q \end{bmatrix} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{5}$$

Note that (5) is in the form of (1) with C, and possibly E, not invertible.

Our results differ in certain key respects from most of the existing literature on this problem. We do not assume the pair (A, B) is controllable. Rather we assume that the matrix $Q_\mu = Q - B^*(\mu + A^*)^{-1}H(\mu - A)^{-1}B$ is invertible for large $|\mu|$. This amounts to being able to *formally* solve (4) uniquely by Laplace transform methods. The matrix Q_μ has appeared before in connection with this control problem. In [20] it is used to describe when (3) on the time interval $[t_0, \infty)$ is bounded below. In [16] it is used to establish existence conditions for solutions of the algebraic Riccati equation. These solutions may then be utilized to rewrite the process (2) so that the eigenvalues of the new A lie either in the right or left half-planes. That is, x satisfies a stability type of condition. Controllability is assumed in both [16] and [20]. The matrix in the cost functional of [16] which corresponds to our Q is invertible. Thus [16] is developed under the assumption that Q_μ is invertible for large $|\mu|$ (see Proposition 4).

Our results proceed as follows. We first establish that (4) provides both necessary and sufficient conditions for optimization. Needed results from [5] are then summarized. Using the results of [5], we find all solutions of (4) under the assumption that Q_μ is invertible for large $|\mu|$. Solution of (4) consists of two parts. One is determining for which x_0, x_1 an optimal control exists. This is given by Theorem 4. For such x_0, x_1, Theorem 3 gives the control u in terms of x, λ. An explicit formula for the constant feedback matrix is given in (27). Both x and λ are explicitly given in terms of x_0, λ_0 in (26). How to calculate λ_0 from x_0, x_1 is given by the discussion immediately preceding Theorem 4. An example is worked. The assumption that Q_μ is invertible is then discussed. It is shown to be equivalent to the uniqueness of optimal controls for those x_0, x_1 which admit an optimal control. It is also shown to be equivalent to the nonexistence of "free" trips with nonzero

controls. In the last section, several additional control problems which can be handled by our techniques are given.

We assume throughout that controls are continuous. All statements concerning optimality are made with respect to the control problem of this section.

3. Optimality of solutions. We shall first show that if (4) has a solution satisfying the boundary conditions, then u must be an optimal control.

THEOREM 1. *Suppose that x, u, λ is a solution of* (4) *and $x(t_0) = x_0$, $x(t_1) = x_1$. Then u is an optimal control.*

Proof. To show that $J[\hat{x}, \hat{u}] \geqq J[x, u]$ for all $\hat{x}, \hat{u}$ satisfying (2) and the boundary conditions, it is clearly equivalent to show that

$$\phi(s) = J[sx + (1-s)\hat{x}, su + (1-s)\hat{u}]$$

has a minimum at $s = 1$ for all $\hat{x}, \hat{u}$. A direct calculation gives that ϕ has a minimum at $s = 1$ if and only if

$$\int_{t_0}^{t_1} \langle Hx, \hat{x} - x \rangle \, dt = \int_{t_0}^{t_1} \langle Qu, u - \hat{u} \rangle \, dt.$$

Using the fact that $\hat{x}, \hat{u}$ satisfy (2) and x, u, λ satisfy (4), we get

$$\langle Hx, x \rangle = \langle -\dot{\lambda} - A^*\lambda, x \rangle = -\langle \dot{\lambda}, x \rangle - \langle \lambda, \dot{x} \rangle + \langle Qu, u \rangle$$

and

$$\langle Hx, \hat{x} \rangle = \langle -\dot{\lambda} - A^*\lambda, \hat{x} \rangle = -\langle \dot{\lambda}, \hat{x} \rangle - \langle \lambda, \dot{\hat{x}} \rangle + \langle Qu, \hat{u} \rangle.$$

Thus

$$\begin{aligned} \int_{t_0}^{t_1} \langle Hx, \hat{x} - x \rangle \, dt &= \int_{t_0}^{t_1} \langle \dot{\lambda}, x \rangle + \langle \lambda, \dot{x} \rangle + \langle Qu, u \rangle - \langle \dot{\lambda}, \hat{x} \rangle - \langle \lambda, \dot{\hat{x}} \rangle - \langle Qu, \hat{u} \rangle \, dt \\ &= \langle \lambda, x \rangle \Big|_{t_0}^{t_1} - \langle \lambda, \hat{x} \rangle \Big|_{t_0}^{t_1} + \int_{t_0}^{t_1} \langle Qu, u \rangle - \langle Qu, \hat{u} \rangle \, dt \qquad (6) \\ &= \int_{t_0}^{t_1} \langle Qu, u - \hat{u} \rangle \, dt \quad \text{as desired.} \end{aligned}$$

□

It is interesting to note that Theorem 1 says that solutions of (4) satisfying the boundary conditions provide optimal controls even if the differential equation (4) has nonunique solutions for consistent initial conditions. Of course, in that case the optimal controls may not be unique.

It would be of interest to have a general form for all solutions of (4) in the case of nonunique solutions.

Note that from (6) one immediately gets that

$$J[x, u] = -\tfrac{1}{2}\langle \lambda, x \rangle \Big|_{t_0}^{t_1}. \qquad (7)$$

4. Summary of needed results. To solve (5) we shall need some results from [5] and a few basic facts about the Drazin inverse [8] of a square matrix.

DEFINITION 1. If F is an $r \times r$ matrix with index k (k is the smallest $l \geqq 0$ such that rank $(A^{l+1}) = \text{rank}\,(A^l)$), then the Drazin inverse of F, denoted F^D, is the unique matrix G such that

(i) $GFG = G$,
(ii) $GF = FG$,
(iii) $GF^{k+1} = F^k$.

The Drazin inverse is *not* an equation solving or (1)-inverse. It has recently been shown to have important applications [5], [14]. Basic properties are developed in [3], [4], [9], [10], [14] and [15]. That there is a connection between the Drazin inverse and control theory was realized independently by Dickinson [7]. Our results are quite different from his.

If there exists a scalar μ such that $(\mu C + E)^{-1}$ exists, then $\tilde{C} = (\mu C + E)^{-1}C$ and $\tilde{E} = (\mu C + E)^{-1}E$ commute [5]. Thus if such a μ exists, (1) can be written as

$$\tilde{C}\dot{x} + \tilde{E}x = \tilde{f}, \quad \text{where } \tilde{C}\tilde{E} = \tilde{E}\tilde{C}. \tag{8}$$

For any matrix F, denote its null space by $\mathcal{N}(F)$. Then (8) is consistent and a particular solution is

$$x = \tilde{C}^D e^{-\tilde{C}^D\tilde{E}t}\int_{t_0}^{t} e^{\tilde{C}^D\tilde{E}s}\tilde{f}(s)\,ds + (I - \tilde{C}\tilde{C}^D)\sum_{n=0}^{k-1}(-1)^n(\tilde{C}\tilde{E}^D)^n\tilde{E}^D\tilde{f}^{(n)}. \tag{9}$$

Here k is the index of $\tilde{C}$. Every solution to the homogeneous equation $\tilde{C}\dot{x} + \tilde{E}x = 0$ is of the form

$$e^{-\tilde{C}^D\tilde{E}t}\tilde{C}^D\tilde{C}x_0, \qquad x_0 \text{ an arbitrary vector [5].} \tag{10}$$

Note that if there exists such a μ, then all but a finite number of scalars can be used for μ. For convenience, we shall say a property holds almost always if it holds except for a finite number of scalars. The existence of such a μ is necessary in order to apply the results of [5]. We shall discuss later how this assumption is related to other types of assumptions. Finally we note that

$$\int e^{Ft}\,dt = F^D e^{Ft} + (I - FF^D)t\left[I + \frac{F}{2}t + \cdots + \frac{F^{k-1}}{k!}t^{k-1}\right] + G, \tag{11}$$

where F is square, $k = \text{Index}\,(F)$ and G is an arbitrary square matrix [5].

5. Solution of the system. Rewrite (5) as

$$\mathcal{A}\dot{z} + \mathcal{B}z = 0, \tag{12}$$

where

$$\mathcal{A} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \qquad \mathcal{B} = \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}.$$

Here I is $2n \times 2n$,

$$B_1 = \begin{bmatrix} A^* & H \\ 0 & -A \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ -B \end{bmatrix}, \quad B_3 = [B^* \quad 0] \quad \text{and } B_4 = Q. \tag{13}$$

Clearly $(\mu + B_1)^{-1}$ exists except for a finite number of μ. Define

$$Q_\mu = B_4 - B_3(\mu + B_1)^{-1}B_2. \tag{14}$$

PROPOSITION 1. *$\mu\mathcal{A}+\mathcal{B}$ is invertible almost always if and only if Q_μ is invertible almost always.*

Proof.

$$\begin{bmatrix} B_3(\mu+B_1)^{-1} & -I \\ I & 0 \end{bmatrix}\begin{bmatrix} \mu+B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}=\begin{bmatrix} 0 & -Q_\mu \\ \mu+B_1 & B_2 \end{bmatrix}$$

and

$$\begin{bmatrix} B_3(\mu+B_1)^{-1} & -I \\ I & 0 \end{bmatrix} \quad \text{is invertible.} \quad \square$$

We now need the following easily verified result whose proof we omit.

PROPOSITION 2. *If $\begin{bmatrix} \mu+B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}$ is invertible, then the inverse is*

$$\begin{bmatrix} (\mu+B_1)^{-1}+(\mu+B_1)^{-1}B_2Q_\mu^{-1}B_3(\mu+B_1)^{-1} & -(\mu+B_2)^{-1}B_2Q_\mu^{-1} \\ -Q_\mu^{-1}B_3(\mu+B_1)^{-1} & Q_\mu^{-1} \end{bmatrix} \tag{15}$$

almost always.

Assume that μ, $\mathcal{A}$, $\mathcal{B}$ are such that $\mu\mathcal{A}+\mathcal{B}$, Q_μ, $\mu+B_1$ are invertible. Then

$$(\mu\mathcal{A}+\mathcal{B})^{-1}\mathcal{A}=\begin{bmatrix} (\mu+B_1)^{-1}+(\mu+B_1)^{-1}B_2Q_\mu^{-1}B_3(\mu+B_1)^{-1} & 0 \\ Q_\mu^{-1}B_3(\mu+B_1)^{-1} & 0 \end{bmatrix}. \tag{16}$$

Define N_μ and M_μ by

$$(\mu\mathcal{A}+\mathcal{B})^{-1}\mathcal{A}=\begin{bmatrix} N_\mu & 0 \\ M_\mu & 0 \end{bmatrix}.$$

Also note that

$$(\mu\mathcal{A}+\mathcal{B})^{-1}\mathcal{B} \tag{17}$$

$$=\left[\begin{array}{c|c} (\mu+B_1)^{-1}B_1+(\mu+B_1)^{-1}B_2Q_\mu^{-1}B_3(\mu+B_1)^{-1}B_1 & \begin{array}{c}(\mu+B_1)^{-1}B_2 \\ +(\mu+B_1)^{-1}B_2Q_\mu^{-1}B_3(\mu+B_1)^{-1}B_2\end{array} \\ \begin{array}{c}-(\mu+B_1)^{-1}B_1+Q_\mu^{-1}B_3\end{array} & \begin{array}{c}-(\mu+B_1)^{-1}B_2Q_\mu^{-1}B_4\end{array} \\ \hline -Q_\mu^{-1}B_3(\mu+B_1)^{-1}B_1+Q_\mu^{-1}B_3 & -Q_\mu^{-1}B_3(\mu+B_1)^{-1}B_2+Q_\mu^{-1}B_4 \end{array}\right].$$

Let Z_μ and W_μ denote the (1, 1) and (2, 1) entries of $(\mu\mathcal{A}+\mathcal{B})^{-1}\mathcal{B}$.

From (14) we have

$$(B_4-B_3(\mu+B_1)^{-1}B_2)Q_\mu^{-1}=Q_\mu^{-1}(B_4-B_3(\mu+B_1)^{-1}B_2)=I.$$

Thus (17) becomes

$$(\mu\mathcal{A}+\mathcal{B})^{-1}\mathcal{B}=\begin{bmatrix}Z_\mu & 0\\ W_\mu & I\end{bmatrix}. \tag{18}$$

In order to use (9), we need to be able to calculate the Drazin inverse of (16).

THEOREM 2 (Meyer and Rose [15]). *Suppose that R, T are square matrices with indices k, l, and S an arbitrary matrix of the appropriate size. Then*

$$\begin{pmatrix}R & S\\ 0 & T\end{pmatrix}^D=\begin{pmatrix}R^D & X\\ 0 & T^D\end{pmatrix},$$

where

$$X=R^{D^2}(S+R^DST+R^{D^2}ST^2+\cdots+R^{D^{l-1}}ST^{l-1})(I-TT^D)$$
$$+(I-RR^D)(S+RST^D+R^2ST^{D^2}+\cdots+R^{k-1}ST^{D^{k-1}})T^{D^2}-R^DST^D.$$

Let $\tilde{\mathcal{A}}=(\mu\mathcal{A}+\mathcal{B})^{-1}\mathcal{A}$, $\tilde{\mathcal{B}}=(\mu\mathcal{A}+\mathcal{B})^{-1}\mathcal{B}$. From Theorem 2 and (16), (18), we get that

$$\begin{aligned}\tilde{\mathcal{A}}^D\tilde{\mathcal{A}}&=\begin{bmatrix}N_\mu & 0\\ M_\mu & 0\end{bmatrix}^D\begin{bmatrix}N_\mu & 0\\ M_\mu & 0\end{bmatrix}\\ &=\begin{bmatrix}N_\mu^D & 0\\ M_\mu N_\mu^{D^2} & 0\end{bmatrix}\begin{bmatrix}N_\mu & 0\\ M_\mu & 0\end{bmatrix}=\begin{bmatrix}N_\mu^DN_\mu & 0\\ M_\mu N_\mu^D & 0\end{bmatrix},\end{aligned} \tag{19}$$

while

$$\tilde{\mathcal{A}}^D\tilde{\mathcal{B}}=\begin{bmatrix}N_\mu^D & 0\\ M_\mu N_\mu^{D^2} & 0\end{bmatrix}\begin{bmatrix}Z_\mu & 0\\ W_\mu & I\end{bmatrix}=\begin{bmatrix}N_\mu^DZ_\mu & 0\\ M_\mu N_\mu^{D^2}Z_\mu & 0\end{bmatrix}.$$

To evaluate $e^{-\tilde{\mathcal{A}}^D\tilde{\mathcal{B}}t}$ note that

$$\begin{bmatrix}N_\mu^DZ_\mu & 0\\ M_\mu N_\mu^{D^2}Z_\mu & 0\end{bmatrix}^r=\begin{bmatrix}[N_\mu^DZ_\mu]^r & 0\\ M_\mu N_\mu^{D^2}Z_\mu[N_\mu^DZ_\mu]^{r-1} & 0\end{bmatrix} \tag{20}$$

for integers $r\geqq 1$. From (20) and the power series expansion of the exponential we see that

$$e^{-\tilde{\mathcal{A}}^D\tilde{\mathcal{B}}t}=\begin{bmatrix}e^{-[N_\mu^DZ_\mu]t} & 0\\ M_\mu N_\mu^D\{e^{-[N_\mu^DZ_\mu]t}-I\} & I\end{bmatrix}. \tag{21}$$

Of course, the quantities N_μ, Z_μ, M_μ, W_μ are closely related.

PROPOSITION 3. *If N_μ, Z_μ, M_μ, W_μ are defined by* (16) *and* (17), *then*

$$Z_\mu=1-\mu N_\mu, \tag{22}$$

$$W_\mu=-\mu M_\mu, \tag{23}$$

and

$$N_\mu = (\mu + B_1)^{-1} - (\mu + B_1)^{-1} B_2 M_\mu. \tag{24}$$

Proof. All the equations follow from the definitions of N_μ, Z_μ, W_μ and M_μ. For example,

$$\begin{aligned} Z_\mu &= [(\mu + B_1)^{-1} + (\mu + B_1)^{-1} B_2 Q_\mu^{-1} B_3 (\mu + B_1)^{-1}] B_1 - (\mu + B_1)^{-1} B_2 Q_\mu^{-1} B_3 \\ &= N_\mu B_1 - \{N_\mu - (\mu + B_1)^{-1}\}(\mu + B_1) \\ &= N_\mu B_1 - N_\mu(\mu + B_1) + I = I - \mu N_\mu. \end{aligned}$$

Thus (22) follows. The proof of (23) and (24) are similar. □

We can now substitute into (10). Using (10) we see that the general solution of (12) is

$$\begin{aligned} \Omega &= e^{-\tilde{\mathcal{A}}^D \tilde{\mathcal{B}} t} \tilde{\mathcal{A}}^D \tilde{\mathcal{A}} \\ &= \begin{bmatrix} e^{-[N_\mu^D Z_\mu]t} & 0 \\ M_\mu N_\mu^D \{e^{-[N_\mu^D Z_\mu]t} - I\} & I \end{bmatrix} \begin{bmatrix} N_\mu^D N_\mu & 0 \\ M_\mu N_\mu^D & 0 \end{bmatrix} \\ &= \begin{bmatrix} e^{-[N_\mu^D Z_\mu]t} N_\mu^D N_\mu & 0 \\ M_\mu N_\mu^D \{e^{-[N_\mu^D Z_\mu]t} - I\} + M_\mu N_\mu^D & 0 \end{bmatrix} \\ &= \begin{bmatrix} e^{-[N_\mu^D Z_\mu]t} N_\mu^D N_\mu & 0 \\ M_\mu N_\mu^D \, e^{-[N_\mu^D Z_\mu]t} & 0 \end{bmatrix}. \end{aligned} \tag{25}$$

Referring back to the original equation we see that

$$\begin{bmatrix} \lambda \\ x \end{bmatrix} = e^{-[N_\mu^D Z_\mu](t - t_0)} N_\mu^D N_\mu \begin{bmatrix} \lambda_0 \\ x_0 \end{bmatrix}, \quad \text{where } \lambda_0 = \lambda(t_0), \tag{26}$$

and

$$u = M_\mu N_\mu^D \begin{bmatrix} \lambda \\ x \end{bmatrix}. \tag{27}$$

Thus we have shown

THEOREM 3. *If Q_μ is invertible, then the optimal control u is given in terms of x, λ by* (27) *if an optimal control exists.*

6. Initial conditions. While (9) gives $\begin{bmatrix} \lambda \\ x \end{bmatrix}$ explicitly, (27) does not give u directly in terms of x. We now turn to this problem.

Let

$$E(t) = e^{-[N_\mu^D Z_\mu](t - t_0)} N_\mu^D N_\mu = \begin{bmatrix} E_1(t) & E_2(t) \\ E_3(t) & E_4(t) \end{bmatrix},$$

where the $E_i(t)$, $i = 1, 2, 3, 4$, are all $n \times n$ matrices.

Suppose that (4) has a solution. Let $\lambda(t_0) = \lambda_0$. Then $\begin{bmatrix} \lambda(t) \\ x(t) \end{bmatrix} = E(t) \begin{bmatrix} \lambda_0 \\ x_0 \end{bmatrix}$. Note that this is possible if and only if $\begin{bmatrix} \lambda_0 \\ x_0 \end{bmatrix}$ is in the range of $N_\mu^D N_\mu$, denoted $R(N_\mu^D N_\mu)$.

Now $\begin{bmatrix}\lambda(t_1)\\ \dot{x}(t_1)\end{bmatrix} = E(t_1)\begin{bmatrix}\lambda_0\\ x_0\end{bmatrix}$ or

$$x_1 = E_3(t_1)\lambda_0 + E_4(t_1)x_0. \tag{28}$$

If $E_3(t_1)$ is invertible, then λ_0 is determined by x_0 and x_1 uniquely. Once λ_0, x_0 are known, x, u follow from (26) and (27). On the other hand if (28) is viewed as defining x_1, then from (26) x will go from x_0 to x_1. Thus we have established the following result.

THEOREM 4. *Suppose that Q_μ is invertible almost always. For a given x_0, x_1 there is an optimal control that takes x from x_0 to x_1 in the time interval $[t_0, t_1]$ if and only if the equation* (28) *has a solution λ_0 such that* $\begin{bmatrix}\lambda_0\\ x_0\end{bmatrix} \in R(N_\mu^D N_\mu)$.

7. An example. It is possible, under our assumptions, for x to be able to go from x_0 to x_1 but not have an optimal control existing if $N_\mu^D N_\mu \begin{bmatrix}\lambda_0\\ x_0\end{bmatrix} = \begin{bmatrix}\lambda_0\\ x_0\end{bmatrix}$, λ_0 satisfying (29), is inconsistent in λ_0. We shall give a simple example that illustrates this. It shall also serve to illustrate our method.

Let $H = I$, $B = I$, $A = 0$, $Q = \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$ be two by two matrices. The process is then simply

$$\dot{x} = u, \tag{29}$$

where x, u are 2-vectors; $x = \begin{bmatrix}\phi_1\\ \phi_2\end{bmatrix}$, $u = \begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix}$, and the cost is $\int_{t_0}^{t_1} |\phi_1|^2 + |\phi_2|^2 + |\psi_1|^2\, dt$. The system (5) becomes

$$\begin{bmatrix}I & 0 & 0\\ 0 & I & 0\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\dot{\lambda}\\ x\\ u\end{bmatrix} + \begin{bmatrix}0 & I & 0\\ 0 & 0 & -I\\ I & 0 & Q\end{bmatrix}\begin{bmatrix}\lambda\\ x\\ u\end{bmatrix} = \begin{bmatrix}0\\ 0\\ 0\end{bmatrix}. \tag{30}$$

Since $\mathcal{B}$ is invertible, we may take $\mu = 0$ in $(\mu\mathcal{A} + \mathcal{B})^{-1}$. Now

$$\begin{bmatrix}0 & I & 0\\ 0 & 0 & -I\\ I & 0 & Q\end{bmatrix}^{-1} = \begin{bmatrix}0 & Q & I\\ I & 0 & 0\\ 0 & -I & 0\end{bmatrix}. \tag{31}$$

Multiplying (30) by (31) gives

$$\begin{bmatrix}0 & Q & 0\\ I & 0 & 0\\ 0 & -I & 0\end{bmatrix}\begin{bmatrix}\dot{\lambda}\\ x\\ u\end{bmatrix} + \begin{bmatrix}\lambda\\ x\\ u\end{bmatrix} = 0. \tag{32}$$

By Theorem 2 and the fact that

$$\begin{bmatrix}0 & Q\\ I & 0\end{bmatrix}^{D^2} = \left(\begin{bmatrix}0 & Q\\ I & 0\end{bmatrix}^2\right)^D = \begin{bmatrix}Q & 0\\ 0 & Q\end{bmatrix}^D = \begin{bmatrix}Q & 0\\ 0 & Q\end{bmatrix}$$

it is straightforward to get that

$$\begin{bmatrix} 0 & Q & 0 \\ I & 0 & 0 \\ 0 & -I & 0 \end{bmatrix}^D = \begin{bmatrix} \begin{bmatrix} 0 & Q \\ I & 0 \end{bmatrix}^D & 0 \\ [0 \quad -Q] & 0 \end{bmatrix} = \begin{bmatrix} 0 & Q & 0 \\ Q & 0 & 0 \\ 0 & -Q & 0 \end{bmatrix}.$$

Thus using (10) the solutions to (32) are given by

$$\begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} = e^{-\begin{bmatrix} 0 & Q & 0 \\ Q & 0 & 0 \\ 0 & -Q & 0 \end{bmatrix}(t-t_0)} \begin{bmatrix} Q & 0 & 0 \\ 0 & Q & 0 \\ -Q & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_0 \\ x_0 \\ u_0 \end{bmatrix}. \tag{33}$$

It is clear from (29) that for any x_0, x_1 there exists a control u sending x_0 to x_1; i.e., (29) is completely controllable. But the x in (33) only takes on values of the form $\begin{bmatrix} c \\ 0 \end{bmatrix}$ for scalar c. Thus in order for an optimal control to exist, x_0, x_1 must be of the form $\begin{bmatrix} c_0 \\ 0 \end{bmatrix}, \begin{bmatrix} c_1 \\ 0 \end{bmatrix}$. A look at the power series for the exponential in (33) shows that

$$\begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} = \begin{bmatrix} \cosh(t-t_0)Q+(I-Q) & -\sinh(t-t_0)Q & 0 \\ -\sinh(t-t_0)Q & \cosh(t-t_0)Q+(I-Q) & 0 \\ -\cosh(t-t_0)Q+Q & -\sinh(t-t_0)Q & I \end{bmatrix}$$

$$\cdot \begin{bmatrix} Q & 0 & 0 \\ 0 & Q & 0 \\ -Q & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_0 \\ x_0 \\ u_0 \end{bmatrix}$$

$$= \begin{bmatrix} \cosh(t-t_0)Q & -\sinh(t-t_0)Q & 0 \\ -\sinh(t-t_0)Q & \cosh(t-t_0)Q & 0 \\ -\cosh(t-t_0)Q+Q & -\sinh(t-t_0)Q & 0 \end{bmatrix} \begin{bmatrix} \lambda_0 \\ x_0 \\ u_0 \end{bmatrix}.$$

If $x_0 = \begin{bmatrix} c_0 \\ 0 \end{bmatrix}$ and $x_1 = \begin{bmatrix} c_1 \\ 0 \end{bmatrix}$, we see that $t = t_0$ gives $u = -Q\lambda_0$. Since

$$\begin{bmatrix} \lambda_0 \\ x_0 \\ u_0 \end{bmatrix} \in R\left(\begin{bmatrix} Q & 0 & 0 \\ 0 & Q & 0 \\ -Q & 0 & 0 \end{bmatrix} \right)$$

we must have $\lambda_0 = \begin{bmatrix} l_0 \\ 0 \end{bmatrix}$, and then $u_0 = \begin{bmatrix} -l_0 \\ 0 \end{bmatrix}$. Letting $t = t_1$ gives

$$c_1 = -\sinh(t_1-t_0)l_0 + \cosh(t_1-t_0)c_0. \tag{34}$$

Solving (34) for l_0 we get

$$u = \begin{bmatrix} \cosh(t-t_0)\{c_1 - \cosh(t_1-t_0)c_0\}/\sinh(t_1-t_0) & -\sinh(t-t_0)c_0 \\ 0 \end{bmatrix}$$

as the optimal control. x can also be easily solved for if desired.

In working a given problem, it is sometimes simpler to solve (5) directly using the techniques used in deriving the formulas (26) and (27) as done in this example, rather than try to use the formulas directly.

8. The assumption that Q_μ is invertible. Let us now examine in more detail our basic assumption that Q_μ is invertible. From (13) and (14) we have

$$\begin{aligned} Q_\mu &= B_4 - B_3(\mu + B_1)^{-1}B_2 \\ &= Q - [B^* \quad 0]\begin{bmatrix} \mu + A^* & H \\ 0 & \mu - A \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ -B \end{bmatrix} \\ &= Q + [B^* \quad 0]\begin{bmatrix} (\mu + A^*)^{-1} & -(\mu + A^*)^{-1}H(\mu - A)^{-1} \\ 0 & (\mu - A)^{-1} \end{bmatrix}\begin{bmatrix} 0 \\ B \end{bmatrix} \\ &= Q - B^*(\mu + A^*)^{-1}H(\mu - A)^{-1}B. \end{aligned} \tag{35}$$

PROPOSITION 4. *If Q is invertible, then Q_μ is almost always invertible.*

Proof. Suppose Q is invertible. By Proposition 1 it suffices to show $\mu\mathscr{A} + \mathscr{B}$ is invertible almost always. But

$$[\mu\mathscr{A} + \mathscr{B}]\begin{bmatrix} I & 0 \\ -B_4^{-1}B_3 & I \end{bmatrix} = \begin{bmatrix} \mu + B_1 - B_2B_4^{-1}B_3 & B_2 \\ 0 & B_4 \end{bmatrix}. \tag{36}$$

Thus $\mu\mathscr{A} + \mathscr{B}$ is invertible almost always since the right side of (36) is invertible almost always. □

We note without proof

PROPOSITION 5. *If F, G are positive semi-definite $r \times r$ matrices, then $F + G$ is invertible if and only if*

$$\mathscr{N}(F) \cap \mathscr{N}(G) = \{0\}.$$

Of course, Q_μ is invertible almost always for real μ if and only if it is almost always invertible for complex μ. Let $\mu = i\omega$ where ω is real. Then (35) becomes

$$\begin{aligned} Q_\mu &= Q - B^*(i\omega + A^*)^{-1}H(i\omega - A)^{-1}B \\ &= Q + B^*(-i\omega + A)^{-1^*}H(-i\omega + A)^{-1}B. \end{aligned} \tag{37}$$

From Proposition 5 we have that Q_μ is invertible almost always if and only if

$$\begin{aligned} \{0\} &= \mathscr{N}(Q) \cap \mathscr{N}(B^*(-i\omega + A)^{-1^*}H(-i\omega + A)^{-1}B) \\ &= \mathscr{N}(Q) \cap \mathscr{N}(H^{1/2}(-i\omega + A)^{-1}B) \\ &= \mathscr{N}(Q) \cap \mathscr{N}(H(-i\omega + A)^{-1}B) \quad \text{for almost all } \omega. \end{aligned}$$

Thus we have proven

THEOREM 5. *Q_μ is invertible for almost all μ if and only if*

$$\mathscr{N}(Q) \cap \mathscr{N}(H(-i\omega + A)^{-1}B) = \{0\} \tag{38}$$

for almost every real ω.

If H is invertible (positive definite), then Q_μ is invertible for almost all μ if and only if

$$\mathcal{N}(Q) \cap \mathcal{N}(B) = \{0\}. \tag{39}$$

We note the following necessary, but not sufficient, condition for the invertibility of Q_μ.

PROPOSITION 6. *Q_μ is invertible for almost all μ if*

$$\bigcap_r \mathcal{N}(HA^rB) = \{0\} \quad \text{for } r = 0, 1, \cdots, \text{Index}(A). \tag{40}$$

Proof. For large ω,

$$H(-i\omega + A)^{-1}B = iH \sum_{n=0}^{\infty} (-iA)^n B\omega^{-n-1}.$$

Thus a vector $v \in \mathcal{N}(H(-i\omega + A)^{-1}B)$ for almost all ω if and only if

$$\psi(\omega) = iH \sum_{n=0}^{\infty} (-iA)^n B\omega^{-n-1} v \tag{41}$$

is zero for large ω. But since (41) is a Laurent series, this happens if and only if $HA^nBv = 0$ for all $n \geqq 0$. Equation (41) now follows. □

The invertibility of Q_μ has two intuitive interpretations. Before developing them we need a result on analytic (1)-inverses.

THEOREM 6. *Suppose that $A(\cdot)$ is an $m \times n$ matrix-valued function such that $A_{ij}(z)$ is a fraction of polynomials for all i and j. Suppose also that $\mathcal{N}(A(z))$ is nontrivial for all z in the domain of $A(\cdot)$. Then for any real number $\omega > 0$, there exists an $n \times n$ matrix-valued function $B(\cdot)$ such that:*

(i) *$B_{ij}(z)$ is a fraction of polynomials,*

(ii) *$R(B(z)) = \mathcal{N}(A(z))$ for almost all z,*

(iii) *the poles of B are integral multiples of ωi, $\omega > 0$, are simple, and*

(iv) *$\|B(z)\| = O(1/|z|^3)$ as $|z| \to \infty$.*

Proof. Suppose that $A(\cdot)$ is an $m \times n$ matrix-valued function such that $A_{ij}(z)$ is a fraction of polynomials for all i and j. Suppose also that $\mathcal{N}(A(z))$ is nontrivial for all z in the domain of $A(\cdot)$. Let X be an $n \times m$ matrix of unknowns x_{ij}. Then

$$AXA = A \tag{42}$$

is a consistent linear system of at most mn equations in mn unknowns. Denote this new system by

$$EX = A. \tag{43}$$

Since the coefficients of (43) are fractions of polynomials, there exists a real number K such that all minors of E are identically zero, or identically nonzero, for $|z| \geqq K$. Thus (43) can be solved by row operations (nonuniquely) to give a $F(\cdot)$ such that F satisfies (42) for $|z| \geqq K$, the entries of $F(z)$ are fractions of polynomials in z, rank $F(z)$ is constant, and rank $F(z)$ is the maximum possible ($\dim \mathcal{N}(A(z))$). Note that $(FA)_{ij}$ is a fraction of polynomials for all i and j. Let $z_1, \cdots, z_q$ be the poles of FA. Let $r_1, \cdots, r_q$ denote their multiplicities. Let r_0 be

such that $\|FA\| = O(|z|^{r_0})$ as $|z| \to \infty$. Set $a = r_0 + r_1 + \cdots + r_q + 3$. Define

$$B(z) = \prod_{j=1}^{q} (z - z_j)^{r_j} \prod_{p=1}^{a} (z - ip\omega)^{-1}(I - F(z)A(z)).$$

Then B clearly satisfies (i), (iii) and (iv). Since (ii) holds for $|z| \geqq K$, it holds for almost all z by analytic continuation. □

We can now give a "physical" interpretation of the invertibility of Q_μ.

THEOREM 7. *The following are equivalent*:

(a) *There exists an* x_0, x_1 *for which optimal controls exist, but are not unique.*

(b) *There is a trajectory from zero to zero of zero cost with nonzero control.*

(c) Q_μ *is not invertible for all* μ.

Proof. Clearly (b) $\Rightarrow$ (a) since $J[0, 0] = 0$. To see that (a) $\Rightarrow$ (b), let (x, u), $(\hat{x}, \hat{u})$ be two optimal solutions from x_0 to x_1. Then there exists λ, $\hat{\lambda}$ so that (λ, x, u) and $(\hat{\lambda}, \hat{x}, \hat{u})$ satisfy (4). Thus $(\lambda - \hat{\lambda}, x - \hat{x}, u - \hat{u})$ satisfies (4) and hence is optimal. But $(x - \hat{x})(t_0) = (x - \hat{x})(t_1) = 0$ and $u - \hat{u}$ is not identically zero. That $J[x - \hat{x}, u - \hat{u}] = 0$ follows from (7).

Suppose now that (b) holds so that there exists x, u such that $J[x, u] = 0$, $x(t_0) = 0$, $x(t_1) = 0$, and u is nonzero. Since $J[x, u] = 0$ it is clear from (2) that $Hx = 0$ and $Qu = 0$. Extend x, μ periodically to $[-\infty, \infty]$ and replace t by $t - t_0$. Call the new functions $\tilde{x}, \tilde{u}$. Thus $H\tilde{x} = 0$, $Q\tilde{u} = 0$, and $\dot{\tilde{x}} = A\tilde{x} + B\tilde{u}$, $t \neq n(t_1 - t_0)$, $n = 0, \pm 1, \pm 2, \cdots$. Since $\tilde{u}$ is bounded and sectionally continuous on finite intervals, $\tilde{x}$ is continuous, and $\tilde{x}$ is of exponential order, we can take Laplace transforms to get $H\mathscr{L}[\tilde{x}] = 0$, $Q\mathscr{L}[\tilde{u}] = 0$ and $\mathscr{L}[\tilde{x}] = (s - A)^{-1}B\mathscr{L}[\tilde{u}]$. Thus $\mathscr{L}[\tilde{u}](s) \in \mathscr{N}(Q) \cap \mathscr{N}(N(s - A)^{-1}B)$ for all s in some right half plane. By (39), we have Q_μ is not invertible for all μ.

Conversely, suppose that Q_μ is not invertible for all μ. Note that $\mathscr{N}(Q) \cap \mathscr{N}(H(\mu - A)^{-1}B) \subseteq \mathscr{N}(Q_\mu)$ for almost all μ. However,

$$\mathscr{N}(Q) \cap \mathscr{N}(H(\mu - A)^{-1}B) = \mathscr{N}(Q_\mu) \tag{44}$$

for $\mu = it$, t real. Thus $\mathscr{N}(Q) \cap \mathscr{N}(H(\mu - A)^{-1}B) = \mathscr{N}(Q_\mu)$ for almost all μ. Now applying Theorem 6 to Q_μ with $\omega = 2\pi/(t_1 - t_0)$ yields a B_μ such that $Q_\mu B_\mu = 0$, and B_μ satisfies (iii), (iv). But (44) then gives us that

$$QB_\mu = 0, \quad \text{and} \quad H(\mu - A)^{-1}BB_\mu = 0. \tag{45}$$

Let ϕ be vector such that $B_\mu\phi$ is not identically zero. Denote $B_\mu\phi$ by $\phi(\mu)$. Let $\tilde{x}(s) = (s - A)^{-1}B\phi(s)$. Then we have from (45) that

$$H\tilde{x}(s) = 0, \quad Q\phi(s) = 0 \quad \text{and} \quad \tilde{x}(s) = (s - A)^{-1}B\phi(s). \tag{46}$$

Let $\hat{x}$ be the inverse Laplace transform of $\tilde{x}$, $\hat{u}$ the inverse Laplace transform of ϕ. From (46) and (iv) we have $H\hat{x} = 0$, $Q\hat{u} = 0$, $\dot{x} = A\hat{x} + B\hat{u}$, $\hat{x}(0) = 0$, and $\hat{u}(0) = 0$ [6, p. 184]. Furthermore, $\hat{u}$ is nonzero. Finally, since the poles of $\phi(s)$ were simple and multiples of $2\pi i/(t_1 - t_0)$ we get that $\hat{x}$, $\hat{u}$ are periodic with period $(t_1 - t_0)$ [6, p. 188]. Replace $\hat{x}$, $\hat{u}$ by $x = \hat{x}(t + t_0)$, $u = \hat{u}(t + t_0)$. Then $x(t_0) = x(t_1) = 0$, $J[x, u] = 0$, and $\dot{x} = Ax + Bu$. Thus (c) $\Rightarrow$ (b). □

It is possible to have Q_μ invertible almost always and still have nonzero optimal trajectories of zero cost. Of course, the control u must then be zero.

Example. Let $Q = I$, $A = I$, $B = 0$, $H = 0$ in (2) and (3). Then Q_μ is invertible for large μ since Q is. Clearly $x = \exp(A(t-t_0))x_0$ is a trajectory of zero cost from x_0 to $x_1 = \exp(A(t-t_0))x_0$. But $u = 0$ and $J[x, u] = 0$. Note also if $x_0 = 0$, then $x \equiv 0$.

Our Theorem 6 is quite possibly contained in the literature on generalized inverses of meromorphic operator-valued functions. However, it is easier to prove what we need directly, than have the reader sort through several pages of notation. Also, we needed to have polynomial growth at infinity, a point not considered in papers such as [2]. The interested reader is referred to [2] both for the state of the art and for a bibliography.

It should be noted that the invertibility of Q_μ is logically independent of the controllability of (2) since for any choice of A, B, setting $Q = I$ makes Q_μ invertible almost always, while setting $Q = H = 0$ makes $Q_\mu \equiv 0$.

Note also that in the example of § 7, the pair (A, B) was completely controllable and Q_μ was invertible. However, optimal controls only existed for certain pairs x_0, x_1. Thus the assumption of controllability does not seem to simplify matters if Q, H are allowed to be singular.

9. Conclusion. We have arrived then at the following procedure for solving the original problem. Given x_0, x_1 determine whether it is possible to go from x_0 to x_1 with an optimal control by solving (if possible) (28) for λ_0 such that $\begin{bmatrix}\lambda_0\\ x_0\end{bmatrix} \in R(N_\mu^D N_\mu)$. If λ_0 is found, use the bottom half of (26) for x if x is needed. Use (26) and (27) to get the optimal control u.

It is interesting to note that for many results the assumption that Q_μ is invertible can be used in place of the assumptions on controllability and invertibility of H and Q. Note also that we have not assumed the invertibility of either H or Q nor A or B.

The results of [5] can be applied, of course, to any problem which leads to a system of the form (1). However, the special form of the $\mathscr{A}$ given in (12) makes most of the calculations of this paper possible since it allowed us to use Theorem 2. The Drazin inverse for a general 2×2 block matrix is very messy. Even if the lower right block is 1×1, the formulas are complicated [19]. Any problem which leads to a system of the form $\mathscr{A}\dot{z} + \mathscr{B}z = f$ with $A = \begin{bmatrix}A_1 & 0\\ A_2 & 0\end{bmatrix}$ can be solved much as was (12), provided, of course, $\mu\mathscr{A} + \mathscr{B}$ is invertible for some μ. We shall now describe several such problems. Since the calculation of the solutions parallels those done earlier, a description of the problem will suffice.

For example, suppose that the cost is given by

$$\int_{t_0}^{t_1} \langle Hx, x\rangle + \langle Qu, u\rangle + \langle x, a\rangle \, dt,$$

where a is a vector. Then the right-hand side of (5) has $\alpha = \begin{bmatrix}a\\ 0\\ 0\end{bmatrix}$ instead of the zero vector.

Equation (9) can be used to solve this nonhomogeneous system to get

$$\Omega = \tilde{\mathcal{A}}^D e^{-\tilde{\mathcal{A}}^D\tilde{\mathcal{B}}t}\int_{t_0}^{t} e^{\tilde{\mathcal{A}}^D\tilde{\mathcal{B}}s}\tilde{\alpha}\,ds + (I-\tilde{\mathcal{A}}\tilde{\mathcal{A}}^D)\tilde{\mathcal{B}}^D\tilde{\alpha} + e^{-\tilde{\mathcal{A}}^D\tilde{\mathcal{B}}t}\tilde{\mathcal{A}}^D\tilde{\mathcal{A}} \tag{47}$$

instead of (25). The integral in (47) can be evaluated by using (11). For this problem, it is important to know whether or not the cost is positive.

Another variation on the same type of problem is process (2) with the cost functional

$$J[x, u] = \int_{t_0}^{t_1} \langle Hx, x\rangle + 2\langle u, Cx\rangle + \langle Qu, u\rangle\, dt,$$

where $\begin{bmatrix} H & C \\ C^* & Q \end{bmatrix}$ is positive semi-definite [1, pp. 461–463]. In this case the system to be solved is

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{\lambda} \\ x \\ u \end{bmatrix} + \begin{bmatrix} A^* & H & C \\ 0 & -A & -B \\ B^* & C & Q \end{bmatrix}\begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

instead of (5). Solution proceeds almost exactly as when $C = 0$, though Q_μ has a slightly different form.

The analysis developed here can be also applied with little change to the following control problem.

Given output y, state vector x and process $\dot{x} = Ax + Bu$, find a control u such that $y = Cx + Du$.

This control problem may be rewritten as

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{x} \\ u \end{bmatrix} + \begin{bmatrix} -A & -B \\ C & D \end{bmatrix}\begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}. \tag{48}$$

If y and u are the same size vectors, then (48) is the nonhomogeneous form of equation (12). It may be solved, under the assumption that Q_μ given in (14) is invertible, by using (9). Here $Q_\mu = D + C(\mu - A)^{-1}B$.

Generalized inverses have been used in [12] to solve (48). However, the assumptions of [12] are of a different nature than ours. Techniques for calculating Drazin inverses may be found in [5], [9], [10], [13], [14], [15], [17], [18] and [19]. The techniques in [13], [14] and [18] are concerned with the index one case. A sequential algorithm is given in [10]. A method based on the eigenvalues of the given matrix is developed in [5]. Formulas for finding the Drazin inverse or index of block triangular matrices are in [15].

It would be desirable to be able to handle (1) when C, E are $m \times n$ instead of square. The results of [5] do not directly apply and the problem of calculating all solutions explicitly appears to be more difficult. There is a tendency for manipulations to cause solutions to be lost or nonsolutions to be introduced. A characterization of consistent initial conditions may be found in [21]. It is a hard characterization to work with, however.

Acknowledgment. Finally, we would like to acknowledge the benefits of several discussions concerning this work with Nicholas J. Rose.

REFERENCES

[1] M. ATHANS AND P. L. FALB, *Optimal Control*, McGraw-Hill, New York, 1966.

[2] H. BART, M. A. KAASHOEK AND D. C. LAY, *Relative inverses of meromorphic operator functions and associated holomorphic projection functions*, Math. Ann., to appear.

[3] S. L. CAMPBELL, *Differentiation of the Drazin inverse*, SIAM J. Appl. Math., 30 (1976), pp. 703–707.

[4] S. L. CAMPBELL AND C. D. MEYER, JR., *Continuity properties of the Drazin pseudoinverse*, Linear Alg. and Appl., 10 (1975), pp. 77–83.

[5] S. L. CAMPBELL, C. D. MEYER, JR. AND N. J. ROSE, *Applications of the Drazin inverse to linear systems of differential equations*, SIAM J. Appl. Math., 31 (1976), to appear.

[6] R. V. CHURCHILL, *Operational Mathematics*, McGraw-Hill, New York, 1958.

[7] B. W. DICKINSON, *Matricial indices and controllability of linear differential systems*, SIAM J. Appl. Math., 25 (1973), pp. 613–617.

[8] M. P. DRAZIN, *Pseudoinverses in associative rings and semigroups*, Amer. Math. Monthly, 65 (1968), pp. 506–514.

[9] T. N. E. GREVILLE, *Spectral generalized inverses of square matrices*, M.R.C. Tech. Sum. Rep. 823, Mathematics Research Center, University of Wisconsin, Madison, 1967.

[10] ———, *The Souriau–Frame algorithm and the Drazin pseudoinverse*, Linear Alg. and Appl., 6 (1973), pp. 205–208.

[11] D. H. JACOBSON, *Totally singular quadratic minimization problems*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 651–657.

[12] V. LOVASS-NAGY AND D. L. POWERS, *On output feedback control*, Internat. J. Control, 1975, to appear.

[13] C. D. MEYER, JR., *Limits and the index of a square matrix*, SIAM J. Appl. Math., 26 (1974), pp. 469–478.

[14] ———, *The role of the group generalized inverse in the theory of finite Markov chains*, SIAM Rev., 17 (1975), pp. 443–464.

[15] C. D. MEYER, JR. AND N. J. ROSE, *The index and the Drazin inverse of block triangular matrices*, SIAM J. Appl. Math., to appear.

[16] B. P. MOLINARI, *The stabilizing solution of the algebraic Riccati equation*, this Journal, 11 (1973), pp. 262–271.

[17] PIERRE ROBERT, *On the group inverse of a linear transformation*, J. Math. Anal. Appl., 22 (1968), pp. 658–669.

[18] N. J. ROSE, *A note on computing the Drazin inverse*, SIAM J. Appl. Math., to appear.

[19] J. M. SHOAF, *The Drazin inverse of a rank-one modification of a square matrix*, Ph.D. dissertation, North Carolina State University, Raleigh, 1975.

[20] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 621–634.

[21] K. T. WONG, *The eigenvalue problem $\lambda Tx + Sx$*, J. Differential Equations, 16 (1974), pp. 270–280.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 6, November 1976

FREQUENCY DOMAIN STABILITY CRITERIA FOR STOCHASTIC NONLINEAR FEEDBACK SYSTEMS*

GILMER L. BLANKENSHIP†

Abstract. This research is concerned with the asymptotic properties of feedback systems containing random parameters and subjected to stochastic perturbations. For the special class of feedback systems formed by the open loop cascade of a multiplicative white noise, a sector nonlinearity and a convolution operator, conditions are given to insure the stability in the mean square sense of the feedback system. These conditions are expressed in terms of the Fourier transform of the convolution kernel, the sector parameters of the nonlinearity, and the mean and variance parameters of the noise. Their form is reminiscent of the familiar Nyquist criterion and the circle theorem for deterministic systems. The approach adopted is functional analytic in flavor and avoids the use of Markov semigroup techniques and auxiliary Lyapunov functionals.

1. Introduction. Within the past decade a number of papers have appeared which develop an input-output approach to the stability problem in dynamical feedback systems. In the original papers of Zames [1] and Sandberg [2] (see also the books [3], [4]) the stability question is considered in the framework of functional analysis and the traditional Lyapunov theory is avoided. Considerable simplification of the definitions and basic criteria for stability is gained by this alternate approach. In essence, a feedback system is stable in a functional context if it may be represented by a bounded operator on a specified function space. Since the output of a feedback system is defined implicitly in terms of the system input, stability is equivalent to the existence of a bounded, causal inverse for the system operator [4]. Stability criteria may then be developed as an application of the mathematical theory relating to the invertibility of operators. Fundamental criteria for stability in the input-output theory require the operators in the open loop system to form a contraction or, when the function space of interest is a Hilbert space, to be positive (dissipative).

One of the most interesting results of this theory is the circle theorem [1], [2]. Consider the deterministic nonlinear integral equation

$$x(t) = u(t) - \int_0^t g(t-s)f(s, x(s))\, ds \tag{1.1}$$

as representing a feedback system composed of a linear convolution with kernel g and a nonlinear, memoryless feedback gain f. The function u is considered as the input, and the properties of x are in question. The system represented by (1.1) is said to be L_∞-stable if $u \in L_\infty(R^+)$ (the set of real-valued measurable functions essentially bounded on $R^+ \triangleq [0, \infty)$) implies $x \in L_\infty(R^+)$ and $\|x\|_{L_\infty} \leqq \beta \|u\|_{L_\infty}$ for some constant $\beta \in R^+$ independent of u. The function $\|\cdot\|_{L_p}$ is the norm on $L_p(R^+)$.

* Received by the editors July 18, 1975, and in revised form February 17, 1976.

† Systems Research Center, Case Western Reserve University, Cleveland, Ohio 44106. This research was supported in part by the National Science Foundation under Grant ENG75-08613.

The criterion is as follows:

THEOREM 1.1 (Circle criterion L_∞-version [5, Thm. 2]).
Assume that the following are true for (1.1):

(i) $u \in L_\infty(R^+)$.

(ii) *There exist constants* $a, b \in R^+$ *such that*

$$0 < a < f(t, y)/y < b$$

for every $t \in R^+$, $y \in R$.

(iii) *There exists a positive constant* r_0 *such that*

$$\int_0^\infty e^{r_0 t}|g(t)|\, dt < \infty.$$

(iv) *For* $G(s) \triangleq \int_0^\infty e^{-st} g(t)\, dt$ *and some* $r \in (0, r_0)$, *the conditions below hold*:

(a) $$(-2(a+b)^{-1}, j0) \notin \bigcup_{\mathrm{Re}(s) \geqq -r} \{G(s)\},$$

(b) $$\inf_{\alpha \in R} |G^{-1}(-r + j\alpha) + \tfrac{1}{2}(a+b)| > \tfrac{1}{2}(b-a).$$

Then $x \in L_\infty(R^+)$ *and there exists a constant* $\beta \in R^+$ *independent of* u *such that* $\|x\|_{L_\infty} \leqq \beta \|u\|_{L_\infty}$.

The decisive condition (iv) is equivalent to the statement: For some $r \in (0, r_0)$ the r-shifted Nyquist locus $\bigcup_{\mathrm{Re}(s)=-r}\{G(s)\}$ does not encircle (iv-a) or intersect (iv-b) the closed disc centered at $(-\frac{1}{2}(a^{-1}+b^{-1}), j0)$ with radius $\frac{1}{2}(a^{-1}-b^{-1})$. If L_p-stability for $p < \infty$ is of interest, then it is not necessary to use shifted Nyquist loci; however, the L_∞-version of the circle is used below and it is stated here for convenience. Notice that if $b = a$ (the feedback gain is a linear constant), then the circle theorem reduces to the Nyquist criterion (except for the shifts) and becomes necessary for stability as well.

The purpose of this paper is to consider a generalization of (1.1) obtained by replacing the deterministic function f with a stochastic function and permitting the disturbance u to be a stochastic process. Specifically, we examine the stability problem associated with stochastic integral equations of the form

$$x(t) = u(t) - \int_0^t g(t-s) f[s, x(s)]\, dl(s). \tag{1.2}$$

Here g and f are deterministic functions, u is a stochastic process whose properties are known, and l is a Levy process (roughly the sum of a Wiener and a Poisson process). The functions g and f are deterministic functions of their arguments. The signal x is a stochastic process and the asymptotic properties of $x(t)$ as t goes to infinity are at issue. Actually a somewhat more general version of (1.2) is treated in this paper; however, (1.2) suffices for a discussion of the ideas to be presented. A precise problem statement and detailed discussion of the stochastic integral in (1.2) are deferred until §§ 2 and 3.

Equations of this kind arise in the study of physical systems with unknown parameters and as models of control systems with humans in the feedback path [6], [7]. Equations similar to (1.2) may be used to model the evolution of round-off errors in a computational algorithm [8] or the travel of tsunami over an uneven bottom topography [9]. In each of these applications, stability, with a suitably relaxed interpretation of the concept, is an important issue. For example, in feedback systems, convergence to zero of almost all the sample paths of the process x is usually desired, whereas in the study of round-off errors, estimates of the magnitude of the moments of the errors are useful.

The present criteria establish bounds on the first and second moments of the process x in terms of the moments of u and of parameters related to f, g, and l. The criteria are expressed in terms of the Fourier transform of g, bounds on the gain of nonlinearity f, and the mean and variance parameters of the noise l. Their form is reminiscent of the familiar Nyquist criterion and the circle theorem for deterministic feedback systems. When the system is linear and time-invariant and the noise process l is stationary, necessary and sufficient stability conditions are given which make the differences between the deterministic and stochastic problems most apparent. This paper may be regarded as a generalization of [10] where only linear systems with multiplicative Gaussian white noise were considered and bounds on the first and second moments of x established. Two papers related to earlier drafts of this paper are [11], [12]. They contain some sharpening of the results here in overlapping regions. A listing of the results of this paper appeared in [13].

2. The main results. In this section the main theorems of the paper are summarized and discussed. Their proofs are in § 4. A number of technical preliminaries needed in the proofs are collected in § 3. The next paragraph gives the minimum of definitions and notations needed to state the results.

2.1. Definitions and notation. Let $R^+ = [0, \infty)$ be the time set; systems defined in continuous time are being considered here. The triple $(\Omega, \mathcal{F}, P)$ is a Borel probability space; i.e., Ω is a topological space, $\mathcal{F}$ a Borel σ-algebra of subsets of Ω, and P is a probability measure on $\mathcal{F}$.

Let $C(R^+; R)$ denote the set of continuous functions mapping R^+ into R. Let $D(R^+; R)$ be the set of functions mapping R^+ into R continuous from the right and possessing left-hand limits at every point. Elements of $D(R^+; R)$ are bounded on compact intervals, and for any $\varepsilon > 0$ have at most a finite number of jumps of amplitude greater than ε on any bounded interval in R^+ [14].

For $(\Omega, \mathcal{F}, P)$ a probability space and $(X, \mathcal{B}(X), d)$ a complete, separable, Borel measurable metric space, let $RV(\Omega; X)$ denote the set of X-valued random variables on Ω (equivalent to the set of measurable functions r; $\Omega \to X$). Each element r of $RV(\Omega; X)$ induces a natural probability measure on $\mathcal{B}(X)$ via

$$\mu_r(A) = P\{r^{-1}(A)\}, \qquad A \in \mathcal{B}(X).$$

2.2 Problem statement. Let $(\Omega, \mathcal{F}, P)$ be a probability space to be identified later. Consider the nonlinear stochastic integral equation

$$x(t,\omega) = u(t,\omega) - \int_0^t g(t-s) f(s, x(s,\omega))\, dw(s,\omega) - \int_0^t g(t-s) \int_{-\infty}^{\infty} h(s, x(s,\omega), y)\nu(\omega, ds, dy), \qquad \omega \in \Omega, \quad t \in R^2. \tag{2.1}$$

Here w is an R-valued Wiener process on R^+ with parameters $m \geqq 0$ and σ^2. The random measure ν is defined below as Poisson (on $R^+ \times R$) with parameter $\Pi(t, dy)\, dt$. The process u is a nonanticipating (with respect to w and ν) random process and may be regarded as the input to a feedback system characterized by g, f, h, w, and ν. Here g is the kernel of a linear deterministic convolution which represents the "plant". The nonlinear functions f and h represent nonlinear feedback elements with stochastic components w and ν respectively. Thus, (2.1) may be used to represent a nonlinear feedback system with two random elements, one acting continuously (w) and the (ν) representing the effect of random "shock phenomena" in the feedback channel.

The input u is said to be an admissible nonanticipating input if $\mathcal{B}_{0t}(u) \vee \mathcal{B}_{0t}(w) \vee \mathcal{B}_{0t}(\nu)$ is independent of $\mathcal{B}_{t\infty}(dw) \vee \mathcal{B}_{t\infty}(\nu(ds))$. Here $\vee$ denotes lattice sum, $\mathcal{B} \vee \mathcal{S}$ is the least σ-algebra containing $\mathcal{B}$ and $\mathcal{S}$, and $\mathcal{B}_{st}(f)$ denotes the least Borel σ-algebra over which the random variables $f(v)$, $v \in [s, t]$ are measureable.

Problem to be considered. For all admissible, nonanticipating input processes u with bounded second moments ($\sup_{t \geqq 0} Eu^2(t) < \infty$) find conditions on $g, f, h, m, \sigma^2, \Pi$ so that the solution x of (2.1) is nonanticipating and satisfies

$$\sup_{t \geqq 0} Ex^2(t) \leqq M \sup_{t \geqq 0} Eu^2(t)$$

for some M independent of u.

Remarks. (a) The kernel $g(\cdot)$ in (2.1) need not have a rational Laplace transform for the results obtained below to apply. Thus, the results obtained "extend" those available using Lyapunov theory which essentially requires that the transform of g be rational (so that a representation of (2.1) in terms of an Itô differential equation is possible).

(b) The restriction of attention to "mean square stability" is perhaps not as severe as it might seem. If g is rational and f and h are linear, then the stability properties of all the moments of the solution may be analyzed (see [15]); however, the techniques required for that investigation are very different from the present one. The stability of the higher order moments for the nonlinear equation will be considered elsewhere. In most engineering applications stability of (almost all) the solution trajectories, or "almost sure stability," is probably the most reasonable requirement. However, known criteria[1] to guarantee almost sure stability are very

[1] See, for example, the general conditions in [16] and the example based on these conditions in [17]-[20].

difficult to evaluate in specific (linear) cases. For nonlinear equations comparably sharp conditions for almost sure stability are not known. Hence, the results presented here should be considered as preliminary to the determination of such conditions. It is known that in some cases mean square stability implies almost sure stability [21]; however, criteria for the former may be very conservative guarantees of the latter [20, p. 588].

2.3 Theorems. In this section solutions to the problem posed are given for various assumptions on the functions g, f, h and the noises w and ν. The first theorem to follow gives sufficient conditions (satisfied in every case by the criteria below) for the existence of a unique solution in the space $RV(\Omega; D(R^+; R))$ for (2.1).

THEOREM 2.1 [22, § 3.3]. *Assume that the functions u, g, f, h satisfy the following conditions*:

(i) *u is an admissible nonanticipating element of $RV(\Omega; D(R^+; R))$ such that*

$$E\{u^2(t)\} < \infty \qquad \text{for } t \in [0, T] \subset R^+.$$

(ii) *There exists a $K < \infty$ such that for all $t \in R^+$,*

$$\sigma^2 \int_0^t |g(t-s)|^2 |f(s,x)-f(s,y)|^2\, ds + \int_0^t |g(t-s)|^2 \int_{-\infty}^{\infty} |h(s,x,z) - h(s,y,z)|^2 \Pi(s, dz)\, ds \leqq K|x-y|^2, \qquad x, y \in R.$$

(iii) *There exists a $K < \infty$ such that for all $t \in R^+$,*

$$\int_0^t |g(t-s)| \int_{-\infty}^{\infty} |h(s,x,y)| \Pi(s, dy)\, ds < K(1+|x|), \qquad x \in R.$$

Then a solution x of the integral equation (1.1) *exists in $D(R^+; R)$. Moreover, if $\sup_{0 \leqq t \leqq T} E\{u^2(t)\} < \infty$, then $\sup_{0 \leqq t \leqq T} E\{x^2(t)\} < \infty$ for any $T \in R^+$. The solution x is nonanticipating and unique at all points of continuity.*

In the sequel the conditions on f, g, h, and Π indicated in Theorem 2.1 will be assumed to be satisfied. In all cases they will be superseded by the conditions given for other properties to hold.

Before proceeding to the analysis of the nonlinear equation (2.1), consider the linear case (corresponding to f and h linear and time-invariant):

$$(2.2)\quad x(t) = u(t) - \int_0^t g(t-s)x(s)\, dw(s) - \int_0^t g(t-s)x(s) \int_{-\infty}^{\infty} h(y)\nu(ds, dy),$$

where

$$Edw(t) = m\, dt,$$
$$E(dw(t) - m\, dt)^2 = \sigma^2\, dt,$$
$$E\nu(dt, A) = \Pi(A)\, dt,$$
$$E(\nu(dt, A) - \Pi(A)\, dt)^2 = \Pi(A)\, dt.$$

Assume that u, w, and ν are independent processes.

If Theorem 1.1 holds, then

(2.3)
$$Ex(t) = Eu(t) - \int_0^t g(t-s)Ex(s)m\,ds - \int_0^t g(t-s)Ex(s)\int_{-\infty}^{\infty} h(y)\Pi(dy)\,dt.$$

This is a deterministic equation, to which the Nyquist criterion applies.

THEOREM 2.2. *Assume that the kernel $g \in L_2(R^+)$ and let $G(s)$ denote the Laplace transform of g,*

$$G(s) = \text{l.i.m.}_{T\to\infty} \int_0^T g(t)\,e^{-st}\,dt, \qquad \text{Re}(s) \geqq 0.$$

Then $\sup_{t\in R^+} |Eu(t)| < \infty$ implies $\sup_{t\in R^+} |Ex(t)| < \infty$ if and only if

$$(-(m+\hat{\pi})^{-1}, j0) \notin \bigcup_{\text{Re}(s)\in R^+} \{G(s)\},$$

where $\hat{\pi} = \int_{-\infty}^{\infty} h(y)\pi(dy)$.

Proof. The proofs of this and the following results are contained in § 4.

Now consider the problem of bounding the second moment of x. A transformation of (2.2) gives

(2.4)
$$\begin{aligned} x(t) = u(t) &- \int_0^t g(t-s)x(s)[m+\hat{\pi}]\,ds \\ &- \int_0^t g(t-s)x(s)\,d\tilde{w}(s) - \int_0^t g(t-s)x(s)\int_{-\infty}^{\infty} h(y)\tilde{\nu}(ds, dy), \end{aligned}$$

where $d\tilde{w}(s) = dw(s) - m\,ds$ and $\tilde{\nu}(ds, dy) = \nu(ds, dy) - \Pi(dy)\,ds$.

THEOREM 2.3. *Assume that the kernel $g \in L_2(R^+)$. Then $\sup_{t\in R^+} Eu(t)^2 < \infty$ implies $\sup_{t\in R^+} Ex(t)^2 < \infty$ if and only if*

(i) $(-(m+\hat{\pi})^{-1}, j0) \notin \bigcup_{\text{Re}(s)\in R^+} \{G(s)\}$

(ii) $$\int_{-\infty}^{\infty} \left| \frac{G(j\alpha)}{1+(\hat{\pi}+m)G(j\alpha)} \right|^2 d\alpha < 2\pi(\hat{\pi}+\sigma^2)^{-1}.$$

(Here $\pi = 3.14\cdots$ and $\hat{\pi}$ is defined in Theorem 2.2.)

Two sufficient conditions were proved in [6] for a special case of (2.2) (corresponding to $\nu \equiv 0$); these may be modified to apply in this case, and they yield conditions more easily checked for a given kernel g than the criterion (ii) of Theorem 2.3.

COROLLARY 2.4. *Assume that $g \in L_1(R^+)$ and denote by $\tilde{g}$ the function whose Fourier transform is $G(j\alpha)[1+(m+\hat{\pi})G(j\alpha)]^{-1}$. Assume $\sup_{t\in R^+} Eu(t)^2 < \infty$. Then $\sup_{t\in R^+} Ex(t)^2 \leqq \beta \sup_{t\in R^+} Eu(t)^2$ for some $\beta \in R^+$ if there exists a $\gamma \in R$ such*

that

(i) $(-(m+\hat{\pi})^{-1}, j0) \notin \bigcup_{\mathrm{Re}(s)\in R^+} \{G(s)\}$,

(ii) $\dfrac{\sigma^2+\hat{\pi}}{m+\hat{\pi}+\gamma}\dfrac{\tilde{g}(0)}{2}<1$,

(iii) *and any of the following conditions is satisfied*:

(a) $(m+\hat{\pi})/\gamma>0$, *and the Nyquist locus* $\bigcup_{\alpha\in R}\{G(j\alpha)\}$ *lies inside the circle centered on the real axis of the complex plane at* $(\frac{1}{2}\gamma^{-1}, j0)$ *and passing through the origin.*

(b) $-1<(m+\hat{\pi})/\gamma<0$, *and the Nyquist locus* $\bigcup_{\alpha\in R}\{G(j\alpha)\}$ *lies inside the circle centered on the real axis at* $(\frac{1}{2}\gamma^{-1}, j0)$ *and passing through the origin.*

(c) $(m+\hat{\pi})/\gamma<-1$, *and the Nyquist locus* $\bigcup_{\alpha\in R}\{G(j\alpha)\}$ *does not intersect or encircle the disc centered at* $(\frac{1}{2}\gamma^{-1}, j0)$ *passing through the origin.*[2]

The next result is a special case of Corollary 2.4 as $\gamma\to 0$.

COROLLARY 2.5. *Assume that* $g\in L_1(R^+)$. *Then for* (2.4), $\sup_{t\in R^+}Ex(t)^2 \leqq \beta\sup_{t\in R^+}Eu(t)^2$ *for some* $\beta\in R^+$ *if*

(i) $m+\hat{\pi}>(\sigma^2+\hat{\pi})\tilde{g}(0)/2$,

(ii) $\mathrm{Re}\ G(j\alpha)\geqq 0$ *for all* $\alpha\in R$.

Returning to the analysis of the nonlinear equation (2.1), assume that

$$Edw(t)=0,$$
$$Edw(t)^2=\sigma^2 dt,$$
$$E\nu(dt, dy)=\Pi(dy)\,dt,$$
$$E(\nu(dt, dy)-\Pi(dy)\,dt)^2=\Pi(dy)\,dt,$$

and that there exist constants a, b, c, d such that

$$0<a<f(t, x)/x<b<\infty, \qquad t\in R^+, \quad x\in R,$$
$$0<c<h(t, x, y)/x<d<\infty, \qquad t\in R^+ \quad x, y\in R.$$

Moreover, assume for simplicity that $Eu(t)=0$ for all t and that u, w, and ν are independent processes. The next result establishes criteria sufficient to guarantee a bound on $Ex^2(t)$.

THEOREM 2.6. *For* (2.1) *subject to the assumptions of the last paragraph,* $\sup_{t\in R^+}Ex^2(t)\leqq\beta\sup_{t\in R^+}Eu^2(t)$ *for some* $\beta\in R^+$ *if*:

(i) *There exists an* $r_0>0$ *such that*

$$\int_0^\infty \exp(r_0 t)|g(t)|\,dt<\infty.$$

(ii) $\tilde{\pi}=\int_{-\infty}^{\infty}\Pi(dy)<\infty$.

[2] This particular result has been improved by J. L. Willems in a paper [11] related to an earlier draft of the present paper.

(iii) $([-\tilde{\pi}(c+d)/2]^{-1}, j0) \notin \bigcup_{\mathrm{Re}(s)\geqq -r_0} \{G(s)\}$.

(iv) *For* $\hat{G}(s) = G(s)[1+\frac{1}{2}\tilde{\pi}(c+d)G(s)]^{-1}$ (see (iii)) *and* $\hat{G}_2 = \hat{G} * \hat{G}$ (* *convolution*), *then*

$$[(\tfrac{1}{2}[\sigma^2(a^2+b^2)+\tilde{\pi}(c^2+d^2)])^{-1}, j0] \notin \bigcup_{\mathrm{Re}(s)\geqq -r_0} \{\hat{G}_2(s)\}.$$

(v) *For some* $\alpha \in (0,1)$ *and* $r \in (0, r_0)$,

$$\inf_{\mathrm{Re}(s)\geqq -r} |\hat{G}_2^{-1}(s) - \tfrac{1}{2}[\sigma^2(a^2+b^2)+\tilde{\pi}(c^2+d^2)]| \geqq \frac{1}{2\alpha}[\sigma^2(b^2-a^2)+\tilde{\pi}(d^2-c^2)]$$

and

(vi) *For some* $r \in (0, r_o)$ *and*

$$H(r+j\xi) \triangleq \int_{-\infty}^{\infty} \frac{\hat{G}(r+j(\xi-\xi_0))\hat{G}(r+j\xi_0)}{(r+j(\xi-\xi_0))(r+j\xi_0)}\, d\xi_0$$

the inequality below holds:

$$\sup_{\xi\in R} \left| \frac{\frac{1}{2}H(r+j\xi)\tilde{\pi}^2(d^2-c^2)}{1-\frac{1}{2}[\sigma^2(a^2+b^2)+\tilde{\pi}(c^2+d^2)]\hat{G}_2(r+j\xi)} \right| < 1-\alpha.$$

2.4. Comments and interpretations. The criteria given in Theorems 2.2 through 2.6 have more or less explicit graphical versions in terms of the frequency response $G(s)$ of the linear part. Theorem 2.2 is a standard Nyquist condition and may be easily checked. The condition (ii) in Theorem 2.3 may be evaluated directly when $G(s)$ has certain forms; for example, if $G(s) = k/(s+p)$, then

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \left| \frac{G(j\alpha)}{1+(m+\hat{\pi})G(j\alpha)} \right|^2 d\alpha = \frac{k^2}{2(p+m+\hat{\pi})}.$$

Hence, in (2.1) $\sup_{t\in R^+} Ex^2(t) \leqq \beta \sup_{t\in R^+} Eu^2(t)$ for some $\beta \in R^+$ if and only if

$$k^2\sigma^2 - 2m + \hat{\pi}(k^2-2) < 2p.$$

See [23, Appendix E] for evaluation of the integrals $(1/(2\pi))\int_{-\infty}^{\infty} |H(j\omega)|^2\, d\omega$ when H is a rational function of degree less than or equal to ten. See [6] for a related interpretation in terms of the Hurwitz polynomials associated with rational $G(s)$. The Corollaries 2.4 and 2.5 give easily verified graphical conditions for condition (ii) of Theorem 2.3 to hold. Corollary 2.4 is reminiscient of the deterministic circle theorem though the intention here is different. Corollary 2.5 is a "positivity condition" and immediately suggests a number of modifications using "multipliers" to shift the phase of $G(s)$; see [4] for the method and details of this procedure.

The derivation of the main criterion, Theorem 2.6, is based directly on the deterministic circle criterion of Zames [1] and Sandberg [2], especially the "L_∞ version" in [5]. This is more easily seen in the following special case of Theorem 2.6. Assume that f satisfies the sector condition with parameters (a, b) and that $\nu \equiv 0$ so that only Gaussian noise (w) feedback is permitted.

COROLLARY 2.7. *Assume the conditions on u and w given in Theorem* 2.6 *hold, in particular,* $Eu(t) = 0 = Edw(t)$. *If* $Eu^2(t) \in L_\infty(R^+)$ *and*

(i) $\int_0^\infty e^{r_0 t} g^2(t)\, dt < \infty$ *for some* $r_0 > 0$,

(ii) *for* $H(s) = \int_0^\infty e^{-st} g^2(t)\, dt$ *and for some* $r \in (0, r_0)$ *the exclusion below holds*:

$$(2\sigma^{-2}(a^2+b^2)^{-1}, j0) \notin \bigcup_{\mathrm{Re}(s) \geqq -r} \{H(s)\},$$

(iii) $$\inf_{\alpha \in R} \left| H^{-1}(-r+j\alpha) - \frac{\sigma^2}{2}(a^2+b^2) \right| > \frac{\sigma^2}{2}(b^2-a^2)$$

for some $r \in (0, r_0)$, *then* $\sup_{t \in R^+} Ex^2(t) \leqq \beta \sup_{t \in R^+} Eu^2(t)$ *for some* $\beta \in R^+$.

In geometric terms conditions (ii) and (iii) above are equivalent to the statement that the r-shifted Nyquist locus of $H(j\alpha)$ (the set $\cup_{\mathrm{Re}(s) \geqq -r} \{H(s)\}$) does not encircle (ii) or intersect (iii), the closed disc in the complex plane centered at $(+2\sigma^{-2}(a^{-2}+b^{-2}), j0)$ with radius $\sigma^{-2}(b^{-2}-a^{-2})$. Compare this with Theorem 2 in [5]. It may be possible to give a similar interpretation to the full Theorem 2.6, and although no attempt to do so will be given here, the promise of such a procedure is acknowledged.

In [10] versions of Theorem 2.3 were given for linear equations analogous to (2.1) with $\nu \equiv 0$ for discrete-time and vector-valued random process $x(k)$, $u(k)$, $w(k)$ defined in discrete time. A review of the results in [10] indicates that these situations may be treated easily in the present context, indeed, discrete processes represent a considerable simplification and multidimensional variables require only a more sophisticated notation.

Referring to the comparatively simple criterion of Corollary 2.7, it is apparent that the phase information in $G(s)$, the Laplace transform of $g(t)$, is used in a somewhat complicated manner. In contrast, in the linear case covered by Theorem 2.3, the phase of $G(j\omega)$ plays no role whatsoever in the decisive condition (ii).[3] Since Theorem 2.3 is necessary and sufficient, it follows that the phase information will be useful only in more stringent (sufficient) criteria, i.e., Corollary 2.4. This peculiar property of independence of the criteria from the phase of $G(s)$ appears to arise as a consequence of the properties of the white noise in the equation, especially the property of orthogonal increments. In a physical sense the use of white noise as a model for feedback gains is artificial, and a more realistic model would result if the noise were allowed to be "colored", i.e., have correlated increments. Investigations (for example, [15] and references therein) have shown that the resulting "stability problem" is considerably more difficult than that considered here. In particular the statistics of the noise seem to enter in a nonlinear fashion, and other aspects, such as the role of the argument of $G(s)$, appear to have a subtler influence than that observed in the present case.

[3] Assuming $\hat{\pi} = 0 = m$, that is, the disturbances are zero mean. Otherwise replace G by $G/[1 + (\hat{\pi}+m)G]$.

Note further that little use has been made of the properties of $G_2(s)$ in Theorem 2.6 (or the simpler $H(s)$ in Corollary 2.8) as the Laplace transform of a nonnegative function $g^2(\cdot)$. As such G_2 has some special properties and use of these should simplify, for example, conditions (v) and (vi) of Theorem 2.6.

The preceding results have been aimed at bounding the second moments of the solution process. Using an inequality due to Zakai [29], we can obtain similar bounds for moments of arbitrary order. For simplicity we consider only the linear case.

THEOREM 2.8. *Let $w(t)$ be a standard R-valued Wiener process, and $g: R^+ \to R$ such that $\|g\|_{L_2} < \infty$. Then $x(t)$ satisfying*

$$x(t) = u(t) - \int_0^t g(t-s)x(s)\, dw(s)$$

has $E|x(t)|^p < \infty$ for some $p \geqq 1$ and each t in R^+ if $u(t)$ has $E|u(t)|^p < \infty$ for each t. Moreover, for $p \geqq 2$, if

$$\sup_{\omega \in R} |G(j\omega)| < \left[\frac{2\pi}{p-1}\right], \tag{2.5}$$

where $G(j\omega) = \int_0^\infty g(t) \exp(-j\omega t)\, dt$, then $\|u\|_p = \sup (E|u(t)|^p)^{1/p} < \infty$ implies

$$\|x\|_p < \left[1 - \left(\frac{p-1}{2\pi}\right)^{1/2} \sup_{\omega} |G(j\omega)|\right]^{-1} \|u\|_p.$$

If $p = 2$, the condition (2.5) *is necessary and sufficient for mean square stability.*

Remarks. (a) These results complement those of our paper [15] where vector Itô equations are considered.

(b) By modifying the hypotheses slightly, we can derive asymptotic stability criteria (for moments of arbitrary order as above) assuming that $(E|u(t)|^p)^{1/p} \to 0$ (or is integrable, etc.).

3. Some properties of stochastic integrals. Let w denote the standard real-valued Wiener process on R^+, normalized so that $w(0) = 0$. *The Wiener measure* w is a probability measure on $(\Omega, \mathscr{F}) = (C(R^+; R), \mathscr{B}(C))$ satisfying two properties. For each $t, s \in R^+$ the random variable $w(t) - w(s)$ is normally distributed (on R) with mean

$$E\{w(t) - w(s)\} = m \cdot (t-s), \qquad m \in R,$$

and variance

$$E\{[w(t) - w(s) - m(t-s)]^2\} = \sigma^2 \cdot |t-s|;$$

for any finite collection of elements $\{t_i\}_{i=1}^n \subset R^+$ such that $t_1 \leqq t_2 \leqq \cdots \leqq t_n$, the random variables $w(t_2) - w(t_1)$, $w(t_3) - w(t_2)$, $\cdots$, $w(t_n) - w(t_{n-1})$ are independent under the normal distribution, i.e., w has independent increments. For any nonanticipating[4] random functional ψ, the Itô stochastic integral versus w is

[4] $\mathscr{B}_{0t}(\psi) \vee \mathscr{B}_{0t}(w)$ independent of $\mathscr{B}_{t\infty}(dw)$. See [25] for details.

denoted by

$$\int_s^t \psi(r, \omega)\, dw(r, \omega).$$

See, for instance, [24] for its properties.

For some given measureable space $(X, \beta(X))$ consider the random measure on $\mathcal{B}(R^+)\times\mathcal{B}(X)$ denoted by $\nu([s, t], A), [s, t]\subset R^+, A\in\mathcal{B}(X)$, as expressing the number of events in the set A during the interval $[s, t]$. Assume that the random variable ν takes on nonnegative integer values independent on disjoint elements of $\mathcal{B}(R^+)\times\mathcal{B}(X)$; moreover, for each set $[s, t]\times A\in\mathcal{B}(R^+)\times\mathcal{B}(X)$, assume that $\nu([s, t], A)$ is Poisson with parameter

$$\int_s^t \Pi(r, A)\, dr.$$

That is,

$$P\{\omega : \nu(\omega, [s, t], A) = n\} = \frac{1}{n!}\left(\int_s^t \Pi(r, A)\, dr\right)^n \exp\left(-\int_s^t \Pi(r, A)\, dr\right).$$

Here $\Pi(t, A)$ is a (given) probability measure on $\mathcal{B}(X)$ for each $t\in R^+$, and a measurable function mapping R^+ into R for each $A\in\mathcal{B}(X)$.

It follows that the random process ν is a process with independent increments (on R^+); so the stochastic integral

$$\int_s^t \int_X l(r, x)\nu(dr, dx)$$

is well-defined as the usual limit of retarded Riemann sums for nonanticipating (with respect to ν) random functionals l on $R^+\times W$ such that

$$\int_s^t \int_X E|l(r, x)|^k \Pi(r, dx)\, dr < \infty, \qquad k = 1, 2.$$

See [26], [13], and [22] for more details.

For the usual Itô integral,

$$y(t) = \int_0^t x(s)\, dw(s),$$

where w is a zero-mean, σ^2-variance parameter Wiener process and x is a nonanticipating (w) random process, it is easily shown [25, p. 24] that if

$$P\left\{\omega : \int_0^t x^2(t, \omega)\, ds < \infty; t\in R^+\right\} = 1,$$

then

$$Ey(t) = \int_0^t Ex(s)Edw(s) = 0,$$

$$Ey^2(t) = \sigma^2 \int_0^t Ex^2(s)\, ds. \tag{3.1}$$

Equally true, but less often used, are similar facts for

$$y(t)=\int_0^t\int_{-\infty}^{\infty} h(s,x)\nu(ds,dx),$$

where ν is defined above with parameter $\pi(s,dx)\,ds$ and h is nonanticipating. If for $t\in R^+$,

$$\int_0^t\int_{-\infty}^{\infty} Eh^k(s,x)\Pi(s,dx)\,ds<\infty,\qquad k=1,2,$$

then

$$y(t)=\int_0^t\int_{-\infty}^{\infty} h(s,x)\lambda(ds,dx)$$

is a martingale in t, and

$$Ey(t)=E\left\{\int_0^t\int_{-\infty}^{\infty} h(s,x)\lambda(ds,dx)\right\}=0,$$

(3.2)

$$Ey^2(t)=\int_0^t\int_{-\infty}^{\infty} Eh^2(s,x)\Pi(s,dx)ds,$$

where $\lambda(ds,dx)=\nu(ds,dx)-\Pi(s,dx)ds$. Properties (3.1) and (3.2) are used in the proofs of Theorems 2.2, 2.3, and 2.6.

4. Proofs of Theorems 2.2 through 2.6.

4.1. Proof of Theorem 2.2. The hypothesis that $g\in L_2(R^+)$ and the local existence Theorem 2.1 prove that

$$\int_0^t Ex^2(s,\omega)\,ds<\infty$$

for any $t\in R^+$. Thus, the properties of the stochastic integral (equations (3.1) and (3.2)) apply. Equation (2.3) is thus valid, and the theorem follows from a result of Davis [26] and the observation that $Ex(t)$ is piecewise continuous. Q.E.D.

4.2. Proof of Theorem 2.3. From the assumption $g\in L_2(R^+)$ and the condition (i) it follows that the operator $[I+\frac{1}{2}(\hat{\pi}+m)G(s)]$ has an inverse (on locally $L_2(R^+)$ functions) represented by the identity minus a convolution [27]. Let G_1 be the linear convolution whose kernel g_1 has Fourier transform

$$G_1(j\alpha)=[1+(m+\hat{\pi})G(j\alpha)]^{-1},$$

and denote by $\tilde{g}$ the function with corresponding transform $\tilde{G}(j\alpha)=G(j\alpha)G_1(j\alpha)$. Then from (2.4) and the preceding comments,

$$x(t)=(G_1u)(t)-\int_0^t \tilde{g}(t-s)x(s)\,d\tilde{w}(s)-\int_0^t \tilde{g}(t-s)x(s)\int_{-\infty}^{\infty} h(y)\tilde{\nu}(ds,dy)$$

and it follows that

$$Ex^2(t) = \int_0^t \int_0^t g_1(t-s)g_1(t-r)E[u(s)u(r)]\,ds\,dt$$
$$+ \int_0^t \tilde{g}^2(t-s)Ex^2(s)(\sigma^2+\hat{\pi})ds.$$

Therefore,

$$\sup_{0\leqq s\leqq t} Ex^2(s) \leqq \left(\int_0^\infty g_1(r)\,dr\right) \sup_{s\in R^+} Eu^2(s)$$
$$+(\sigma^2+\hat{\pi})\|\tilde{g}\|_{L_2} \sup_{0\leqq s\leqq t} Ex^2(s).^5$$

Parseval's theorem and condition (ii) of Theorem 2.3 guarantee that $(\sigma^2+\hat{\pi})\|\tilde{g}\|_{L_2}<1$. Therefore $\sup_{0\leqq s\leqq t} Ex^2(s)\leqq\beta \sup_{s\in R^+} Eu^2(s)$ for β independent of $t\in R^+$ and u. The desired conclusion follows from this independence. Q.E.D.

4.3. Proof of Corollary 2.4. By Theorem 2.3 it suffices to show that

$$(\sigma^2+\hat{\pi})\frac{1}{2\pi}\int_{-\infty}^{\infty} |G(j\alpha)/[1+(m+\hat{\pi})G(j\alpha)]|^2\,d\alpha<1.$$

Note that for an arbitrary complex function H, if

$$|H(j\alpha)-\delta/2|<\delta/2$$

(condition like that stated in (a), (b), (c) of (iii)), then $|H(j\alpha)|^2\leqq\delta\ \mathrm{Re}\ H(j\alpha)$. Using this fact and the restrictions on the graph of $G(j\alpha)$, it follows that

$$\left|\frac{(m+\hat{\pi})G(j\alpha)}{1+(m\hat{\pi})G(j\alpha)}\right|^2 \leqq [1+\gamma(m+\hat{\pi})^{-1}]^{-1}\,\mathrm{Re}\frac{(m+\hat{\pi})G(j\alpha)}{1+(m+\hat{\pi})G(j\alpha)}.$$

Thus,

$$\frac{(\sigma^2+\hat{\pi})}{2\pi}\int_{-\infty}^{\infty} |\tilde{G}(j\alpha)|^2 f\alpha \leqq \frac{\sigma^2+\hat{\pi}}{m+\hat{\pi}+\gamma}\frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{G}(j\alpha)\,d\alpha$$
$$=\frac{\sigma^2+\hat{\pi}}{m+\hat{\pi}+\gamma}\frac{\tilde{g}(0)}{2}.$$

The last step uses the fact that $g\in L_1(R^+)$ and that zero is a Lebesgue point of $\tilde{g}$. Condition (ii) of the Corollary implies the conclusion through Theorem 2.3. Q.E.D.

As noted in the text, Corollary 2.5 is a limiting case of Corollary 2.4 as $\gamma\to 0$.

4.4. Proof of Theorem 2.6. This result is proved in exactly the same manner as the deterministic circle criterion. The first few steps below are intended to transform the system equation (2.1) into a form where the deterministic techniques may be applied.

[5] Note that $g_1\in L_1(R^+;R)$ from, for example, Hille and Phillips [28, p. 155].

A transformation of (2.1) gives

$$x(t)=u(t)-\int_0^t g(t-s)\int_{-\infty}^{\infty}\tilde{h}(s,x(s),y)\Pi(dy)\,ds$$

$$-\frac{1}{2}(c+d)\tilde{\pi}\int_0^t g(t-s)x(s)\,ds-\int_0^t g(t-s)f(s,x(s))\,dw(s)$$

$$-\int_0^t g(t-s)\int_{-\infty}^{\infty}h(s,x(s),y)\tilde{\nu}(ds,dy),$$

where $\tilde{\nu}(ds,dy)=\nu(ds,dy)-\Pi(dy)\,dt$ and $\tilde{h}(s,x,y)=h(s,x,y)-\frac{1}{2}(c+d)x$. The constant $\tilde{\pi}$ is defined in (ii). Let $W(s)=[1+\frac{1}{2}\tilde{\pi}(c+d)G(s)]$. Then by (ii) and (iii) and from, for example, [26], W^{-1} exists on $L_\infty(R^+)$ functions. Hence

$$x(t)=(W^{-1}u)(t)-\int_0^t \hat{g}(t-s)\int_{-1}^{\infty}\tilde{h}(s,x(s),y)\Pi(dy)\,ds$$

$$-\int_0^t \hat{g}(t-s)f(s,x(s))\,dw(s)$$

$$-\int_0^t \hat{g}(t-s)\int_{-\infty}^{\infty}h(s,x(s),y)\tilde{\nu}(ds,dy).$$

Here $\hat{g}$ is the inverse Laplace transform of $\hat{G}(s)=G(s)W^{-1}(s)$.

Taking into account the assumptions on u, w, and ν:

$$Ex^2(t)=E\{(W^{-1}u)^2(t)\}+E\left\{\left(\int_0^t \hat{g}(t-s)\int_{-\infty}^{\infty}\tilde{h}(s,x(s),y)\Pi(dy)\,ds\right)^2\right\}$$

$$+\sigma^2\int_0^t \hat{g}^2(t-s)E\{f^2(s,x(s))\}\,ds$$

$$+\int_0^t \hat{g}^2(t-s)\int_{-\infty}^{\infty}E\{h^2(s,x(s),y)\}\Pi(dy)\,ds.$$

Adding and subtracting the terms

$$\frac{1}{2}\sigma^2(a^2+b^2)\int_0^t \hat{g}^2(t-s)Ex^2(s)\,ds,$$

$$\frac{1}{2}\tilde{\pi}(c^2+d^2)\int_0^t \hat{g}^2(t-s)Ex^2(s)\,ds,$$

the result is

$$
\begin{aligned}
Ex^2(t) &- \frac{1}{2}[\sigma^2(a^2+b^2)+\tilde{\pi}(c^2+d^2)]\int_0^t \hat{g}^2(t-s)Ex^2(s)\,ds \\
&= E\{(W^{-1}u)^2(t)\}+E\left(\int_0^t \hat{g}(t-s)\int_{-\infty}^{\infty}\tilde{h}(s,x(s),y)\Pi(dy)\,ds\right)^2 \\
&\quad +\sigma^2\int_0^t \hat{g}^2(t-s)E\hat{f}^2(s,x(s))\,ds \\
&\quad +\int_0^t \hat{g}^2(t-s)\int_{-\infty}^{\infty} E\hat{h}^2(s,x(s),y)\Pi(dy)\,ds,
\end{aligned}
$$

where $\hat{f}^2(s,x)=f^2(s,x)-\frac{1}{2}(a^2+b^2)x^2$ and $\hat{h}^2(s,x,y)=h^2(s,x,y)-\frac{1}{2}(c^2+d^2)x^2$.

Let K be the convolution operator whose transform is

$$K(s)=[1-\tfrac{1}{2}[\sigma^2(a^2+b^2)+\tilde{\pi}(c^2+d^2)]\hat{G}_2(s)]^{-1},$$

where $\hat{G}_2(s)$ is defined in the theorem statement (iv). The operator K is well-defined (Re $(s)\geqq -r_0$) by (iv). Let k denote the inverse transform of K.

Then

$$
\begin{aligned}
Ex^2(t) &= K(E\{(W^{-1}u)^2\})(t) \\
&\quad +\int_0^t k(t-s)E\left\{\left(\int_0^s \hat{g}(s-v)\int_{-\infty}^{\infty}\tilde{h}(v,x(v),y)\Pi(ds)\,dv\right)^2\right\}ds \\
&\quad +\sigma^2\int_0^t n(t-s)E\hat{f}^2(s,x(s))\,ds \\
&\quad +\int_0^t n(t-s)\int_{-\infty}^{\infty} E\hat{h}^2(s,x(s),y)\Pi(dy)\,ds.
\end{aligned}
\tag{4.1}
$$

Here n is the kernel whose transform is $N(s)=\hat{G}_2(s)K(s)$.

Using the bounds,

$$
\begin{aligned}
|\hat{f}^2(s,x)| &\leqq \tfrac{1}{2}(b^2-a^2)x^2 \quad \text{for every } s\in R^+, \\
|\hat{h}^2(s,x,y)| &\leqq \tfrac{1}{2}(d^2-c^2)x^2 \quad \text{for every } s\in R^+,\quad y\in R,
\end{aligned}
$$

and condition (v) it is clear that the last two terms in (4.1) are bounded by $\alpha \sup_{0\leqq s\leqq t} Ex^2(s)$. Closer consideration of the decisive term (T2) second on the right of (4.1) will yield the desired conclusion. Expanding the square in T2,

$$
\begin{aligned}
&\int_0^s\int_0^s \hat{g}(s-v)\hat{g}(s-z)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E[\tilde{h}(s,x(s),y)\tilde{h}(z,x(z),q)]\Pi(dy)\Pi(dq)\,ds\,dz \\
&\quad \leqq \iint \hat{g}(s-v)\hat{g}(s-z)\iint [E\tilde{h}^2(s,y)]^{1/2}[E\tilde{h}^2(z,q)]^{1/2}\Pi(dy)\Pi(dq)\,ds\,dz \\
&\quad \leqq \frac{1}{2}\tilde{\pi}^2(d^2-c^2)\left[\int_0^s |\hat{g}(s-v)|\,dv\right]^2 \sup_{0\leqq z\leqq s} Ex^2(z).
\end{aligned}
$$

Hence

$$\mathrm{T2} \leqq \frac{1}{2}\tilde{\pi}^2(d^2-c^2)\int_0^t |k(t-s)|\Big(\int_0^s \hat{g}(v)\,dv\Big)^2 \sup_{0\leqq z\leqq s} Ex^2(z)\,dz.$$

Thus some standard inequalities and condition (vi) imply that

$$|\mathrm{T2}| \leqq (1-\alpha)\sup_{0\leqq s\leqq t} Ex^2(s)$$

and that the operator composed of T2 and the sum of the last two terms is a contraction on the Banach space defined by the norm $\|x\| = [\sup_{s\in R^+} Ex^2(s)]^{1/2}$. The conclusion follows now from arguments identical to those used in the last stages of the proof of Theorem 2.3. Q.E.D.

4.5. Proof of Theorem 2.8. Using $(E|a+b|^p)^{1/p} \leqq (E|a|^p)^{1/p} + (E|b|^p)^{1/p}$ for $p \geqq 1$, we have for $x(t)$ satisfying (2.5) and $\|x\|p = (E|x|^p)^{1/p}$,

$$\|x(t)\|_p \leqq \|u(t)\|_p + \Big\|\int_0^t g(t-s)x(s)\,dw(s)\Big\|_p.$$

Now from Zakai [29], for $p \geqq 2$, $T < \infty$ and any nonanticipating functional F,

$$E\Big|\int_0^t F(t)\,dw(t)\Big|^p \leqq (p-1)^{p/2}\Big(\int_0^t (E|F(t)|^p)^{2/p}\,dt\Big)^{p/2}.$$

Hence,

$$\|x(t)\|_p \leqq \|u(t)\|_p + (p-1)^{1/2}\Big(\int_0^t g^2(t-s)\|x(s)\|_p^2\,ds\Big)^{1/2}.$$

So if (2.5) holds, then for any t in R^t,

$$\sup_{s\leqq t}\|x(s)\|_p \leqq [1-a(p)]^{-1}\sup_{s\leqq t}\|u(s)\|_p,$$

where $a(p) = (p-1)^{1/2}\|g\|_{L_2} < 1$ is assumed.

The condition $a(p) < 1$ is equivalent to (2.5) by Parseval's theorem. Q.E.D.

REFERENCES

[1] G. ZAMES, *On the input-output stability of time-varying nonlinear feedback systems—Part I: Conditions derived using concepts of loop gain, conicity, and positivity; Part II: Conditions involving circles in the frequency plane and sector nonlinearities*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238, and 465–476.

[2] I. W. SANDBERG, *Some results on the theory of physical systems governed by nonlinear functional equations*, Bell System Tech. J., 44 (1965), pp. 871–898.

[3] J. HOLTZMAN, *Nonlinear System Theory: A Functional Analysis Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1970.

[4] J. C. WILLEMS, *The Analysis of Feedback Systems*, MIT Press, Cambridge, Mass., 1971.

[5] G. ZAMES, *Nonlinear time varying feedback systems—Conditions for L_∞-boundedness derived using conic operators on exponentially weighted spaces*, Proc. Third Allerton Conf. on Circuit and System Theory, Urbana, Ill., 1965, p. 460–471.

[6] G. BEKEY, *Description of the human operator in control systems*, Modern Control Systems Theory, C. T. Leondes, ed., McGraw-Hill, New York, 1965, pp. 431–462.

[7] D. KLEINMAN, S. BARON AND W. LEVISON, *A control theoretic approach to manned-vehicle systems analysis*, IEEE Trans. Automatic Control, AC-16 (1972), pp. 824-832.
[8] P. HENRICE, *Error Propagation for Difference Methods*, John Wiley, New York, 1963.
[9] G. CARRIER, *Stochastically driven dynamical systems*, J. Fluid Mech, 44 (1970), part 2, pp. 249–264.
[10] J. WILLEMS AND G. BLANKENSHIP, *Frequency domain stability criteria for stochastic systems*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 292–299.
[11] J. L. WILLEMS, *Lyapunov functions and global frequency domain stability criteria for a class of stochastic feedback systems*, Stability of Stochastic Dynamical Systems, Lecture Notes in Mathematics, #294, Springer-Verlag, New York, 1972, pp. 139–146.
[12] R. BROCKETT AND J. C. WILLEMS, *Average value criterion for stochastic stability*, Stability of Stochastic Dynamical Systems, Lecture Notes in Mathematics, #294, Springer-Verlag, New York, 1972, pp. 252–272.
[13] G. BLANKENSHIP, *A circle criterion for nonlinear stochastic feedback systems*, Proc. 1972 Joint Automatic Control Conference, Stanford, Calif., 1972, pp. 212–219.
[14] P. BILLINGSLY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
[15] G. BLANKENSHIP, *Lie Theory and the Moment Stability Problem for Stochastic Differential Equations*, IFAC Proceedings, 1975.
[16] R. KHAS'MINSKII, *Necessary and sufficient conditions for the asymptotic stability of linear stochastic systems*, Theor. Probability Appl., 1 (1967), pp. 144–147.
[17] F. KOZIN AND C. WU, *On the stability of linear stochastic differential equations*, ASME J. Appl. Mech., 40 (1973), pp. 87–92.
[18] R. MITCHELL, *Sample stability of second order stochastic differential equations with nonsingular phase diffusion*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 706–707.
[19] F. KOZIN AND S. PRODROMOU, *Necessary and sufficient conditions for almost sure sample stability of linear Ito equations*, SIAM J. Appl. Math., 21 (1971), pp. 413–424.
[20] R. MITCHELL AND F. KOZIN, *Sample stability of second order linear differential equations with wide band noise coefficients*, Ibid., 27 (1974), pp. 571–605.
[21] F. KOZIN, *On almost sure asymptotic sample properties of diffusion processes defined by stochastic differential equations*, J. Math. Kyoto Univ., 4 (1965), p. 575.
[22] A. V. SKOROKHOD, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Mass., 1965.
[23] G. NEWTON, L. GOULD, AND J. KAISER, *Analytic Design of Linear Feedback Controls*, John Wiley, New York, 1957.
[24] H. P. MCKEAN, *Stochastic Integrals*, Academic Press, New York, 1969.
[25] K. ITO, *On stochastic differential equations*, Mem. Amer. Math. Soc., No. 4, 1951.
[26] J. DAVIS, *Stability of linear feedback systems*, Ph.D. thesis, Dept. of Mathematics, Mass. Inst. of Tech., Cambridge, Mass., 1970.
[27] V. E. BENEŜ, *A nonlinear integral equation in the Marcinkiewicz space M_2*, J. Math. and Phys., 44 (1965), pp. 24–35.
[28] E. HILLE AND R. PHILLIPS, *Functional Analysis and Semi-Groups*, American Mathematical Society, Providence, R.I., 1957.
[29] M. ZAKAI, *Some moment inequalities for stochastic integrals and for solutions of stochastic differential equations*, Israel J. Math., 5 (1967), pp. 170–176.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 6, November 1976

ESTIMATION THEORY FOR ABSTRACT EVOLUTION EQUATIONS EXCITED BY GENERAL WHITE NOISE PROCESSES*

RUTH F. CURTAIN

Abstract. The filtering smoothing and prediction problems are solved for a general class of linear infinite-dimensional systems. The dynamical system is modeled as an abstract evolution equation, which includes linear ordinary differential equations, classes of linear partial differential equations and linear differential delay equations. The noise process is modeled using a stochastic integral with respect to a class of Hilbert space-valued stochastic processes, which includes the Wiener process and the Poisson process as special cases. The observation process is finite-dimensional and is corrupted by Gaussian-type white noise, which is modeled using the Wiener integral. The theory is illustrated by an application to an environmental problem.

Introduction. We solve the filtering, smoothing and prediction problem for a very general class of linear infinite-dimensional systems, where the system disturbance is a general noise process, which includes Gaussian and Poisson-type white noise. Earlier papers on infinite-dimensional filtering including [2], [4], [6], [18], [8], consider the filtering problem for various systems where the disturbance is white Gaussian noise. Our approach to the filtering problem follows the ideas first introduced by Falb in [13] and later papers [5], [6], [18] and [8]; that is, the noise process is modeled using a stochastic integral with respect to a Hilbert space-valued Wiener process. By introducing the notion of stochastic integration with respect to a wider class of general Hilbert space-valued stochastic processes we are able to model more general types of noise disturbances. The infinite-dimensional system is modeled using the evolution operator approach, which has proved so successful in solving deterministic control problems in abstract spaces (see [10]) and allows for a very wide class of systems including those described by partial differential equations and delay equations. Using the ideas of Kailath [14], an innovations approach is used to solve the prediction and smoothing problems as was done in [8] for the Gaussian white noise case. It is found that optimal linear estimators exist under very weak conditions on the system operators, but in order to obtain differential forms, extra conditions must be imposed. However these conditions are not unduly restrictive in applications, and this is illustrated by considering examples. In particular, a river pollution problem is considered where the noise process is Poisson-like.

1. Preliminaries.

1.1. Evolution operators. We summarize the theory of evolution operators developed in [9] to model a large class of linear infinite-dimensional systems, which are appropriate for control theory applications.

DEFINITION 1.1. *Mild evolution operator.* Let H be a real Hilbert space and $T=[0, T]$ a real finite time interval and denote $\Delta(T)=\{(s, t): 0 \leqq s<t \leqq T\}$. Then $\mathcal{U}(\cdot, \cdot): \Delta(T) \to \mathcal{L}(H)$ is a *mild evolution operator* if

(a) $\mathcal{U}(t, r)\mathcal{U}(r, s)=\mathcal{U}(t, s)$ for $0 \leqq s \leqq r \leqq t \leqq T$, $\mathcal{U}(t, t)=I$.

(b) $\mathcal{U}(t, s)$ is weakly continuous in s on $[0, t]$ and in t on $[s, T]$.

* Received by the editors October 22, 1975.

† Control Theory Centre, University of Warwick, Coventry CV4 7AL, Warwickshire, England.

The following theorem allows us to define perturbations of mild evolution operators.

THEOREM 1.1. *Let* $\mathcal{U}(\cdot,\cdot)$ *be a mild evolution operator and* $D \in \mathcal{B}_\infty(T;\mathcal{L}(H))$ *the space of* $\mathcal{L}(H)$*-valued functions which are strongly measurable on* T *with* $\operatorname{ess\,sup}_{t\in T}\|D(t)\| < \infty$. *Then the following integral equation on* $\mathcal{L}(H)$ *has a unique solution* $\mathcal{U}_D(\cdot,\cdot)$ *in the class of weakly continuous* $\mathcal{L}(H)$*-valued functions on* $\Delta(T)$:

$$\mathcal{U}_D(t,s)x = \mathcal{U}(t,s)x + \int_s^t \mathcal{U}(t,r)D(r)\mathcal{U}_D(r,s)x\,dr, \qquad x \in H. \tag{1.1}$$

$\mathcal{U}_D(\cdot,\cdot)$ is a mild evolution operator and is called the perturbation of $\mathcal{U}(t,s)$ corresponding to the perturbation D. A stronger concept, closely related to semigroup concept is

DEFINITION 1.2. *Strong evolution operator. A strong evolution operator* is a mild evolution operator $\mathcal{U}(t,s)$ with an associated generator $\mathcal{A}(t)$, which for each $t \in T$ is a closed, densely-defined linear operator on H such that

(a) $\mathcal{U}(t,s):\mathcal{D}(\mathcal{A}(s)) \to \mathcal{D}(\mathcal{A}(t))$ for $t > s$,

(b) $(\partial/\partial t)\mathcal{U}(t,s)x = \mathcal{A}(t)\mathcal{U}(t,s)x$ for $x \in \mathcal{D}(\mathcal{A}(s))$, $t > s$.

We remark that if $\mathcal{U}(t,s)$ is a strong evolution operator, then the abstract evolution equation

$$\begin{aligned} \dot{u}(t) &= \mathcal{A}(t)u(t) + g(t), \\ u(s) &= u_0 \in \mathcal{D}(\mathcal{A}(s)) \end{aligned} \tag{1.2}$$

has a unique, strongly continuous solution, given by

$$u(t) = \mathcal{U}(t,0)u_0 + \int_s^t \mathcal{U}(t,r)g(r)\,dr \tag{1.3}$$

provided g is Hölder continuous on T (see [9]). If $u_0 \in H$, $g \in L_2(T;H)$, then (1.3) is still well-defined and we call it the *mild solution* of (1.2).

For stochastic differential equations, it happens that a more useful concept is that of an *almost strong evolution* operator where (b) is replaced by

$$\int_s^t \mathcal{A}(r)\mathcal{U}(r,s)x\,dr = \mathcal{U}(t,s)x - x \tag{b}'$$

$$\text{for } x \in \mathcal{D}_{s,t}(\mathcal{A}) = \{x : \mathcal{U}(r,s)x \in \mathcal{D}(\mathcal{A}(r));\ s \leqq r \leqq t\}$$

which implies that

$$\frac{\partial}{\partial t}\mathcal{U}(t,s)x = \mathcal{A}(t)\mathcal{U}(t,s)x \quad \text{a.e. for } x \in \mathcal{D}_{s,t}(\mathcal{A}), \quad t > s.$$

The concepts of strong and almost strong evolution operators are obviously closely related and often coincide, as in the case of analytic semigroups, where $\mathcal{D}_{s,t}(\mathcal{A}) = H$ and (b) holds for all $x \in H$. If we suppose that $\mathcal{A}(t)\mathcal{U}(t,s)x$ is Bochner integrable for $x \in \mathcal{D}(\mathcal{A}(s))$, then (b)′ holds for $x \subset \mathcal{D}(\mathcal{A}(s))$. However, in general, $\mathcal{D}_{s,t}(\mathcal{A}) \supset \mathcal{D}(\mathcal{A}(s))$ and so (b)′ holds for a wider subset of H. This is very important in applications as is illustrated in [7] and [8].

1.2. Abstract probability theory. Let $(\Omega, \mathcal{P}, \mu)$ be a complete probability space, H, K real separable Hilbert spaces and $T=[0, T]$, a real finite time interval. Then we shall use the following standard definitions.

DEFINITION 1.3. An *H-valued random variable* is a map $u:\Omega\to H$ which is measurable with respect to the μ-measure. If $u\in L_1(\Omega, \mu; H)$, we define its expectation by

$$E\{u\}=\int_\Omega u\, d\mu.$$

If $u\in L_2(\Omega, \mu; H)$, we define its covariance operator by

$$E\{(u-E\{u\})\circ(u-E\{u\})\},$$

where $u\circ v\in\mathcal{L}(H, K)$ is defined for all $u\in H$, $v\in K$, by

$$(u\circ v)h=u\langle v, h\rangle \quad \text{for } h\in K.$$

We note that $u\circ u$ is a self adjoint nuclear operator, with trace $\{u\circ u\}=\|u\|^2$.

DEFINITION 1.4. An *H-valued stochastic process* is a map $u(\cdot,\cdot):T\times\Omega\to H$ which is measurable on $T\times\Omega$ using the Lebesgue measure on T.

DEFINITION 1.5. H and K-valued random variables u and v are *independent* if $\{\omega: u(\omega)\in A\}$ and $\{\omega: v(\omega)\in B\}$ are independent sets in $\mathcal{P}$ for any Borel sets A in H and B in K.

DEFINITION 1.6. An H-valued stochastic process $\{m(t)\}$ is a *martingale* relative to an increasing sigma-field $\{\mathcal{F}_t\}$ if

$$E\{m(t)|\mathcal{F}_s\}=m(s) \quad \text{with probability one (w.p.1) for } t>s. \tag{1.4}$$

DEFINITION 1.7. An H-valued random variable $h\in L_2(\Omega, \mu; H)$ is Gaussian if $\langle h, e_i\rangle$ is a real Gaussian random variable for all i, where $\{e_i\}$ is a complete orthonormal basis for H.

The following theory of estimation for H-valued random variables is from Bensoussan [2].

The estimation problem is to estimate a random variable $x\in L_2(\Omega, \mu; H)$ from a random variable $y\in L_2(\Omega, \mu; K)$. $L_2(K, \sigma; H)$ is isometric to $\tilde{L}_2(K, \sigma; H)$, a closed subspace of $L_2(\Omega, \mu; H)$, where (K, σ) is the probability space induced by the random variable y.

DEFINITION 1.8. The *best global estimate* $\hat{x}=E\{x|y\}$ of x on y is the projection of x on $\tilde{L}_2(K, \sigma; H)$:

$$\hat{x} \quad \text{always exists and is unique.}$$

The *best linear estimate* $\bar{x}$ of x based on y is $\bar{x}=Xy$ where $X\in\mathcal{L}(K, H)$ minimizes $E\{\|x-\Lambda y\|^2\}$ over all $\Lambda\in\mathcal{L}(K, H)$.

If x and y are Gaussian, $\bar{x}=\hat{x}$.

We now introduce a class of stochastic processes of particular interest to us in applications.

DEFINITION 1.9. An H-valued orthogonal increments process $\{q(t), t\in T\}$ is such that

$$q(t)=\sum_{i=0}^{\infty} q_i(t)\, e_i \tag{1.5}$$

where $q_i(t)$ are "orthogonal increments" processes of the same type, and $\{e_i\}$ is an orthonormal basis for H. More specifically,

$$E\{q_i(t)\} = \mu_i \rho(t),$$

where $\rho(t)$ is a monotone nonincreasing real function and $\sum \mu_i < \infty$ and

$$E\{(\bar{q}_i(t_2) - \bar{q}_i(s_2))(\bar{q}_j(t_1) - \bar{q}_j(s_1))\} = 0, \qquad 0 \leqq s_1 < t_1 < s_2 < t_2 \leqq T,$$

$$E\{(\bar{q}_i(t) - \bar{q}_i(s))(\bar{q}_j(t) - \bar{q}_j(s))\} = \lambda_{ij}(f(t) - f(s)), \qquad 0 \leqq s < t \leqq T,$$

where $\bar{q}_i(t) = q_i(t) - \mu_i p(t)$, and f is a monotone nonincreasing function, $\sum_{i=0}^{\infty} \lambda_i < \infty$; $\lambda_{ij} = \lambda_i$ and $\lambda_{ij}^2 \leqq \lambda_i \lambda_j$.

More simply we can write

$$E\{q(t)\} = r(t) = \left(\sum_{i=0}^{\infty} \mu_i e_i\right) \rho(t),$$

$$\bar{q}(t) = q(t) - r(t), \tag{1.6}$$

$$E\{\bar{q}(t) - \bar{q}(s) \circ \bar{q}(t) - \bar{q}(s)\} = \Lambda(f(t) - f(s)), \qquad 0 \leqq s < t \leqq T,$$

where Λ is a nuclear operator with $\Lambda e_j = \sum_{i=0}^{\infty} \lambda_{ij} e_i$, and trace $\Lambda = \sum_{i=0}^{\infty} \lambda_i$. $r(t)$ is called the expectation function and $\Lambda f(t)$ is called the covariance function of $q(t)$. We note that

$$E\{\|\bar{q}(t) - \bar{q}(s)\|^2\} = (f(t) - f(s)) \text{ trace } \Lambda, \qquad 0 \leqq s \leqq t \leqq T. \tag{1.7}$$

If $r(t) = 0$, $q(t)$ is called a *centered* orthogonal increments process; a particular example is

DEFINITION 1.10. An H-valued Wiener process is a centered H-valued orthogonal increments process on $T \times \Omega$ given by

$$w(t, \omega) = \sum_{i=0}^{\infty} \beta_i(t, \omega)\, e_i,$$

where $\beta_i(t, \omega)$ are real mutually independent Wiener processes, with $E\{\beta_i(t)^2\} = t\lambda_i$, and $\sum_{i=0}^{\infty} \lambda_i < \infty$. So

$$E\{w(t)\} = 0,$$

$$E\{w(t) - w(s) \circ w(t) - w(s)\} = W(t - s), \qquad 0 \leqq s < t \leqq T,$$

where W is a positive nuclear operator with $We_i = \lambda_i e_i$.

In [2], [3] it is shown that $w(t)$ actually has independent increments and has continuous sample paths.

Another example of an orthogonal increments process which is useful in applications is the Poisson process.

DEFINITION 1.11. An *H-valued Poisson process* is defined by

$$p(t, \omega) = \sum_{i=0}^{\infty} \pi_i(t, \omega)\, e_i,$$

where $\{e_i\}$ is a complete orthonormal basis for H and π_i are mutually independent real Poisson processes with parameter μ_i and $\sum_{i=0}^{\infty} \mu_i < \infty$.

Using the properties of the real Poisson process, we see that $\rho(t, w)$ is a well-defined stochastic process with

$$E\{p(t)\} = t \sum_{i=0}^{\infty} \mu_i e_i, \tag{1.8}$$

$$E\{\|p(t)-p(s)\|^2\} = \sum_{i=0}^{\infty} \mu_i(t-s) + \sum_{i=0}^{\infty} \mu_i^2(t-s). \tag{1.9}$$

Defining $m(t) = p(t) - E\{p(t)\}$, we see that $m(t)$ is a centered orthogonal increments process with

$$\begin{aligned} &E\{m(t)\} = 0, \\ &E\{m(t)-m(s) \circ m(t)-m(s)\} = M(t-s), \end{aligned} \tag{1.10}$$

where $Me_i = \mu_i e_i$.

In [3] a theory for integrals for Hilbert space-valued Wiener processes was developed and a study of stochastic evolution equations excited by Gaussian white noise was made. This theory was exploited in [2], [4], [6], [11], [18] to study the filtering problem for linear infinite-dimensional systems with Gaussian white noise disturbance. Recently more general stochastic integration theories have been developed, the most general by Métivier [16], [17] who uses a martingale approach. For our purposes the orthogonal increments noise approach in [7] is more appropriate and, in fact, was developed for application to the filtering problem for general infinite-dimensional systems. Here we outline the theory of stochastic integration with respect to a general orthogonal increments process.

DEFINITION 1.12. Let $q(t)$ be a centered orthogonal increments process and let $\Phi \in \mathscr{F}_2(T; \mathscr{L}(H, K))$, the class of strongly measurable $\mathscr{L}(H, K)$-valued functions with $\int_T \|\Phi(s)\|^2 \, df(s) < \infty$. For $\Phi \in \mathscr{F}_2(T; \mathscr{L}(H, K))$, we define

$$\int_0^t \Phi(s) \, dq(s) = \sum_{i=0}^{\infty} \int_0^t \Phi(s) e_i \, dq_i(s),$$

where $\{e_i\}$ is a complete orthonormal basis for H. Note that $\int_0^t \Phi(s) e_i \, dq_i(s)$ is a well-defined stochastic integral since q_i is a real orthogonal increments process (see [12]). $\int_0^t \Phi(s) \, dq(s) \in \mathscr{C}(T; \mathscr{L}_2(\Omega; K))$ is an H-valued martingale, with the following properties:

$$E\left\{\int_0^t \Phi(s) \, dq(s)\right\} = 0, \tag{1.11}$$

$$\begin{aligned} E\left\{\left\|\int_0^t \Phi(s) \, dq(s)\right\|^2\right\} &= \operatorname{trace} \int_0^t \Phi(s) M \Phi^*(s) \, df(s) \\ &\leqq \operatorname{trace} M \int_0^t \|\Phi(s)\|^2 \, df(s), \end{aligned} \tag{1.12}$$

$$E\left\{\int_0^t \Phi_1(\alpha) \, dq(\alpha) \circ \int_0^s \Phi_2(\alpha) \, dq(\alpha)\right\} = \int_0^{\min(t,s)} \Phi_1(\alpha) M \Phi_2^*(\alpha) \, df(\alpha), \tag{1.13}$$

$$E\left\{\int_{t_1}^{t_2} \Phi_1(\alpha)\, dq(\alpha) \circ \int_{t_3}^{t_4} \Psi_2(\alpha)\, dq_2(\alpha)\right\} = 0, \tag{1.14}$$

where $q_2(t)$ is an H_2-valued centered orthogonal increments process independent of $q(t)$, Φ_1, $\Phi_2 \in \mathscr{F}_2(T; \mathscr{L}(H, K))$ and $\Psi \in \mathscr{F}_2(T; \mathscr{L}(H_2, K_2))$.

In the special case, where $q(t) = w(t)$, the Wiener process, $f(t) = t$ and $\int_0^t \Phi(s)\, dw(s)$ has continued sample paths.

The stochastic integral can be easily extended to a general orthogonal increments process $\{q(t)\}$, by defining

$$\int_0^t \Phi(s)\, dq(s) = \int_0^t \Phi(s)\, d(q(s) - r(s)) + \sum_{i=0}^{\infty} \mu_i \int_0^t \Phi(s) e_i\, d\rho(s),$$

where $r(t) = E\{q(t)\} = \sum_{i=0}^{\infty} \mu_i e_i p(t)$ as in Definition 1.9.

A particular example is the Poisson process, where we have for $\Phi \in \mathscr{B}_2(T; \mathscr{L}(H, K))$[1],

$$\int_0^t \Phi(s)\, dp(s) = \int_0^t \Phi(s)\, d(p(s) - E\{p(s)\}) + \sum_{i=0}^{\infty} \mu_i \int_0^t \Phi(s) e_i\, ds.$$

The following stochastic Fubini theorem holds for a general orthogonal increments process.

THEOREM 1.2. *Let $\Phi(\cdot, \cdot): T \times T \to \mathscr{L}(H)$ be strongly measurable on $T \times T$ and such that $\int_{T \times T} \|\Phi(t, s)\|^2\, ds\, df(t) < \infty$. Then the following integrals exist and are equal w.p.*1:

$$\int_T \left(\int_T \Phi(t, s)\, ds\right) dq(t) = \int_T \left(\int_T \Phi(t, s)\, dq(t)\right) ds \quad w.p.1.$$

1.3. Stochastic evolution equations. Here we outline the results of [7] on stochastic evolution equations excited by orthogonal increments type noise processes. Using Definition 1.12 for stochastic integration with respect to the orthogonal increments process q, we consider

$$\begin{aligned} du(t) &= \mathscr{A}(t)u(t)\, dt + \Phi(t)\, dq(t) + g(t)\, dt, \\ u(0) &= u_0, \end{aligned} \tag{1.15}$$

where $\mathscr{A}(t)$ is the generator of an evolution operator $\mathscr{U}(t, s)$, $\Phi \in \mathscr{F}_2(T; \mathscr{L}(K, H))$, $g \in L_2(T \times \Omega; H)$, $u_0 \in L_2(\Omega; H)$ and $q(t)$ is a K-valued orthogonal increments process.

First we define the mild solution of (1.15) to be

$$u(t) = \mathscr{U}(t, 0)u_0 + \int_0^t \mathscr{U}(t, s)\Phi(s)\, dq(s) + \int_0^t \mathscr{U}(t, s)g(s)\, ds. \tag{1.16}$$

Even when $\mathscr{U}(t, s)$ is only a mild evolution operator, (1.16) is a well-defined stochastic process and $\langle h, u(t)\rangle$ is continuous in the mean-square on T for all $h \in H$.

[1] $\mathscr{B}_2(T; \mathscr{L}(H, K))$ is the space of strongly measurable $\mathscr{L}(H, K)$-valued functions such that $\int_T \|\Phi(s)\|^2\, ds < \infty$.

We need to impose extra conditions on $\mathcal{U}(t, s)$, Φ, u_0 and g in order that (1.16) be a strong solution of (1.15) in the following sense.

DEFINITION 1.13. Equations (1.15) have a strong solution u if $u \in \mathcal{C}(T; L_2(\Omega; H))$, $u(t) \in \mathcal{D}(\mathcal{A}(t))$ w.p. 1 and $u(t)$ satisfies (1.15) almost everywhere on $T \times \Omega$.

We say that u is unique if whenever u_1 and u_2 are strong solutions,

$$\mu\left\{\omega : \sup_{t \in T} \|u_1(t) - u_2(t)\| = 0\right\} = 1.$$

THEOREM 1.3. *If $\mathcal{A}(t)$ generates an almost strong evolution operator $\mathcal{U}(t, s)$, $g \in L_2(\Omega \times T; H)$, $B \in \mathcal{B}_2(T; \mathcal{L}(K, H))$, $u_0 \in L_2(\Omega; H)$, and the following extra assumptions are satisfied*:

$$\mathcal{U}(t, s)B(s)\, e_i \in \mathcal{D}(\mathcal{A}(t)) \quad \textit{for almost all } t > s \in T$$

and

$$\begin{aligned} &\sum_{i=0}^{\infty} \lambda_i \int_0^t \|\mathcal{A}(t)\mathcal{U}(t, r)B(r)e_i\|^2\, df(r) < \infty, \\ &\sum_{i=0}^{\infty} \mu_i \int_0^t \|\mathcal{A}(t)\mathcal{U}(t, r)B(r)e_i\|\, dp(r) < \infty, \end{aligned} \tag{1.17}$$

$$\mathcal{U}(t, s)g(s) \in \mathcal{D}(\mathcal{A}(s)) \quad \textit{for almost all } t > s \in T \tag{1.18}$$

and

$$\int_0^t \|\mathcal{A}(t)\mathcal{U}(t, s)g(s)\|\, ds < \infty \quad \textit{w.p.}1,$$

$$\mathcal{U}(t, 0)u_0 \in \mathcal{D}(\mathcal{A}(t)) \quad \textit{w.p.}1; \tag{1.19}$$

then (1.15) *has a unique strong solution given by* (1.16).

2. The abstract filtering problem. Motivated by possible applications to delay equations and partial differential equations excited by Poisson-or Gaussian-type white noise (see [8], [9]), we consider the following abstract signal and observation process:

$$\begin{aligned} u(t, \omega) &= \mathcal{U}(t, 0)u_0 + \int_0^t \mathcal{U}(t, \sigma)B(\sigma)\, dq(\sigma) \\ &= \mathcal{U}(t, 0)u_0 + \int_0^t \mathcal{U}(t, \sigma)B(\sigma)\, dr(\sigma) \\ &\quad + \int_0^t \mathcal{U}(t, \sigma)B(\sigma)\, dm(\sigma), \end{aligned} \tag{2.1}$$

$$z(t, \omega) = \int_0^t C(\tau)u(\tau)\, d\tau + \int_0^t F(\tau)\, dw(\tau), \tag{2.2}$$

where $\mathcal{U}(t, s)$ is a mild evolution operator on a Hilbert space, H,

$$B \in \mathcal{B}_\infty(T; \mathcal{L}(K, H)), \quad C \in \mathcal{B}_\infty(T; \mathcal{L}(H, R^k)), \quad F, F^{-1} \in L_\infty(T; \mathcal{L}(R^k)),$$

u_0 is an H-valued random variable with zero expectation and covariance operator P_0, w is a k-dimensional Wiener process with incremental covariance matrix W and q is a K-valued orthogonal increments process, with $r(t)=E\{q(t)\}=\sum_{i=0}^{\infty}\mu_i e_i \rho(t)$ and $m(t)=q(t)-r(t)$ has the covariance function $Mf(t)$ with $Me_i=\sum\lambda_{ij}e_i$. $\rho(t)$ and $f(t)$ are both monotone nondecreasing real functions, $\sum_{l=0}^{\infty}\mu_i^2<\infty$, $\sum_{l=0}^{\infty}\lambda_{ii}<\infty$. Furthermore, we assume that q, w and u_0 are mutually independent.

The assumption that the observation process is finite-dimensional is necessary because the incremental covariance operator of the observation noise, W, is both nuclear and invertible. However as one can only hope to make finite observations in practice, this is not a serious restriction.

From the properties of integrals with respect to orthogonal increments processes and of mild evolution operators, we have for $h\in H$, $\langle h, u(t)\rangle\in \mathscr{C}(T;L_2(\Omega))$. So (2.1), (2.2) is a well-defined system model.

The state estimation problem is then to find the best unbiased estimate of the state $u(t)$ at time t, based on the observation $z(s)$, $0\leqq s\leqq t_0$, which has the form

$$\hat{u}(t|t_0)=\int_0^{t_0}\mathscr{K}(t,s)\,dz\,(s,\omega)+v(t,t_0),$$

where $\mathscr{K}(t,\cdot)\in\mathscr{B}_2(0,t_0;\mathscr{L}(R^k,H))$ for each $t\in T$ and such that $E\{\langle h,\tilde{u}(t|t_0)\rangle^2\}$ is a minimum for all $h\in H$. $\tilde{u}(t|t_0)=u(t)-\hat{u}(t|t_0)$ is the error process and for an unbiased estimate,

$$v(t,t_0)=E\{u(t)\}-E\left\{\int_0^{t_0}\mathscr{K}(t,s)\,dz(s)\right\}.$$

For $t_0<t$ we have the smoothing problem, for $t_0>t$ the prediction problem and for $t_0=t$ the filtering problem. For the filtering problem we write $\hat{u}(t|t)=\hat{u}(t)$, $\tilde{u}(t|t)=\tilde{u}(t)$ and $v(t,t)=v(t)$. From (1.12) it is easily verified that $\hat{u}(t|t_0)$ is a well-defined stochastic process with $E\{\|\hat{u}(t|t_0)\|^2\}<\infty$.

We now establish a series of lemmas along the lines of those established in [4] and [8] for the Gaussian white noise case.

LEMMA 2.1. $\Lambda(t,s)=E\{(u(t)-E\{u(t)\})\circ(u(s)-E\{u(s)\})\}$ *is given by*

$$\Lambda(t,s)x=\mathscr{U}(t,0)P_0\mathscr{U}^*(s,0)x+\int_0^{\min(t,s)}\mathscr{U}(t,\tau)B(\tau)MB^*(\tau)\mathscr{U}^*(s,t)x\,df(\tau)$$

for each $x\in H$.

Proof. Direct substitution from (2.1) using the independence of and u_0, q, $E\{u_0\}=0$ and (1.13).

LEMMA 2.2. Orthogonal projections lemma. $\hat{u}(t|t_0)=\int_0^t\mathscr{K}(t,s)\,dz\,(s)+v(t,t_0)$ *is a solution to the estimation problem if*

$$E\{\tilde{u}(t|t_0)\circ z(\sigma)-z(\tau)\}=0\quad \textit{for all}\ \sigma,\tau\ \textit{such that}\ 0\leqq\tau\leqq\sigma\leqq t_0\leqq T.$$

Proof. For fixed $h\in H$, define the Hilbert space

$$X(h)=\left\{\begin{array}{c}\langle u,h\rangle: u\in L_2(\Omega;H)\ \text{with inner product}\\ [\langle u,h\rangle,\langle v,h\rangle]=E\{\langle u,h\rangle\langle u,h\rangle\}\end{array}\right\}.$$

Note that $E\{u \circ v\} = 0$ iff $[\langle u, h\rangle, \langle v, h\rangle] = 0$ for all $h \in H$. For fixed t_0, define the subspace

$$X(h; t_0, t) = \left\{ \begin{array}{l} \langle y(y, t_0), h\rangle \text{ where } y(t, t_0) = \int_0^{t_0} B(t, s)\, dz(s) \\ B(t, \cdot) \in \mathscr{B}_2(0, t_0; \mathscr{L}(R^k, H)) \text{ for all } t \leqq t_0 \end{array} \right\}.$$

Then $\langle \tilde{u}(t|t_0), h\rangle$ is in the manifold $\langle u(t) - v(t, t_0), h\rangle + X(h; t_0, t)$. We seek to minimize $\tilde{u}(t|t_0) = u(t) - \hat{u}(t|t_0)$ in the $X(h)$ norm for all h. By the orthogonal projections lemma for Hilbert spaces this is equivalent to requiring $\langle \tilde{u}(t|t_0), h\rangle \perp X(h; t_0, t)$ in $X(h)$, i.e., $E\{\langle h, \tilde{u}(t|t_0)\rangle\langle h, y(t, t_0)\rangle\} = 0$ for all $\langle h, g(t, t_0)\rangle \in X(h; t_0, t)$. So it remains to establish that

$$E\{\tilde{u}(t|t_0) \circ y(t, t_0)\} = 0 \quad \text{if and only if} \tag{2.3}$$

$$E\{\tilde{u}(t|t_0) \circ z(\sigma) - z(\tau)\} = 0 \quad \text{for } 0 \leqq \tau \leqq \sigma \leqq t_0. \tag{2.4}$$

Supposing that (2.4) holds, then it is easily verified that (2.3) holds when $y(t, t_0) = \int_0^{t_0} B_0(t, s)\, dz(s)$, where $B_0(t, s)$ is a step function in the s variable: $(E\{\tilde{u}(t|t_0) \circ B_0(z(\sigma) - z(\tau))\} = E\{\tilde{u}(t|t_0) \circ z(\sigma) - z(\tau)\}B_0^*)$. For general $B(t, \cdot)$, we approximate it by a sequence of step functions $\{B_n(t, \cdot)\}$ such that $\int_0^{t_0} \|B(t, s) - B_2(t, s)\|^2\, ds \to 0$ as $n \to \infty$, and in the usual way extend the result to general $B(t, \cdot) \in \mathscr{B}_2(0, t_0; \mathscr{L}(R^k, H)$, (see [4]).

Conversely, suppose that (2.3) holds but

$$E\{\tilde{u}(t|t_0) \circ z(\sigma) - z(\tau)\} \neq 0 \quad \text{for some } \sigma, \tau.$$

Defining

$$B_0(t, s) = \left\{ \begin{array}{ll} E\{\tilde{u}(t|t_0) \circ (z(\sigma) - z(\tau))\} & \text{for } \tau \leqq s \leqq \sigma, \\ 0 & \text{otherwise,} \end{array} \right.$$

$$\int_0^{t_0} \|B_0(t, s)\|^2\, ds = (\sigma - \tau)\|E\{\tilde{u}(t|t_0) \circ (z(\sigma) - z(\tau))\}\|^2$$

$$\leqq (\sigma - \tau)E\{\|\tilde{u}(t|t_0)\|^2\}E\{\|z(\sigma) - z(\tau)\|^2\}$$

(by the Schwarz inequality)

$$< \infty,$$

so $y(t, t_0) = \int_0^{t_0} B_0(t, s)\, dz(s)$ is such that $\langle h, y(t, t_0)\rangle \in X(h; t, t_0)$. But

$$\begin{aligned} \langle h, E\{\tilde{u}(t|t_0) \circ y(t, t_0)\}h\rangle &= \langle h, E\{\tilde{u}(t|t_0) \circ z(\sigma) - z(\tau)\}E\{\tilde{u}(t|t_0) \circ z(\sigma) \\ &\quad - z(\tau)\}^* h\rangle \\ &= \|E\{\tilde{u}(t|t_0) \circ z(\sigma) - z(\tau)\}^* h\|^2 \\ &\neq 0 \quad \text{for some } h \in H. \end{aligned}$$

So $E\{\tilde{u}(t|t_0) \circ y(t|t_0)\} \neq 0$ and the lemma is established.

COROLLARY 2.1. $E\{\tilde{u}(t|t_0) \circ \hat{u}(\sigma)\} = 0$ *for* $\sigma < t_0$.

As in infinite-dimensional estimation theory we define the innovations process by

$$\gamma(t, \omega) = z(t, \omega) - \int_0^t C(s)\hat{u}(s)\, ds, \tag{2.5}$$

and we also define the second innovation process

$$\gamma_0(t, \omega) = z(t, \omega) - \int_0^t C(s)(\hat{u}(s) - v(s))\, ds.$$

$\gamma_0(t, w)$ is also a k-dimensional stochastic process, and in the following sense it is equivalent to the observation process.

LEMMA 2.3. *For all $M \in V = \mathcal{B}_2(0, t; \mathcal{L}(R^k, H))$ there exists a unique $N \in V$ such that*

$$\int_0^t N(s)\, d\gamma_0(s, \omega) = \int_0^t M(s)\, dz(s, \omega),$$

and conversely.

Proof. (This is essentially the proof from [2].) Take $N \in V$. Then

$$\begin{aligned}\int_0^t N(s)\, d\gamma_0(s) &= \int_0^t N(s)\, dz(s, \omega) - \int_0^t N(s)C(s)(\hat{u}(s) - v(s))\, ds \\ &= \int_0^t N(s)\, dz(s) - \int_0^t N(s)C(s) \int_0^s \mathcal{K}(s, \alpha)\, dz(\alpha)\, ds \\ &= \int_0^t N(s)\, dz(s) - \int_0^t \int_\alpha^t N(s)C(s)\mathcal{K}(s, \alpha)\, ds\, dz(\alpha),\end{aligned}$$

interchanging the order of integration by Theorem 1.2.

Given $M \in V$, consider the integral equation

$$M(\alpha)x = N(\alpha)x - \int_\alpha^t N(s)C(s)\mathcal{K}(s, \alpha)x\, ds \quad \text{for } x \in R^k, \tag{2.6}$$

or equivalently,

$$N^*(\alpha)h = M^*(\alpha)h + \int_\alpha^t \mathcal{K}^*(s,\alpha)C^*(s)N^*(s)h\, ds \quad \text{for } h \in H.$$

This is a Volterra integral equation for $N^*(s)h \in L_2(0, t; R^k)$, and since $\int_0^t \int_\alpha^t \|\mathcal{K}^*(s, \alpha)C^*(s)\|^2\, ds\, d\alpha < \infty$, it has a unique solution $N^*(s)h$ for any given $M^*(\alpha)h$. So given $M \in V$, there is a unique $N \in V$ satisfying (2.6) and hence $\int_0^t N(s)\, d\gamma_0(s) = \int_0^t M(s)\, dz(s)$. The converse is proved similarly.

COROLLARY 2.2. *The optimal estimator, if it exists, is also given by*

$$\hat{u}(t|t_0) = v(t, t_0) + \int_0^{t_0} G(t, s)\, d\gamma_0(s, \omega)$$

for some $G(t, \cdot) \in \mathcal{B}_2(0, t_0; \mathcal{L}(R^k, H))$.

COROLLARY 2.3. *$\hat{u}(t|t_0)$ is a solution to the estimation problem if $E\{\tilde{u}(t|t_0) \circ \gamma_0(\sigma) - \gamma_0(\tau)\} = 0$ for all σ, τ such that $0 \leqq \tau \leqq \sigma \leqq t_0$.*

We state the following result proved in [2], [5] and [8].

LEMMA 2.4. Wiener–Hope equation. *Under the assumptions of our problem, the following integral equation has a unique solution* $\mathcal{K}(t, \cdot) \in \mathcal{B}_2(0, t; \mathcal{L}(R^k, H))$ *for* $t \in T$:

$$\int_0^t \mathcal{K}(t, s)C(s)\Lambda(s, \sigma)C^*(\sigma)x\, ds + K(t, \sigma)F(\sigma)WF^*(\sigma)x = \Lambda(t, \sigma)C^*(\sigma)x \tag{2.7}$$

for $x \in H$.

We now consider the filtering problem.

LEMMA 2.5. *There is a unique solution to our filtering problem under the stated assumptions.*

Proof. (This is an extension of the proof in [5]. We show that (2.7) has a solution if and only if there is an optimal filter $\hat{u}(t) = v(t) + \int_0^t \mathcal{K}(t, s)\, dz(s)$ and then Lemma 2.3 guarantees the existence of an optimal filter. Suppose there is a $\mathcal{K}_0(t, \cdot) \in \mathcal{B}_2(0, t; \mathcal{L}(R^k, H))$ such that .

$$\hat{u}(t) = v(t) + \int_0^t \mathcal{K}_0(t, s)\, dz(s) \tag{2.8}$$

is optimal. Let

$$\begin{aligned} y(\sigma) &= \int_0^\sigma C(s)(u(s) - E\{u(s)\})\, ds \\ &= z(\sigma) - z(0) - \int_0^\sigma F(s)\, dw(s) - \int_0^\sigma C(s)E\{u(s)\}\, ds. \end{aligned} \tag{2.9}$$

Then

$$\frac{d}{d\sigma}E\{\tilde{u}(t) \circ y(\sigma)\} = E\left\{(u(t) - v(t) - \int_0^t \mathcal{K}(t, s)\, dz(s)) \circ (u(\sigma) - E\{u(\sigma)\})\right\}C^*(\sigma)$$

a.e. since $\sigma < t$

$$\begin{aligned} &= E\{(u(t) - v(t)) \circ (u(\sigma) - E\{u(\sigma)\})\}C^*(\sigma) \\ &\qquad - E\left\{\int_0^t \mathcal{K}(t, s)\, dz(s) \circ u(\sigma) - E\{u(\sigma)\}\right\}C^*(\sigma) \\ &= \Lambda(t, \sigma)C^*(\sigma) - \int_0^t \mathcal{K}(t, s)C(s)\Lambda(s, \sigma)C^*(\sigma)\, ds, \end{aligned} \tag{2.10}$$

substituting for $z(s)$ from (2.2) and using Lemma 2.1 and the independence of u_0, q and w. But

$$E\{\tilde{u}(t) \circ y(\sigma)\} = E\left\{\tilde{u}(t) \circ \int_0^\sigma F(s)\, dw(s)\right\} - E\left\{\tilde{u}(t) \circ \int_0^\sigma C(s)E\{u(s)\}\, ds\right\}$$

(by Lemma 2.2)

$$= E\left\{\int_0^t \mathcal{K}(t, s)\, dz(s) \circ \int_0^\sigma F(s)\, dw(s)\right\},$$

substituting for $\tilde{u}(t)$ and using the independence of q, u_0 and w and

$$E\{\tilde{u}(t)\} = 0$$

$$= \int_0^\sigma \mathcal{K}(t, s)F(s)WF^*(s)\, ds \quad \text{(by (1.13))}.$$

So

$$\frac{d}{d\sigma}E\{\tilde{u}(t) \circ y(\sigma)\} = \mathcal{K}(t, \sigma)F(\sigma)WF^*(\sigma)$$

$$= \Lambda(t, \sigma)C^*(\sigma) - \int_0^t \mathcal{K}(t, s)C(s)\Lambda(s, \sigma)C^*(\sigma)\, ds$$

from (2.10), and so $\mathcal{K}(t, \cdot)$ is a solution to (2.3).

Suppose now that (2.7) has a solution $\mathcal{K}(t, \cdot)$. We show that $\hat{u}(t) = v(t) + \int_0^t \mathcal{K}(t, s)\, dz(s)$ satisfies Lemma 2.2. From the linearity we may assume $\tau = 0$. Now $E\{\tilde{u}(t) \circ z(\sigma) - z(0)\} = E\{\tilde{u}(t) \circ y(\sigma)\} + E\{\tilde{u}(t) \circ \int_0^\sigma F(s)\, dw(s)\}$ since $E\{\tilde{u}(t) = 0\}$, where y satisfies (2.9):

$$= \int_0^\sigma \left[\Lambda(t, \alpha)C^*(\alpha) - \int_0^t \mathcal{K}(t, s)C(s)\Lambda(s, \alpha)C^*(\alpha)\, ds\right] d\alpha$$

$$- \int_0^\sigma \mathcal{K}(t, s)F(s)WF^*(s)\, ds$$

from (2.10) and expanding $\hat{u}(t)$ using the independence of u_0, p and w and (1.11), (1.13):

$$= 0$$

since $\mathcal{K}(t, \cdot)$ satisfies (2.7).

We now quote a result from [8] which will enable us to establish recursive formulas defining the optimal filter.

THEOREM 2.1. *Under the stated assumptions the following integral equations are equivalent and have a unique solution in the class of self-adjoint positive operator functions in* $\mathcal{B}_\infty(T; \mathcal{L}(H))$:

$$P(t)x = \mathcal{Y}(t, 0)P_0\mathcal{U}^*(t, 0)x + \int_0^t \mathcal{Y}(t, s)B(s)MB^*(s)\mathcal{U}^*(t, s)x\, ds, \quad x \in H, \tag{2.11}$$

$$P(t)x = \mathcal{Y}(t, 0)P_0\mathcal{Y}^*(t, 0)x + \int_0^t \mathcal{Y}(t, s)(B(s)MB^*(s) + P(s)C^*(s)(F(s)WF^*(s))^{-1}C(s)P(s))\mathcal{Y}^*(t, s)x\, ds, \tag{2.12}$$

where $\mathcal{Y}(t, s)$ *is the perturbation of the mild evolution operator* $\mathcal{U}(t, s)$ *corresponding to the perturbation* $-P(t)C^*(t)(F(t)WF^*(t))^{-1}C(t)$. *Furthermore,* $\mathcal{K}(t, s) = \mathcal{Y}(t, s)P(s)C^*(s)(F(s)WF^*(s))^{-1}$ *is the unique solution of* (2.7).

So from Lemma 2.5 we see that there is a unique optimal filter

$$\hat{u}(t) = v(t) + \int_0^t \mathcal{Y}(t, s)P(s)C^*(s)(F(s)WF^*(s))^{-1}\, dz(s)$$

and

$$v(t) = E\{u(t)\} - E\left\{\int_0^t \mathcal{K}(t,s)\,dz(s)\right\}$$

$$= \int_0^t \mathcal{U}(t,\sigma)B(\sigma)\,dr(\sigma) - E\left\{\int_0^t \mathcal{K}(t,s)C(s)u(s)\,ds\right\} \quad \text{(by (1.11), (2.1))}$$

$$= \int_0^t \mathcal{U}(t,\sigma)B(\sigma)\,dr(\sigma)$$
$$- \int_0^t \mathcal{K}(t,s)C(s)\int_0^s \mathcal{U}(s,\sigma)B(\sigma)\,dr(\sigma)\,ds \quad \text{(by (1.11), (2.1))}$$

$$= \int_0^t \Big[\mathcal{U}(t,\sigma) - \int_\sigma^t \mathcal{Y}(t,s)P(s)C^*(s)$$
$$\cdot (F(s)WF^*(s))^{-1}C(s)\mathcal{U}(s,\sigma)\,ds\Big]B(\sigma)\,dr(\sigma)$$

(interchanging the order of integration)

$$= \int_0^t \mathcal{Y}(t,\sigma)B(\sigma)\,dr(\sigma) \quad \text{(by Theorem 1.1).}$$

Using Lemma 2.3 we may also express the filter in terms of the innovations process.

LEMMA 2.6. *The unique optimal filter is given by*

$$\hat{u}(t) = \int_0^t \mathcal{Y}(t,\sigma)B(\sigma)\,dr(\sigma) + \int_0^t \mathcal{Y}(t,s)P(s)C^*(s)(F(s)WF^*(s))^{-1}\,dz(s)$$

$$= \int_0^t \mathcal{U}(t,\sigma)B(\sigma)\,dr(\sigma) + \int_0^t \mathcal{U}(t,s)P(s)C^*(s)(F(s)WF^*(s))^{-1}\,dv(s).$$

Proof. By Lemma 2.3, there exists a unique $G(t,\cdot) \in \mathcal{B}_2(0,t;\mathcal{L}(R^k,H))$ such that

$$\int_0^t \mathcal{K}(t,s)\,dz(s) = \int_0^t G(t,s)\,d\gamma_0(s)$$

where G satisfies

$$\mathcal{K}(t,s)x = G(t,s) - \int_s^t G(t,\alpha)C(\alpha)\mathcal{K}(\alpha,s)x\,d\alpha \quad \text{for } x \in R^K.$$

Letting $\mathcal{K}(t,s) = \mathcal{Y}(t,s)P(s)C^*(s)(F(s)WF^*(s))^{-1}$ we can verify that $G(t,s) = \mathcal{U}(t,s)P(s)C^*(s)(F(s)WF^*(s))^{-1}$ is the solution, since

$$\mathcal{Y}(t,s)x - \mathcal{U}(t,s)x + \int_s^t \mathcal{U}(t,\alpha)P(\alpha)C^*(\alpha)(F(s)WF^*(s))^{-1}C(\alpha)\mathcal{Y}(\alpha,s)x\,d\alpha = 0$$

from the definition of $\mathcal{Y}(t, s)$ as a perturbed mild evolution operator (see Theorem 1.1). So

$$\begin{aligned}\hat{u}(t) &= \int_0^t \mathcal{Y}(t, \sigma)B(\sigma)\, dr(\sigma) + \int_0^t \mathcal{U}(t, s)P(s)C^*(s)(F(s)WF^*(s))^{-1}\, d\gamma_0(s) \\ &= \int_0^t \mathcal{Y}(t, \sigma)B(\sigma)\, dr(\sigma) - \int_0^t \mathcal{U}(t, s)P(s)C^*(s) \\ &\qquad \cdot (F(s)WF^*(s))^{-1}[v(s)\, ds - d\gamma(s)] \\ &= \int_0^t \mathcal{U}(t, \sigma)B(\sigma)\, dr(\sigma) + \int_0^t \mathcal{U}(t, s)P(s)C^*(s)(F(s)WF^*(s))^{-1}\, d\gamma(s)\end{aligned}$$

again since $\mathcal{Y}(t, s)$ is a perturbed mild evolution operator.

The unique solution $P(t)$ of the Riccati equation (2.11) is the covariance operator of the error process.

LEMMA 2.7.

(i) $E\{\tilde{u}(t) \circ \tilde{u}(t)\} = P(t)$,

(ii) $E\{\tilde{u}(t) \circ \tilde{u}(s)\} = P(t, s) = \begin{cases} \mathcal{Y}(t, s)P(s) & \text{for } t > s, \\ P(t)\mathcal{Y}^*(s, t) & \text{for } s > t. \end{cases}$

Proof.

(i) is proved exactly as in [8].

(ii) Suppose $t > s$. Then

$$u(t) = \mathcal{U}(t, s)u(s) + \int_s^t \mathcal{U}(t, \alpha)B(\alpha)\, dq(\alpha)$$

and by Theorem 2.1,

$$\begin{aligned}\hat{u}(t) = \mathcal{Y}(t, s)\hat{u}(s) &+ \int_s^t \mathcal{Y}(t, \alpha)P(\alpha)C^*(\alpha) \\ &\cdot (F(\alpha)WF^*(\alpha))^{-1}[C(\alpha)u(\alpha)\, d\alpha - F(\alpha)\, dw(\alpha)].\end{aligned}$$

But

$$\mathcal{U}(t, s)x = \mathcal{Y}(t, s)x + \int_s^t \mathcal{Y}(t, \alpha)P(\alpha)C^*(\alpha)(F(\alpha)WF^*(\alpha))^{-1}C(\alpha)\mathcal{U}(\alpha, s)x\, d\alpha.$$

So

$$\begin{aligned}\tilde{u}(s) = u(s) - \int_0^s \mathcal{Y}(s, \alpha)P(\alpha)C^*(\alpha) \\ \cdot (F(\alpha)WF^*(\alpha))^{-1}[C(\alpha)u(\alpha)\, d\alpha - F(\alpha)\, dw(\alpha)].\end{aligned}$$

Hence

$$\tilde{u}(t) = \mathcal{Y}(t, s)\tilde{u}(s) + \int_s^t \mathcal{U}(t, \alpha)B(\alpha)\, dq(\alpha) - \int_s^t F(\alpha)\, dw(\alpha)$$

and

$$E\{\tilde{u}(t) \circ \tilde{u}(s)\} = \mathcal{Y}(t, s)E\{\tilde{u}(s) \circ \tilde{u}(s)\}$$

since q and w have independent increments. Similarly,

$$P(t, s) = P(t)\mathcal{Y}^*(s, t) \quad \text{for } s > t.$$

In the Gaussian white noise case with u_0 Gaussian and q a Wiener process, it is proved in [8] that $\gamma(t)$ is a martingale relative to $\mathcal{Z}t = \sigma\{z(s); 0 \leqq s \leqq t\}$ and has the representation

$$\gamma(t) = \int_0^t F(\alpha)\, dv(\alpha),$$

where $\{v_t, \mathcal{Z}t\}$ is a k-dimensional Wiener process with incremental covariance matrix W. The proof relies crucially on the fact that $\hat{u}(t) = E\{u(t)|\mathcal{Z}t\}$ or the linear optimal filter is also the best global filter, a property of Gaussian estimators. Here we need a different approach.

LEMMA 2.8. *The innovations process $\{\gamma(t)\}$ is a martingale relative to the observation field $\mathcal{Z}t$ and has the representation*

$$\gamma(t) = \int_0^t F(\alpha)\, dv(\alpha),$$

where $\{v(t), \mathcal{Z}t\}$ is a k-dimensional orthogonal increments process with incremental covariance matrix W.

Proof. Let $v(t) = \int_0^t F^{-1}(\alpha)\, d\gamma(\alpha)$, which is well-defined expanding γ by (2.5).

(a) $v(t)$ has orthogonal increments and zero mean.

We shall instead prove the equivalent result:

$$E\{\gamma(s) \circ \gamma(t) - \gamma(s)\} = E\left\{\gamma(s) \circ \int_s^t F(\alpha)\, dw(\alpha) + \int_s^t C(\alpha)\tilde{u}(\alpha)\, d\alpha\right\}$$

$$= E\left\{\gamma(s) \circ \int_s^t F(\alpha)\, dw(\alpha)\right\} + E\left\{\gamma_0(s) \circ \int_s^t C(\alpha)\tilde{u}(\alpha)\, d\alpha\right\}$$

since $E\{\tilde{u}(\alpha)\} = 0$

$$= E\left\{\gamma(s) \circ \int_s^t F(\alpha)\, dw(\alpha)\right\} \qquad \text{(by Corollary 2.3)}$$

$$= E\left\{\int_0^s C(\alpha)\tilde{u}(\alpha)\, d\alpha + \int_0^s F(\alpha)\, dw(\alpha) \circ \int_s^t F(\alpha)\, dw(\alpha)\right\}$$

$$= 0,$$

since w has independent increments and since w, u_0 are independent. Also

$$E\{\gamma(s)\} = E\left\{\int_s^t C(s)\tilde{u}(s)\, ds\right\} + E\left\{\int_0^t F(s)\, dw(s)\right\}$$

$$= 0$$

since $E\{\tilde{u}(s)\}=0$, and by (1.11),

$$\text{(b)}\quad E\{v(t)\circ v(t)\}=E\Big\{\int_0^t dw(\alpha)\circ\int_0^t dw(\alpha)\Big\}$$
$$+E\Big\{\int_0^t dw(\alpha)\circ\int_0^t F^{-1}(\alpha)C(\alpha)\tilde{u}(\alpha)\,d\alpha\Big\}$$
$$+E\Big\{\int_0^t F^{-1}(\alpha)C(\alpha)\tilde{u}(\alpha)\,d\alpha\circ\int_0^t dw(\alpha)\Big\}$$
$$+E\Big\{\int_0^t F^{-1}(\alpha)C(\alpha)\tilde{u}(\alpha)\,d\alpha\circ\int_0^t F^{-1}(\alpha)C(\alpha)\tilde{u}(\alpha)\,d\alpha\Big\}$$
$$=Wt-E\Big\{\int_0^t dw(\alpha)\circ\int_0^t F^{-1}(\alpha)C(\alpha)\int_0^\alpha \mathcal{K}(\alpha,\beta)\,dz(\beta)\,d\alpha\Big\}$$
$$-E\Big\{\int_0^t F^{-1}(\alpha)C(\alpha)\int_0^\alpha \mathcal{K}(\alpha,\beta)\,dz(\beta)\,d\alpha\circ\int_0^t dw(\alpha)\Big\}$$
$$+\int_0^t\int_0^t F^{-1}(\beta)C(\beta)E\{\tilde{u}(\beta)\circ\tilde{u}(\alpha)\}C^*(\alpha)F^{-1*}(\alpha)\,d\beta\,d\alpha;$$

by (1.11) and since w is independent of u_0 and q,

$$=Wt-E\Big\{\int_0^t dw(\beta)\circ\int_0^t\int_\beta^t F^{-1}(\alpha)C(\alpha)\mathcal{K}(\alpha,\beta)F(\beta)\,d\alpha\,dw(\beta)\Big\}$$
$$-E\Big\{\int_0^t\int_\beta^t F^{-1}(\alpha)C(\alpha)\mathcal{K}(\alpha,\beta)F(\beta)\,d\alpha\,dw(\beta)\circ\int_0^t dw(\beta)\Big\}$$
$$+\int_0^t\Big[\int_0^\beta+\int_\beta^t F^{-1}(\beta)C(\beta)\mathcal{K}(\beta,\alpha)C^*(\alpha)F^{*-1}(\alpha)\,d\alpha\Big]\,d\beta;$$

since w is independent of $u(\cdot)$ and interchanging the order of integration,

$$=Wt-\int_0^t\Big[W\Big(\int_\beta^t F^{-1}(\alpha)C(\alpha)\mathcal{K}(\alpha,\beta)F(\beta)\,d\alpha\Big)^*$$
$$+\int_\beta^t F^{-1}(\alpha)C(\alpha)\mathcal{K}(\alpha,\beta)F(\beta)\,d\alpha W\Big]\,d\beta$$
$$+\int_0^t F^{-1}(\beta)C(\beta)\Big[\int_0^\beta \mathcal{Y}(\beta,\alpha)P(\alpha)C^*(\alpha)F^{-1*}(\alpha)\,d\alpha$$
$$+\int_\beta^t p(\beta)\mathcal{Y}^*(\alpha,\beta)C^*(\alpha)F^{-1*}(\alpha)\,d\alpha\Big]\,d\beta;$$

by (1.11) and Lemma 2.7,

$$=Wt.$$

Substituting for $\mathcal{K}(t,s)$ from Theorem 2.1 and interchanging the order of integration once more, then $v(t)$ is a centered orthogonal increments process of Definition 1.9 with incremental covariance matrix Wt.

This is a rigorous proof of a claim made by Kailath in [14] in the finite-dimensional case that the innovations process was a Gaussian white noise process of the same type as $w(t)$.

COROLLARY 2.4.

$$E\left\{\int_0^t M(\alpha)\,d\gamma(\alpha)\circ\int_0^s N(\alpha)\,dv(\alpha)\right\}x=\int_0^{\min(t,s)} M(\alpha)F(\alpha)WF^*(\alpha)N^*(\alpha)x\,d\alpha$$

for $x\in H$, $M, N\in B_2(T;\mathscr{L}(R^k,H))$.

Proof. Lemma 2.8 and (1.13) for the orthogonal increments process v.

We can now establish our smoothing result which is analagous to that in [8].

THEOREM 2.2. *Consider the smoothing problem for* (2.1) *and* (2.2) *under the stated assumptions and given* $t_0>t$. *Then the best smoothed estimate is given by*

$$\hat{u}(t|t_0)=\hat{u}(t)+P(t)\lambda(t), \tag{2.13}$$

where $\hat{u}(t)$ *is the optional filter at time* t *and* $P(t)$ *is the unique solution of* (2.11) *and*

$$\lambda(t)=\int_t^{t_0}\mathscr{Y}^*(s,t)C^*(s)(F(s)VF^*(s))^{-1}\,d\gamma(s). \tag{2.14}$$

Proof. By Corollary 2.2,

$$\hat{u}(t|t_0)=v(t,t_0)+\int_0^{t_0}G(t,s)\,d\gamma_0(s) \tag{2.15}$$

for some

$$G(t,\cdot)\in\mathscr{B}_2(0,t_0;\mathscr{L}(R^k,H))$$

and

$$\hat{u}(t|t_0)=a(t,t_0)+\int_0^{t_0}G(t,s)\,d\gamma(s), \tag{2.16}$$

where

$$a(t,t_0)=v(t,t_0)+\int_0^{t_0}C(s)v(s)\,ds.$$

Suppose $t\leqq s\leqq t_0$. Then

$$E\{u(t)\circ\gamma(s)\}=E\{u(t)\circ\gamma_0(s)\}-E\left\{u(t)\circ\int_0^s C(s)v(s)\,ds\right\}\quad\text{(by (2.5))}$$

$$=E\{\hat{u}(t)\circ\gamma_0(s)\}-E\left\{u(t)\circ\int_0^s C(s)v(s)\,ds\right\}\quad\text{(by Corollary 2.3)}$$

$$=E\left\{\int_0^{t_0}G(t,s)\,d\gamma(s)\circ\gamma(s)\right\}+E\{a(t,t_0)\circ\gamma(s)\}$$

$$+E\left\{\hat{u}(t|t_0)\circ\int_0^s C(\alpha)v(\alpha)\,d\alpha\right\}$$

$$-E\{u(t)\}\circ\int_0^s C(\alpha)v(\alpha)\,d\alpha\quad\text{(by (2.16))}$$

$$= E\left\{\int_0^t G(t, s)\, d\gamma(s) \circ \gamma(s)\right\} \quad (\text{since } E\{\tilde{u}(t)\} = 0,\ E\{\gamma(s)\} = 0)$$

$$= \int_0^s G(t, \alpha) F(\alpha) WF^*(\alpha)\, d\alpha \quad (\text{by Corollary 2.4}).$$

Therefore

$$\frac{\partial}{\partial s} E\{u(t) \circ \gamma(s)\} = G(t, s) F(s) WF^*(s).$$

But

$$E\{u(t) \circ \gamma(s)\} = E\left\{u(t) \circ \int_0^s C(\alpha)\tilde{u}(\alpha)\, d\alpha + \int_0^s F(\alpha)\, dw(\alpha)\right\}$$

$$= E\left\{u(t) \circ \int_0^s C(\alpha)\hat{u}(\alpha)\, d\alpha\right\}$$

since q, w and u_0 are independent.
Therefore

$$\frac{\partial}{\partial s} E\{u(t) \circ \gamma(s)\} = E\{u(t) \circ C(s)\tilde{u}(s)\}$$

$$= E\{\tilde{u}(t) + \hat{u}(t) \circ \tilde{u}(s)\} C^*(s)$$

$$= E\{\tilde{u}(t) \circ \tilde{u}(s)\} C^*(s)$$

by Corollary 2.1 since $t_0 > s > t$.

So

$$\hat{u}(t|t_0) = v(t, t_0) + \int_0^{t_0} G(t, s)\, d\gamma_0(s)$$

$$= v(t, t_0) + \int_0^t G(t, s)\, d\gamma_0(s) + \int_t^{t_0} P(t)\mathcal{Y}^*(s, t) C^*(s) \cdot (F(s) WF^*(s))^{-1}\, d\gamma_0(s)$$

$$= \hat{u}(t) + v(t, t_0) - v(t) + \int_t^{t_0} P(t)\mathcal{Y}^*(s, t) C^*(s)(F(s) WF^*(s))^{-1}\, d\gamma_0(s)$$

$$= \hat{u}(t) - E\left\{\int_{t_0}^t P(t)\mathcal{Y}^*(s, t) C^*(s)(F(s) WF^*(s))^{-1}\, d\gamma_0(s)\right\} + \int_t^{t_0} P(t)\mathcal{Y}^*(s, t) C^*(s)(F(s) WF^*(s))^{-1}\, d\gamma_0$$

$$= \hat{u}(t) + \int_{t_0}^t P(t)\mathcal{Y}^*(s, t) C^*(s)(F(s) WF^*(s))^{-1}\, d\gamma(s) \quad \text{by (2.5)}.$$

THEOREM 2.3. *Consider the prediction for* (2.1), (2.2) *under the stated assumptions and given* $t_0 < t$. *Then the best predictor is given by*

$$\hat{u}(t|t_0) = E\{u(t)\} + \mathcal{U}(t, t_0)\hat{u}(t_0) \quad \text{for } t > t_0. \tag{2.17}$$

Proof. By Corollary 2.2,

$$\hat{u}(t|t_0) = v(t, t_0) + \int_0^{t_0} G(t, s)\, d\gamma_0(s) \tag{2.15}$$

for some

$$G(t, \cdot) \in \mathcal{B}_2(0, t_0; \mathcal{L}(R^k, H)).$$

Now

$$\begin{aligned} E\{u(t) \circ \gamma(s)\} &= E\{u(t) \circ \gamma_0(s)\} - E\left\{u(t) \circ \int_0^s C(\alpha)v(\alpha)\, d\alpha\right\} \quad \text{(by (2.5))} \\ &= E\{\hat{u}(t|t_0) \circ \gamma_0(s)\} - E\left\{u(t) \circ \int_0^s C(\alpha)v(\alpha)\, d\alpha\right\} \quad \text{(by Corollary 2.3)} \\ &= E\left\{\int_0^{t_0} G(t, s)\, d\gamma(s, w) \circ \gamma(s, w)\right\} + E\{a(t, t_0) \circ \gamma(s, w)\} \\ &\quad + E\left\{\hat{u}(t|t_0) \circ \int_0^s C(\alpha)v(\alpha)\, d\alpha\right\} \\ &\quad - E\left\{u(t) \circ \int_0^s C(\alpha)v(\alpha)\, d\alpha\right\} \quad \text{(by (2.15))} \\ &= \int_0^s G(t, \alpha)F(\alpha)WF^*(\alpha)\, d\alpha \end{aligned}$$

(by Corollary 2.4 and since $E\{\tilde{u}(t)\} = 0$, $E\{\gamma(t)\} = 0$).

Therefore

$$G(t, s)F(s)WF^*(s) = \frac{\partial}{\partial s} E\{u(t) \circ \gamma(s)\} \quad \text{for } s < t_0 < t.$$

But

$$\begin{aligned} E\{u(t) \circ \gamma(s)\} &= E\left\{u(t) \circ \int_0^s C(\alpha)\tilde{u}(\alpha)\, d\alpha + \int_0^s F(\alpha)\, dw(\alpha)\right\} \\ &= E\left\{u(t) \circ \int_0^s C(\alpha)\tilde{u}(\alpha)\, d\alpha\right\} \end{aligned}$$

since u_0, q and w are independent. Therefore

$$\frac{\partial}{\partial s} E\{u(t) \circ \gamma(s)\} = E\{u(t) \circ \tilde{u}(s)\}C^*(s).$$

But

$$u(t) = \mathcal{U}(t, s)u(s) + \int_s^t \mathcal{U}(t, \alpha)B(\alpha)\, dq(\alpha) \quad \text{for } t > s.$$

Hence

$$E\{u(t) \circ \tilde{u}(s)\} = \mathcal{U}(t, s)E\{u(s) \circ \tilde{u}(s)\}$$

since q is independent of u_0 and w and has orthogonal increments

$$G(t, s) = \mathcal{U}(t, s)P(s)C^*(s)(F(s)WF^*(s))^{-1} \quad (\text{for } s \leqq t_0 < t)$$
$$= \mathcal{K}(t, s).$$

Therefore

$$\hat{u}(t|t_0) = v(t, t_0) + \int_0^{t_0} \mathcal{K}(t, s)\, d\gamma_0(s)$$
$$= v(t, t_0) + \int_0^{t_0} \mathcal{K}(t, s)\, d\gamma(s) + \int_0^t \mathcal{K}(t, s)C(s)E\{u(s)\}\, ds \quad (\text{by } (2.5))$$
$$= E\{u(t)\} + \mathcal{U}(t, t_0)\hat{u}(t_0).$$

As in [8] we can show that the optimal estimate $\hat{u}(t|t_0)$ is the best linear estimate of $u(t)$ based on $z(s)$, $0 \leqq s \leqq t_0$, in the sense of Definition 1.8. However, $\hat{u}(t|t_0)$ is only the best global filter when $q(t)$ is a Wiener process and u_0 is Gaussian.

Furthermore, as in [8] we can express $\hat{u}(t|t_0)$ as the strong solution of a stochastic evolution equation under additional assumptions on $\mathcal{A}(t)$, $B(t)$ and the noise processes.

THEOREM 2.4. *Consider the estimation problem* (2.1), (2.2) *under the following additional assumptions*:

(i) $\mathcal{U}(t, s)$ *is an almost strong evolution operator with generator* $\mathcal{A}(t)$,

(ii) $\mathcal{U}(t, r)B(t)e_i \in \mathcal{D}(\mathcal{A}(t))$ *for* $t > r$ *and*

$$\sum_{i=0}^{\infty} \lambda_i^2 \int_0^t \|\mathcal{A}(t)\mathcal{U}(t, r)B(r)\, e_i\|^2\, dr < \infty;$$

(iii) $\mathcal{U}(t, s)B(s)\, e_i \in \mathcal{D}(\mathcal{A}(s))$ *for almost all* $t > s \in T$ *and all* i:

$$\sum_{i=0}^{\infty} \mu_i \int_0^t \|\mathcal{A}(t)\mathcal{U}(t, s)B(s)\, e_i\|\, d\rho(r) < \infty;$$

(iv) $\mathcal{U}(t, 0)P_0\, e_i \in \mathcal{D}(\mathcal{A}(t))$ *for* $t > 0$ *and* $\sum_{i=0}^{\infty} \|\mathcal{A}(t)\mathcal{U}(t, 0)P_0\, e_i\|^2 < \infty$.

Writing

$$\hat{u}(t) = v(t) + x(t),$$

where $v(t) = \int_0^t \mathcal{Y}(t, \sigma)B(\sigma)\, dr(\sigma)$ *is deterministic by Lemma* 2.6, *we have that* $x(t)$ *is the unique solution of the stochastic evolution equation*

$$dx(t) = (\mathcal{A}(t) - P(t)K(t)C(t))x(t)\, dt + P(t)K(t)\, dz(t),$$
$$x(0) = 0, \tag{2.18}$$

where

$$K(t) = C^*(t)(F(t)WF^*(t))^{-1}$$

and $v(t)$ *is the unique solution of the deterministic equation*

$$dv(t) = (\mathcal{A}(t) - P(t)K(t)C(t))v(t)\, dt + B(t)\, dr(t),$$
$$v(0) = 0. \tag{2.19}$$

For the prediction problem $t > t_0$, writing

$$\hat{u}(t|t_0) = \bar{u}(t) + x(t|t_0)$$

where $\bar{u}(t) = E\{u(t)\}$ from Theorem 2.3, *we have that $x(t|t_0)$ is the unique strong solution of the stochastic evolution equation*

$$\begin{aligned} dx(t|t_0) &= \mathcal{A}(t)x(t|t_0)\,dt, \\ x(t|t_0) &= x(t_0), \end{aligned} \tag{2.20}$$

and $\bar{u}(t)$ is the unique solution of the deterministic equation

$$\begin{aligned} d\bar{u}(t) &= \mathcal{A}(t)\bar{u}(t)\,dt + B(t)\,dr(t), \\ \bar{u}(0) &= 0. \end{aligned} \tag{2.21}$$

For the smoothing problem $t_0 > t$, writing $\hat{u}(t|t_0) = v(t) + y(t|t_0)$, we have that $y(t|t_0)$ is the unique strong solution of the stochastic evolution equation

$$\begin{aligned} dy(t|t_0) &= (\mathcal{A}(t) - P(t)K(t)C(t))y(t|t_0)\,dt \\ &\quad + P(t)K(t)C(t)\hat{u}(t)\,dt + \mathrm{B(t)}MB^*(t)\lambda(t)\,dt, \\ y(t_0|t_0) &= x(t_0). \end{aligned} \tag{2.22}$$

Proof.

(i) The stochastic evolution equations (2.18), (2.20) and (2.22) are the equations one obtains for the optimal estimators in the Gaussian case, where $v(t) = 0 = \bar{u}(t)$. So we refer the reader to [8] for the proof that they have a unique strong solution. (The proof uses Theorem 1.3.)

(ii) In [8] it is shown that $\mathcal{A}(t) - P(t)K(t)C(t)$ generates the almost strong evolution operator $\mathcal{Y}(t, s)$, that is

$$\int_s^t (\mathcal{A}(r) - P(r)K(r)C(r))\mathcal{Y}(r, s)x\,dr = (\mathcal{Y}(t, s) - I)x \tag{2.23}$$

$$\text{for } x \in \bigcap_{s \leqq r \leqq t} \mathcal{D}(\mathcal{A}(s)).$$

By Theorem 1.1, $\mathcal{Y}(t, s)$ is the unique solution of

$$\mathcal{Y}(t, s)x = \mathcal{U}(t, s)x - \int_s^t \mathcal{U}(t, \alpha)P(\alpha)K(\alpha)C(\alpha)\mathcal{Y}(\alpha, s)x\,d\alpha, \tag{2.24}$$

and in [8] it is shown that

$$\mathcal{U}(t, s)P(s) : H \to \mathcal{D}(\mathcal{A}(t)) \quad \text{for } t > s \tag{2.25}$$

and

$$\mathcal{A}(t)\mathcal{U}(t, s)P(s)x$$

is Bochner integrable on $(0, t)$ for $x \in H$. From (2.24), (2.25) and assumption (iii), we deduce that

$$\begin{aligned} &\mathcal{Y}(t, s)B(s)e_i \in \mathcal{D}(\mathcal{A}(s)) \quad \text{for almost all } t > s \in T, \quad \text{all } i, \\ &\sum_{i=0}^{\infty} \mu_i \int_0^t \|\mathcal{A}(t)\mathcal{Y}(t, s)B(s)\,e_i\|\,d\rho(r) < \infty. \end{aligned} \tag{2.26}$$

Now the mild solution of (2.19) is

$$v(t)=\sum_{i=0}^{\infty}\mu_i\int_0^t \mathcal{Y}(t,s)B(s)\,e_i\,d\rho(s),$$

and since $\mathcal{A}(t)-P(t)K(t)C(t)$ is closed, from (2.26) we have that $v(t)\in\mathcal{D}(\mathcal{A}(t))$ and

$$(\mathcal{A}(t)-P(t)K(t)C(t))v(t)=\sum_{i=0}^{\infty}\mu_i\int_0^t (\mathcal{A}(t)-P(t)K(t)C(t))\mathcal{Y}(t,s)B(s)\,e_i\,d\rho(s).$$

Since $\mathcal{A}(t)-P(t)K(t)C(t)$ generates an almost strong evolution operator, we have

$$\begin{aligned}
&\int_0^t (\mathcal{A}(r)-P(r)K(r)C(r))v(r)\,dr\\
&\quad=\sum_{i=0}^{\infty}\mu_i\int_0^t\int_0^r (\mathcal{A}(r)-P(r)K(r)C(r))\mathcal{Y}(r,s)B(s)\,e_i\,d\rho(s)\,dr\\
&\quad=\sum_{i=0}^{\infty}\mu_i\int_0^t\int_0^t (\mathcal{A}(r)-P(r)K(r)C(r))\mathcal{Y}(r,s)B(s)\,e_i\,dr\,d\rho(s)\\
&\qquad\qquad\text{(interchanging the order of integration)}\\
&\quad=\sum_{i=0}^{\infty}\mu_i\int_0^t (\mathcal{Y}(t,s)-I)B(s)\,e_i\,d\rho(s)\\
&\qquad\qquad\text{(from (2.23)) since } \mathcal{Y}(t,s)B(s)\,e_i\in\mathcal{D}(\mathcal{A}(s))\\
&\quad=v(t)-\int_0^t B(s)\,dr(s).
\end{aligned}$$

So $v(t)$ satisfies (2.19). Similarly it is shown that $\bar{u}(t)=\int_0^t \mathcal{U}(t,s)B(s)\,d\rho(s)$ is the unique solution of (2.21).

Remarks.

1. From Theorem 1.3, conditions (1.17) for the state $u(t)$ to be the strong solution of the stochastic evolution equation (2.27), $du(t)=\mathcal{A}(t)u(t)\,dt+B(t)\,dq(t)$, are stronger than (i), (ii) and (iii) of Theorem 2.5. This is because the white noise process in (2.27), $B(t)\,dq(t)$, is infinite-dimensional, whereas the white noise process in (2.18), $P(t)K(t)F(t)\,dw(t)$, is finite-dimensional.

2. An alternative assumption to (iii) is the following:

(iii) $\quad \mathcal{U}(t,s)B(s)x\in\mathcal{D}(\mathcal{A}(t))$ for all $t>s$, $x\in H$,

and $B(s)x$ is Hölder continuous on T.

3. From [10], if $\mathcal{U}^*(T-s,T-t)$, or any perturbation, is also a strong evolution operator, then $P(t)$ is the unique solution of the following differential

equation in the class of weakly continuous operators such that $\langle P(t)x, y\rangle$ is absolutely continuous for all $x, y \in \bigcap_{t\in T}\mathscr{D}(\mathscr{A}^*(t))$:

$$\frac{d}{dt}\langle P(t)x, y\rangle - \langle P(t)x, \mathscr{A}^*(t)y\rangle - \langle \mathscr{A}^*(t)x, P(t)y\rangle$$

$$+\langle P(t)C^*(t)(F(t)WF^*(t))^{-1}C(t)P(t)x, y\rangle = \langle B(t)MB(t)^*x, y\rangle \tag{2.27}$$

for $x, y \in \bigcap_{t\in T}\mathscr{D}(\mathscr{A}^*(t))$,

$$P(0) = P_0.$$

If $\mathscr{A}$ is the infinitesimal generator of a strongly continuous semigroup, then $P(t)$ is the unique solution of (2.23)

3. Applications. Of course the first possible application is to finite-dimensional linear systems, and so this includes the work of Kailath [14]. In [8] the same problem was considered for the special case where the noise process in the system model was assumed to be of the form $B(t)\,dw(t)$ where $w(t)$ is a K-valued Wiener process. Several examples were considered where the system was a partial differential equation and also where the system consisted of delay equations. Because of the similar nature of the orthogonal increments process $\{q(t)\}$, the same system models can be considered corrupted by this general noise process; we refer the reader to [8]. To motivate the use of general noise processes we give an application of the theory to the environmental problem of river pollution. This application is examined in detail in [9], so only the outline is given here.

The problem was originally considered by Kwakernaak [15] and concerns the estimation of the concentration of chemical pollution in a river based on measurements at a finite number of points along the river. The time evolution of the concentration of the chemical at location x at time t is $y(t, x)$ and is assumed to be given by

$$\frac{\partial y(t, x)}{\partial x} = D\frac{\partial^2 y(t, x)}{\partial x^2} - V\frac{\partial y(t, x)}{\partial x} + \xi(t, x), \tag{3.1}$$

where D is the dispersion coefficient, V is the water velocity and $\zeta(t, x)$ is the rate of increase of concentration at (t, x) due to the deposits of the chemical wastes. It is assumed that the number of deposits in a section of the river of infinitesimal length dx (x being the distance along the river) behaves according to a Poisson process with rate parameter $\lambda(x)\,dx$, where λ is a given function; the number of deposits in nonoverlapping sections are independent processes and the amounts of chemical deposited at x are independent stochastic variables H_x with $E\{H_x^2\} < \infty$. In [9] it is shown that after a change of variable $y(t, x) = e^{ax}u(t, x)$ $(a = v/2D)$ a convenient model for the polluted river is

$$\begin{aligned} du(t) &= \mathscr{A}u(t)\,dt + B\,dq(t) \qquad \text{on } H = L_2(0, l), \\ u(0) &= u_0 \end{aligned} \tag{3.2}$$

where $\mathcal{A}$ is a self-adjoint operator given by

$$\mathcal{A}\mathcal{L} = \frac{\partial^2 h}{\partial x^2} - a^2 h \quad \text{for } h \in \mathcal{D}(\mathcal{A}),$$

$$\mathcal{D}(\mathcal{A}) = \left\{ h \in H : \frac{\partial h}{\partial x}, \frac{\partial^2 h}{\partial x^2} \in H ; ah + \frac{\partial h}{\partial x} = 0 \text{ at } x = 0, l \right\} \tag{3.3}$$

and $B \in \mathcal{L}(H)$ is given by

$$(\mathcal{B}h)(x) = e^{-ax} h(x) \quad \text{for } h \in H.$$

$u_0 \in L_2(\Omega; H)$ has zero expectation and covariance operator P_0 given by $P_0 e_k = \alpha_k e_k$; $\sum_{k=1}^{\infty} \alpha_k = 0$, where $\{e_i\}$ is the following orthonormal basis for H:

$$e_k(x) = \sqrt{\frac{2}{e}} \sin\left(\frac{k\pi x}{l} + \varepsilon_k\right), \quad \tan \varepsilon_k = -\frac{\pi k}{al}. \tag{3.4}$$

$\{e_k ; k \geqq 1\}$ are the eigenfunctions of $\mathcal{A}$ and $\mathcal{A}$ generates an analytic semigroup $\{\mathcal{T}_t\}$ given by

$$(\mathcal{T}_t h)(x) = \sum_{k=1}^{\infty} \langle h, e_k \rangle e^{(a-(k^2/l^2))t} e_k(x), \qquad h \in H. \tag{3.5}$$

$q(t)$ is an H-valued orthogonal increments process, $q(t) = \sum_{k=1}^{\infty} q_k(t) e_k$, where $q_k(t)$ is a real compound Poisson process for each k and the parameters μ_k, λ_k for $q_k(t)$ are given by

$$\mu_k = \int_0^l \lambda(x) E\{Hx\} e_k(x)\, dx,$$
$$\lambda_k = \int_0^l \lambda(x) E\{H_x^2\} e_k^2(x)\, dx. \tag{3.6}$$

In order that $\sum_{k=1}^{\infty} \mu_k < \infty$ and $\sum_{k=1}^{\infty} \lambda_k < \infty$, we need to have $(\lambda(x) E\{Hx\})^2$ and $\lambda(x) E\{H^2 x\} \in L_2(0, l)$ with

$$(\lambda(x) E\{Hx\})^2 = \sum_{k=1}^{\infty} \beta_k e_{2k}(x)$$

and

$$\lambda(x) E\{H^2 x\} = \sum_{k=1}^{\infty} \gamma_k e_{2k}(x).$$

It is also shown that (3.2) has the unique strong solution

$$u(t) = \mathcal{T}_t u_0 + \int_0^t \mathcal{T}_{t-s} B\, dq(s). \tag{3.7}$$

The observation model is taken to be

$$z(t) = \int_0^t Cu(s)\, ds + w(t), \tag{3.8}$$

where $w(t)$ is a k-dimensional Wiener process with covariance matrix the identity and $C \in \mathscr{L}(H; R^K)$ is given by

$$(Cu)_j = \frac{1}{2\varepsilon} \int_{x_j-\varepsilon}^{x_j+\varepsilon} u(x)\, dx \quad \text{for small } \varepsilon > 0. \tag{3.9}$$

This approximates the situation where you continuously measure the concentration of a chemical at fixed stations $x_1, x_2, \cdots, x_k$, along the river.

Hence (3.7), (3.8) satisfies all the assumptions of our theory in § 2 and so there exists a unique optimal estimator $\hat{u}(t|t_0)$, at least in integral form. Since $\mathscr{A}$ generates a strongly continuous semigroup, $P(t)$ is the unique solution of the differential Riccati equation

$$\begin{aligned} &\frac{d}{dt}\langle P(t)f, h\rangle - \langle P(t)f, \mathscr{A}h\rangle - \langle P(t)h, \mathscr{A}f\rangle \\ &\quad + \langle P(t)C^*CP(t)f, h\rangle = \langle B\Lambda B^* f_1 h\rangle, \\ &\quad P(0) = P_0 \quad \text{and} \quad f, g \in \mathscr{D}(\mathscr{A}). \end{aligned} \tag{3.10}$$

(3.10) may be reduced to an infinite system of ordinary differential equations by expanding

$$\begin{aligned} &P(t) = \sum_{i,j=1}^{\infty}\sum p_{ij}(t)\, e_i(t)\langle e_j(t), \cdot\rangle, \qquad p_{ij}(t) = p_{ji}(t), \\ &\frac{d}{dt}p_{ij}(t) + \frac{\pi^2}{l^2}(i^2+j^2)p_{ij}(t) + \sum_{m,n=1}^{\infty}\sum p_{ni}(t)p_{mj}(t)A_{mn} = p_{ij}, \\ &p_{ij}(0) = \delta_{ij}\lambda_i, \end{aligned} \tag{3.11}$$

where

$$A_{mn} = \sum_{r=1}^{k} a_{mr}a_{nr}$$

and

$$a_{sr} = \frac{\sqrt{2l}}{\delta\pi\varepsilon} \sin\frac{\delta\pi\varepsilon}{l} \sin\left(\frac{\delta\pi x r}{l} + \varepsilon_s\right) = (Ce_s)_r$$

and

$$\rho_{ij} = \sum_{k=1}^{\infty}\sum_{n=1}^{\infty} \lambda_{kn}\langle Be_i, e_n\rangle\langle Be_j, e_k\rangle.$$

In [8] it is shown that assumptions (i), (ii) and (iv) of Theorem 2.4 are satisfied, since $\mathscr{T}_t$ is an almost strong evolution operator and (iii) is proved similarly (or use (iii)), and so $\hat{u}(t|t_0)$ are the unique solutions of the appropriate differential equations (2.18)–(2.22). By expanding $\hat{u}(t|t_0)$ in terms of e_i, it is possible to obtain

the coefficients as solutions of a system of ordinary stochastic differential equations. For example, writing

$$\hat{u}(t) = v(t) + x(t)$$
$$= \sum_{k=1}^{\infty} v_k(t)\, e_k + \sum_{k=1}^{\infty} \beta_k(t)\, e_k,$$

we obtain the systems

$$d\beta_k(t) = -\frac{k^2\pi^2}{l^2}\beta_k(t)\, dt + \sum_{n,r=1}^{\infty}\sum A_{nr}\beta r(t) p_{kn}(t)\, dt$$
$$+ \sum_{n=1}^{\infty} p_{nk}(t) \sum_{r=1}^{k} a_{nr}\, dz_r(t), \tag{3.12}$$
$$\beta_k(0) = 0,$$

$$\dot{v}_k(t) = -\frac{k^2\pi^2}{l^2}v_k(t) + \sum_{n,r=1}^{\infty}\sum A_{nr} v_r(t) p_{kn}(t) + \sum_{r=1}^{\infty} \mu_r \langle Be_r, e_k\rangle, \tag{3.13}$$
$$v_k(0) = 0.$$

So we have obtained an infinite system of Kalman–Bucy-type recursive equations for the filtering problem. These may be solved by truncation, and similarly one can obtain solutions for the smoothing and prediction problem.

Conclusions. The filtering smoothing and prediction problems for a very general class of linear infinite-dimensional systems has been solved. The types of infinite-dimensional systems which generate mild evolution operators is wide, including delay equations, parabolic partial differential equations and hyperbolic partial differential equations, for example.

The noise process is allowed to be of a fairly general type including Gaussian-type white noise and Poisson-type noise, for example. However it should be possible to consider ever more general noise processes using the results on stochastic integration with respect to H-valued martingales developed by Métivier in [16], [17].

It is also important to allow for point observations where the operator C is unbounded. In [11] Curtain and Pritchard have solved the filtering problem with Gaussian white noise with point observations. It should be possible to extend these results to the case of general orthogonal processes introduced in this paper. This is currently under investigation.

REFERENCES

[1] A. V. Balakrishnan, *Stochastic Differential Systems*, Springer-Verlag, Berlin, 1973.
[2] A. Bensoussan, *Filtrage Optimal des Systèmes Linéaires*, Dunod, Paris, 1971.
[3] R. F. Curtain and P. L. Falb, *Stochastic differential equations in Hilbert space*, J. Differential Equations, 10 (1971), pp. 412–430.
[4] R. F. Curtain, *Infinite-dimensional filtering*, this Journal, 13 (1975), pp. 89–104.
[5] ———, *A survey of infinite-dimensional filtering*, SIAM Rev., 17 (1975), pp. 395–411.

[6] R. F. CURTAIN, *A Kalman–Bucy filtering theory for affine differential equations*, Internat. Symp. on Control Theory, Numerical Methods and Computer Systems Modelling, I.R.I.A., Rocquencaut, France, 1974, Lecture Notes in Economics and Math. Systems, Springer-Verlag, Berlin, 1974.

[7] ———, *Stochastic evolution equations with general white noise disturbance*, Control Theory Centre Rep. 41, University of Warwick, Coventry, England, 1975.

[8] ———, *Infinite dimensional estimation theory for linear systems*, Control Theory Centre Rep. 38,, University of Warwick, Coventry, England, 1975.

[9] ———, *Infinite dimensional estimation theory applied to a water pollution problem*, Presented at the 7th IFIP Conference, Nice, Italy, 1975.

[10] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite-dimensional Riccati equations for systems defined by evolution operators*, this Journal, to appear.

[11] ———, *Boundary control and filtering with point observations for finite dimensional systems*, Control Theory Centre Rep. 42, University of Warwick, Coventry, England.

[12] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1952.

[13] P. L. FALB, *Infinite dimensional filtering: the Kalman–Bucy filter in Hilbert space*, Information and Control, 11 (1967), pp. 102–137.

[14] T. KAILATH, *An innovations approach to least-square estimation. Part I: Linear filtering in additive white noise. Part II: Linear smoothing in additive white noise* (with P. Frost), IEEE Trans. Automatic Control, AC-13 (1968), pp. 646–660.

[15] H. KWAKERNAAK, *Filtering for systems excited by Poisson white noise*, Internat. Symp. on Control Theory, Numerical Methods and Computer Systems Modelling, I.R.I.A., Rocquencaut, France, 1974, Lecture Notes in Economics and Math. Systems, 107, Springer-Verlag, Berlin, 1974.

[16] M. MÉTIVIER, *Integrale stochastique par rapport à des martingales hilbertiennes*, C. R. Acad. Sci. Paris, 276, pp. 1009–1012.

[17] M. MÉTIVIER AND G. PISTONE, *Une formule d'isométrie pour l'integrale stochastique hilbertienne et équations d'évolution linéaires stochastiques*, to appear.

[18] S. K. MITTER AND R. B. VINTER, *Filtering for linear stochastic hereditary differential systems*, Internat. Symp. on Control Theory, Numerical Methods and Computer Systems Modelling, I.R.I.A., Rocquencaut, France, 1974, Lecture Notes in Economics and Math. Systems, 107, Springer-Verlag, Berlin, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 16, November 1976

RELATIONS AMONG THE MULTIPLIERS FOR PROBLEMS WITH BOUNDED STATE CONSTRAINTS*

I. BERT RUSSAK†

Abstract. In previous articles, the author established certain necessary conditions for control problems with constraints of the form $\psi^\alpha(t, x) \leqq 0$ $\alpha = 1, \cdots, m$. These conditions involve certain multiplier functions $\mu_\alpha(t)$ of the derivatives of the above constraints together with multiplier constants K^α used in the transversality relation. In this paper, it is shown that these terms satisfy $\mu_\alpha(t^0) \leqq K^\alpha$ with $\mu_\alpha(t^0) = K^\alpha$ if $\psi^\alpha(t^0) < 0$.

1. Introduction. We consider the following problem. Let A be the class of arcs a:

$$a: \quad \underset{i=1,\cdots,N}{x^i(t)} \quad \underset{k=1,\cdots,K}{u^k(t)} \quad \underset{\sigma=1,\cdots,r}{b^\sigma}, \qquad t^0 \leqq t \leqq t^1,$$

which have points $t, x(t), u(t)$ in a region R in t-x-u space, b in a region B in b space and $u(t)$ piecewise continuous, and which satisfy the conditions

$$\dot{x}^i(t) = f^i(t, x(t), u(t)), \qquad i = 1, \cdots, N \tag{1.1}$$

$$\psi^\alpha(t, x(t)) \leqq 0, \qquad \alpha = 1, \cdots, m, \tag{1.2}$$

$$I_\gamma(a) \leqq 0, \quad 1 \leqq \gamma \leqq p', \qquad I_\gamma(a) = 0, \quad p' < \gamma \leqq p, \tag{1.3}$$

$$x^i(t^s) = X^{is}(b), \qquad s = 0, 1, \quad 1 \leqq i \leqq N, \tag{1.4}$$

where

$$I_\gamma(a) = g_\gamma(b) + \int_{t^0}^{t^1} L_\gamma(t, x(t), u(t))\, dt, \qquad \gamma = 1, \cdots, p.$$

It is desired to minimize the functional

$$I_0(a) = g_0(b) + \int_{t^0}^{t^1} L_0(t, x(t), u(t))\, dt \tag{1.5}$$

on the class A.

The functions ψ^α are assumed to be of class C^2 on R while the functions f^i, L_γ, g_γ, X^{is} are of class C^1 on R or B as the case may be. We shall assume familiarity with the notation and conventions of [1] through [3]. Furthermore unless otherwise specified the values i, k, σ, α will have their above indicated ranges.

Assume, next, that the arc

$$a_0: \quad x_0(t) \quad u_0(t) \quad b_0, \qquad t^0 \leqq t \leqq t^1,$$

is a solution to our problem and define the functions

$$\phi^\alpha(t, x, u) = \psi_t^\alpha + \psi_{x^i}^\alpha f^i, \qquad \alpha = 1, \cdots, m. \tag{2}$$

* Received by the editors December 3, 1975.

† Department of Mathematics, Naval Postgraduate School, Monterey, California 93940. This work was supported by an NPS Foundation grant.

For arcs in the class A, these functions act as $d\psi^\alpha/dt$ along these arcs. We assume that the matrix

$$[\phi^\alpha_{u^k} \quad \delta_{\alpha\beta}\psi^\beta], \qquad \alpha, \beta = 1, \cdots, m, \tag{3}$$

(where $\delta_{\alpha\beta}$ is the Kronecker delta) has rank m on the set R_0 of points $(t, x_0(t), u)$ satisfying

$$\begin{gathered}\psi^\alpha \leqq 0, \\ \phi^\alpha \geqq 0 \text{ for all } \alpha \text{ with } \psi^\alpha = 0 \quad \text{or} \quad \phi^\alpha \leqq 0 \text{ for all } \alpha \text{ with } \psi^\alpha = 0, \\ 1 \leqq \alpha \leqq m.\end{gathered} \tag{4}$$

Referring to Theorem 3.1 of [1] and to the quantities $\mu_\alpha(t)$, K^α of that theorem, we prove the following result:

LEMMA. *For each α we have*

$$\mu_\alpha(t^0) \leqq K^\alpha \quad \textit{with } \mu_\alpha(t^0) = K^\alpha \quad \textit{if } \psi^\alpha(t^0) < 0. \tag{5}$$

In Theorem 3.2 of [1], the multipliers $\mu_\alpha(t)$ are modified (by the addition of additive constants) from those of Theorem 3.1 of [1]. The results of this paper then imply associated results to the multipliers of that theorem. Similar remarks hold in the Theorems of [2].

2. Proof of the lemma. It is convenient to prove this result by first transforming the problem. In § 4 of [1] the problem stated above is shown to be equivalent to a reformulated problem (with superscript bars used on quantities in the reformulated problem to distinguish them from the original problem so that for example, $\bar\psi^\alpha$ replaces ψ^α) with functions $\bar\psi^\alpha$, $\bar\phi^\alpha$ formed from the functions ψ^α, ϕ^α and with the major distinction from the above problem being that the assumption involving (3) is replaced by the statement that the matrix

$$[\bar\phi^\alpha_{u^k}] \tag{6}$$

has rank m at points in $\bar D$. Here $\bar D$ is the set of points $(t, \bar x_0(t), u)$ in $\bar R_0$ with $u = \bar u_0(t)$ or for arbitrary u with t interior to an interval of continuity of $\bar u_0(t)$. Now $\bar\phi^\alpha = d\bar\psi^\alpha/dt$ and so (6) implies in particular that

$$[\bar\psi^\alpha_{x^i}(t^0)] \quad \text{has rank } m. \tag{7}$$

The argument t^0 in (7) means evaluation at the point t^0, $x_0(t^0)$ on the arc $\mathscr{a}_0$. We shall use an analogous convention at other points along $\mathscr{a}_0$ and for other functions.

The theorem for this latter problem is Theorem 6.1 of [1] and as shown in § 7 of [1], the terms $\mu_\alpha(t)$, K^α of that theorem and of Theorem 3.1 of [1] for the original problems are the same. In addition, $\bar\psi^\alpha(t^0) = 0$ iff $\psi^\alpha(t^0) = 0$ $\alpha = 1, \cdots, m$, as shown in (36) of [1]. Thus proving our lemma for the reformulated problem will prove it also for the original problem.

We concentrate on the reformulated problem of § 4 of [1].

In order now to prove the first inequality of (5), assume that η is an index such that

$$\bar\psi^\eta(t^0) < 0, \tag{8}$$

and let h be any N-dimensional vector such that $\bar{\psi}^{\eta}_{x^i}(t^0)h^i \neq 0$. Now, according to (7), we can select a vector d such that

$$\begin{aligned} &\bar{\psi}^{\eta}_{x^i}(t^0)\, d^i = \bar{\psi}^{\eta}_{x^i}(t^0)h^i, \\ &\bar{\psi}^{\alpha}_{x^i}(t^0)\, d^i = 0, \qquad \alpha \neq \eta. \end{aligned} \tag{9}$$

Next, select a constant $\delta > 0$ so small that

$$\bar{\psi}^{\eta}(t) < 0, \qquad t^0 \leqq t \leqq t^0 + \delta, \tag{10}$$

and define the k-dimensional arc w such that

$$w^{\alpha}(t^0) = \bar{\psi}^{\alpha}_{x^i}(t^0)\, d^i,$$

$$\dot{w}^{\alpha}(t) = \begin{cases} (-\bar{\psi}^{\alpha}_{x^i}(t^0)\, d^i)2/\delta, & t^0 \leqq t \leqq t^0 + \delta/2, \quad \alpha = 1, \cdots, m, \\ 0 & t^0 + \delta/2 \leqq t \leqq t^1, \quad \alpha = 1, \cdots, m, \end{cases} \tag{11.1}$$

$$w^{\Gamma}(t) \equiv 0, \quad \Gamma = m+1, \cdots, K, \quad t^0 \leqq t \leqq t^1. \tag{11.2}$$

Then w is in the class W of § 13 of [1] and by Lemma 13.1 of [1], we can find an admissible variation

$$\delta a: \quad \delta x(t) \quad \delta u(t) \quad \delta b \qquad\qquad t^0 \leqq t \leqq t^1, \tag{12}$$

satisfying

$$\delta x^{j_s}(t^0) = d^{j_s}, \qquad j_s \neq i_\rho, \quad s = 1, \cdots, N-m, \tag{13.1}$$

$$\delta b = 0, \tag{13.2}$$

where i_ρ are the indices of (108) of [1] and also satisfying

$$\begin{aligned} &\bar{\psi}^{\alpha}_{x^i}(t)\delta x^i(t) = \delta\bar{\psi}^{\alpha}(t) = w^{\alpha}(t), \qquad \alpha = 1, \cdots, m, \\ &\delta\bar{\phi}^{\Gamma}(t) = w^{\Gamma}(t), \quad \Gamma = m+1, \cdots, K \quad t^0 \leqq t \leqq t^1, \end{aligned} \tag{14}$$

where $\delta\bar{\psi}^{\alpha}(t)$, $\delta\bar{\phi}^{\Gamma}(t)$ indicate the variations in these quantities due to the variation δa and where $\bar{\phi}^{\Gamma}$ are the functions of § 8 of [1]. According to the above and by the admissibility of δa, we have that

$$\delta\bar{\phi}^{\alpha}(t) = \frac{d}{dt}\,\delta\bar{\psi}^{\alpha}(t) = \dot{w}^{\alpha}(t) = \begin{cases} (-\bar{\psi}^{\alpha}_{x^i}(t^0)\, d^i)\dfrac{2}{\delta}, & \left[t^0, t^0 + \dfrac{\delta}{2}\right], \\ 0 & \left[t^0 + \dfrac{\delta}{2}, t^1\right] \end{cases} \tag{15}$$

and by (14) and (11.2) also

$$\delta\bar{\phi}^{\Gamma}(t) \equiv 0, \qquad \Gamma = m+1, \cdots, K, \quad t^0 \leqq t \leqq t^1. \tag{16}$$

In addition, by (11.1), (13.1), and (14) evaluated at $t = t^0$, we have

$$\bar{\psi}^{\alpha}_{x^{i_\rho}}(t^0)[d^{i_\rho} - \delta x^{i_\rho}(t^0)] = 0, \qquad \rho, \alpha = 1, \cdots, m, \tag{17}$$

where i_ρ are the indices of (108) of [1]. Then by the nonsingularity of the matrix $[\bar{\psi}^{\alpha}_{x^{i}_{\rho}}(t^0)]$ (see (108) of [1]), we see that $\delta x^{i_\rho}(t^0) = d^{i_\rho} \rho = 1, \cdots, m$, so that together

with (13.1) we obtain

$$\delta x^j(t^0) = d^j, \qquad j = 1, \cdots, N. \tag{18}$$

Next, by (155.2) and Lemmas 11.1 and 15.1 all of [1], together with (15), (16) and (18), we get by computing the variation of the functionals introduced in (69) and (70) of [1] that

$$\tilde{\lambda}_\rho \int_{t^0}^{t^0+\delta/2} F_{\rho u^k} \zeta^k_\alpha (-\bar{\psi}^\alpha_{x^i}(t^0)\, d^i) \frac{2}{\delta}\, dt - \tilde{\lambda}_{p+N+i}\, d^i \geqq 0, \tag{19}$$

$$\rho = 0, 1, \cdots, p+N,$$

where F_ρ, ζ^k_α, $\tilde{\lambda}_{p+N+i}$ are quantities introduced in § 8 of [1].

Using the relations (76.1) of [1] (between $\tilde{\lambda}_{p+N+i}$ and K^α) and (9), we see that (19) becomes

$$\tilde{\lambda}_\rho \int_{t^0}^{t^0+\delta/2} F_{\rho u^k} \zeta^k_\eta (-\bar{\psi}^\eta_{x^i}(t^0) h^i) \frac{2}{\delta}\, dt - K^\eta \bar{\psi}^\eta_{x^i}(t^0) h^i \geqq 0 \quad (\eta \text{ not summed}), \tag{20}$$

where K^η is that term referred to in our present lemma which is associated with $\bar{\psi}^\eta$. Furthermore, by the definition of $\mu_\alpha(t)$ in (74) and (76) of [1] then (20) is

$$(\bar{\psi}^\eta_{x^i}(t^0) h^i) \left[\frac{2}{\delta} \int_{t^0}^{t^0+\delta/2} \mu_\eta(t)\, dt - K^\eta \right] \geqq 0 \quad (\eta \text{ not summed}). \tag{21}$$

According to the properties of the multipliers $\mu_\alpha(t)$, we can by reducing δ if necessary, guarantee that $\mu_\eta(t)$ is continuous on $[t^0, t^0+\delta/2]$. Then by taking the limit of the expression in (21), we get that

$$\bar{\psi}^\eta_{x^i}(t^0) h^i [\mu_\eta(t^0) - K^\eta] \geqq 0 \quad (\eta \text{ not summed}). \tag{22.1}$$

Now we can repeat this same construction with $-h$ replacing h and so get

$$\bar{\psi}^\eta_{x^i}(t^0)(-h^i)[\mu_\eta(t^0) - K^\eta] \geqq 0 \quad (\eta \text{ not summed}). \tag{22.2}$$

Thus, (22) implies that for any vector h with $\bar{\psi}^\eta_{x^i}(t^0) h^i \neq 0$, then

$$\bar{\psi}^\eta_{x^i}(t^0) h^i [\mu_\eta(t^0) - K^\eta] = 0 \quad (\eta \text{ not summed}) \tag{23}$$

which implies that

$$\mu_\eta(t^0) = K^\eta. \tag{24}$$

Since $\bar{\psi}^\eta$ was an arbitrary constraint such that $\bar{\psi}^\eta(t^0) < 0$, then the second statement of our lemma is proved.

In order to prove the first statement of our lemma, let η be an index such that

$$\bar{\psi}^\eta(t^0) = 0 \tag{25}$$

and let h be a vector such that

$$\bar{\psi}^\eta_{x^i}(t^0) h^i \leqq 0. \tag{26}$$

Then as above, pick a vector d such that (9) is true and define the arc w as in (11) where δ is selected so that the multiplier $\mu_\eta(t)$ is continuous on $[t^0, t+\delta/2]$. The construction follows identical steps to the above to yield (22.1) which together

with (26) and the arbitrariness of η, proves the first statement of our lemma and hence also the lemma.

REFERENCES

[1] I. B. RUSSAK, *On problems with bounded state variables*, J. Optimization Theory Appl., 5 (1970), pp. 114–157.
[2] ———, *On general problems with bounded state variables*, Ibid., 6 (1970), pp. 424–452.
[3] ———, *Second order necessary conditions for problems with state inequality constraints*, this Journal, 13 (1975), pp. 372–388.