SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

ON OPTIMAL CONTROL OF LINEAR STOCHASTIC EQUATIONS WITH A LINEAR-QUADRATIC CRITERION*

JEAN-MICHEL BISMUT†

Abstract. The purpose of this paper is to apply the stochastic maximum principle previously obtained by the author to the control of a linear quadratic criterion.

1. Introduction. We consider a stochastic differential equation:

$$\begin{aligned} dx &= (Ax + Cu)\,dt + (Bx + Du)\cdot dw, \\ x(0) &= 0, \end{aligned} \tag{1.1}$$

and a criterion to minimize

$$I(u) = E\left\{ \int_0^T |M_t x_t|^2\,dt + \int_0^T \langle N_t u_t, u_t\rangle\,dt + |M_1 x_T|^2 - 2\langle h, x_T\rangle \right\}, \tag{1.2}$$

where h is a random variable and coefficients are random.

The purpose of this paper is to find the optimal control in feedback form, by using the results obtained by the author in [1] and [2].

In § 2, as in [2], we introduce a dual state, and we discuss some of the problems related to this dual state. In § 3, we find the control in random feedback form.

2. The problem. Assumptions and notations are taken from [2], to which we refer constantly.

Equation (1.1) and criterion (1.2) satisfy the same assumptions as in [2]. We also assume that $h \in L_2^T$.

THEOREM 2.1. *I has a unique optimum.*

Proof. The argument is the same as in [2, Thm. 3.1]. □

We apply the stochastic maximum principle given in [1, Thm. V-1]. The maximum principle equations are

$$\begin{aligned} dp &= (M^*Mx - A^*p - B^*H)\,dt + H\cdot dw + dM, \\ p_T &= -M_1^* M_1 x_T + h, \\ Nu &= C^*p + D^*H, \end{aligned} \tag{2.1}$$

with (p_0, H, M) in $L_2^0 \times L_{22} \times W^\perp$.

As in [2], for $t \geqq 0$ we consider the system

$$\begin{aligned} dx &= (Ax + Cu)\,dt + (Bx + Du)\cdot dw, \\ x(0) &= 0, \\ dp &= (M^*Mx - A^*p - B^*H)\,dt + H\cdot dw + dM, \\ p_t &= h, \\ Nu &= C^*p + D^*H, \end{aligned} \tag{2.2}$$

* Received by the editors June 6, 1975, and in revised form February 23, 1976.

† Paris, France.

where $h \in L_2^t$.

It is easily checked as in [2] that

$$E\left\{\int_0^t |M_s x_s|^2\, ds + \int_0^t \langle N_s u_s, u_s\rangle\, ds\right\} = E\langle p_t, x_t\rangle. \tag{2.3}$$

As in [2], we can then prove that the mapping $Q_t : h \to x_t$ has the following properties:

(a) Q_t is linear and continuous from L_2^t into L_2^t.

(b) Q_t is self-adjoint.

(c) Q_t is a positive operator.

(d) The operators Q_t are uniformly bounded on compact sets of R^+.

However, in total contrast to [2], we *do not have*

$$Q_t(1_A h + 1_{CA} h') = 1_A Q_t h + 1_{CA} Q_t h' \tag{2.4}$$

when $A \in \mathcal{F}_t$. The operators Q_t are of interest because they would allow us to write

$$x_t = Q_t p_t. \tag{2.5}$$

However, this is *not* a feedback relation in the sense that this operator generally acts on the whole random variable and not only on its values at time t. (This last assumption is verified only in the deterministic case where Q_t solves a Riccati equation.)

We then have to use other methods.

3. The feedback form. p_0 is the unique solution in the sense of [2] of

$$\begin{aligned} dp_0 &= -(A^* p_0 + B^* H_0)\, dt + H_0 \cdot dw + dM_0, \\ p_{0_T} &= h. \end{aligned} \tag{3.1}$$

Then $p_1 = p - p_0$ must verify

$$\begin{aligned} dp_1 &= (M^* M x - A^* p_1 - B^* H_1)\, dt + H_1 \cdot dw + dM_1 \\ p_{1_T} &= -M_1^* M_1 x_T. \end{aligned} \tag{3.2}$$

We then have

$$Nu = (C^* p_0 + D^* H_0) + (C^* p_1 + D^* H_1). \tag{3.3}$$

If u_0 and u_1 are defined by

$$u_0 = N^{-1}(C^* p_0 + D^* H_0), \tag{3.4}$$

$$u_1 = N^{-1}(C^* p_1 + D^* H_1), \tag{3.5}$$

then we have the following system:

$$\begin{aligned} dx &= (Ax + Cu_1 + Cu_0)\, dt + (Bx + Du_1 + Du_0) \cdot dw, \\ x(0) &= 0, \\ dp_1 &= (M^* M x - A^* p_1 - B^* H_1)\, dt + H_1 \cdot dw + dM_1, \\ p_{1_T} &= -M_1^* M_1 x_T, \\ Nu_1 &= C^* p_1 + D^* H_1. \end{aligned} \tag{3.6}$$

But this system is a system of the type already studied in [2], with

$$f = Cu_0; \qquad g = Du_0. \tag{3.7}$$

We have then

$$p_{1_T} = -(P_t x_t + r_t), \tag{3.8}$$

where P_t and r_t are defined in [2].

We assume that $(\Omega, \mathcal{F}_t, P) = (\Omega' \times \Omega'', \mathcal{F}'_t \otimes \mathcal{F}''_t, P' \otimes P'')$, that A, B, C, D, M, M_1, N are defined on Ω' and adapted to $\{\mathcal{F}'_t\}_{t \geqq 0}$, that w is defined on Ω'' and adapted to $\{\mathcal{F}''_t\}_{t \geqq 0}$, and that h is defined on Ω, is square integrable and $\mathcal{F}_T$-measurable. We have then in the sense of Theorems 6.1 and 6.2 of [2]:

$$\begin{aligned}
&dP + \{PA + A^*P + B^*PB - (B^*PD + PC)(N + D^*PD)^{-1} \\
&\qquad \cdot (D^*PB + C^*P) + M^*M\}\, dt - d\mathcal{M} = 0, \\
&P_T = M_1^* M_1, \\
&dr = \{(PC + B^*PD)(N + D^*PD)^{-1} C^* - A^*\} r\, dt \\
&\qquad + [\{(PC + B^*PD)(N + D^*PD)^{-1} D^* - B^*\}(PDu_0 + \tilde{h}) - PCu_0]\, dt \\
&\qquad + \tilde{h} \cdot dw + dM' \\
&r_T = 0, \\
&u_1 = -(N + D^*PD)^{-1}\{(C^*P + D^*PB)x + C^*r + D^*(PDu_0 + \tilde{h})\}.
\end{aligned} \tag{3.9}$$

Knowing u_0 we find the optimal control u:

$$u = u_0 + u_1. \tag{3.10}$$

Then

$$p = p_0 - r - Px. \tag{3.11}$$

4. An example: The deterministic coefficients. We assume that all the coefficients $A, B, \cdots, h$ are deterministic. In this case, p_0, r and u_0 are deterministic. Then p and u will be sums of a deterministic process and of a process in feedback form.

Remark. When B and D are null, it is easily proved that the operators Q_t are found by solving a simple Riccati equation:

$$\begin{aligned}
&dQ = AQ + QA^* + CN^{-1}C^* - QM^*MQ, \\
&Q(0) = 0.
\end{aligned} \tag{4.1}$$

Then

$$x_t = Q_t p_t. \tag{4.2}$$

In the general case, we are not able to construct the Q_t directly.

REFERENCES

[1] J. M. BISMUT, *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. Appl., 44 (1973), pp. 384–404.

[2] ———, *Linear quadratic optimal stochastic control with random coefficients*, this Journal, 14 (1976), pp. 419–444.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

ON THE UNIFORM ASYMPTOTIC STABILITY OF CERTAIN LINEAR NONAUTONOMOUS DIFFERENTIAL EQUATIONS*

A. P. MORGAN† AND K. S. NARENDRA‡

Abstract. In this paper we give a simple characterization of the uniform asymptotic stability of equations $\dot{x} = -P(t)x$ where $P(t)$ is a bounded piecewise continuous symmetric positive semi-definite matrix. In the course of developing this characterization, a new and general sufficient condition is given for uniform asymptotic stability in terms of Lyapunov functions. The stability of this type of equation has come up in various control theory contexts (identification, optimization and filtering).

1. Introduction.

The stability of the ordinary differential equation

$$\dot{x} = -P(t)x, \tag{1}$$

where $P(t)$ is symmetric positive semi-definite time-varying matrix arises often in mathematical control theory. (See, for example, Narendra and McBride [8, p. 34, (20)], Lion [7, p. 1837, (10)], and Sondhi and Mitra [11, p. 5, (7)].)

In this paper we consider the stability properties (in the sense of Lyapunov) of the equilibrium state $x = 0$. Since for $V(x) = x^T x$, $\dot{V}(x) \leqq 0$, the origin is uniformly stable. However (uniform) asymptotic stability does not generally hold unless $P(t)$ is positive definite. The semi-definite case arises much more frequently in practice than the definite one, and the main effort in this paper is directed towards finding conditions characterizing uniform asymptotic stability in such a case.

The treatment of uniform asymptotic stability (u.a.s.) rather than mere asymptotic stability is important here. This uniformity assures the "stability under persistent disturbances" of the system. (See Hahn [3, p. 275]; also see Hale [4, pp. 86, 313].) On the other hand, this type of stability is not necessarily possessed by (nonuniform) asymptotically stable systems. (See Hale [4, p. 87] for an example.) Further, u.a.s. proofs yield "rate of convergence" information, and this is frequently not the case if only asymptotic stability is established. Note also that since (1) is linear, all stability properties are global.

The principal results are stated in Theorems 1 and 2, Proposition 1, and the Lemma. The following theorem, which is a part of Theorem 1, gives a simple and complete characterization of uniform asymptotic stability and is illustrative of the type of result derived in this paper.

THEOREM. *Suppose $P(t)$ is a symmetric positive semi-definite matrix of bounded piecewise continuous functions. Then the equation*

$$\dot{x} = -P(t)x$$

is uniformly asymptotically stable if and only if there are real numbers $a > 0$ and b

* Received by the editors June 11, 1975, and in final revised form November 3, 1975. The research reported in this document was sponsored in part by support extended to Yale University by the U.S. Office of Naval Research under Contract N00014-67-A-0097-0020.

† Department of Mathematics, University of Miami, Coral Gables, Florida. Now at Medical College of Georgia, Augusta, Georgia 30902.

‡ Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520.

such that

$$\int_{t_0}^{t} |P(s)w|\, ds \geqq a(t-t_0)+b$$

for all $t \geqq t_0 \geqq 0$ *and all fixed unit vectors* w.

In § 2 we discuss some examples. In §§ 3 and 4 the principal results for uniform and nonuniform asymptotic stability are stated. A key lemma used to establish the results is given in great generality in § 3 and should be useful to show uniform asymptotic stability for other classes of linear and nonlinear systems of equations. Sections 5 and 6 contain the proofs of the theorems in §§ 3 and 4, respectively.

2. Preliminary discussion. Before stating all our main results, we will discuss some implications of the Theorem above. Our discussion divides naturally into five parts ((a), (b), (c), (d) and (e) below). First however, we state the following.

DEFINITION. The equilibrium state $x \equiv 0$ of the uniformly stable differential equation $\dot{x} = f(x, t)$ is uniformly asymptotically stable (u.a.s.) if for some $\varepsilon_1 > 0$ and all $\varepsilon_2 > 0$ there is a $T = T(\varepsilon_1, \varepsilon_2) > 0$ such that if $x(t)$ is a solution and $|x(t_0)| < \varepsilon_1$, then $|x(t)| < \varepsilon_2$ if $t \geqq t_0 + T$. If T depends on t_0, then $\dot{x} = f(x, t)$ is (nonuniformly) asymptotically stable (a.s.). (See Fig. 1.)

We should also make the following comment on notation. We use the n-tuple notation $(x_1, x_2, \cdots, x_n)$ for the column matrix $[x_1, x_2, \cdots, x_n]^T$.

(a) If $P(t) = P$ is a constant matrix, then (1) is u.a.s. if and only if P has rank n.

If $P(t)$ is periodic and continuous, then (1) is u.a.s. if and only if, for each unit w, $|P(t)w| > 0$ for some t.

(b) Let $\lambda(t)$ denote the eigenvalue of minimal length of $P(t)$. Then u.a.s. holds if there are $a > 0$ and b such that

$$\int_{t_0}^{t} |\lambda(s)|\, ds \geqq a(t-t_0)+b$$

for all $t \geqq t_0$.

In particular, if $P(t)$ has (maximal) rank n for all t and $\lambda(t)$ is bounded above

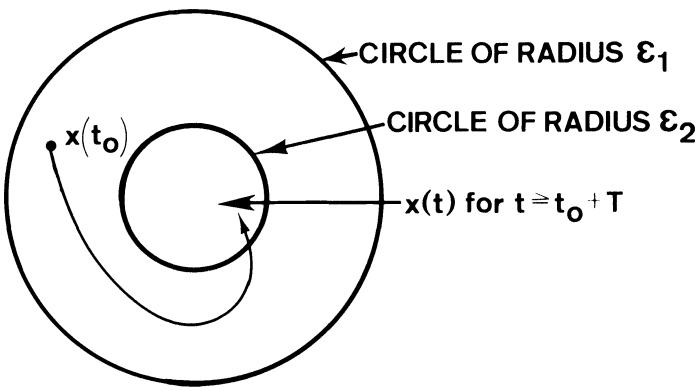


FIG. 1

zero or periodic, then $\dot{x} = -P(t)x$ is u.a.s. Thus if $P(t)$ is rank n and periodic, then u.a.s. holds. However,

$$\int_{t_0}^{t} |\lambda(s)|\, ds \geqq a(t-t_0)+b$$

is not necessary but only sufficient. This will be clear from the discussion of the 2×2 rank 1 case in part (c) below.

(c) Suppose there is $u\colon [0,\infty)\to R^2$ such that

$$P(t) = u(t)\cdot u(t)^T = \begin{bmatrix} u_1^2 & u_1u_2 \\ u_1u_2 & u_2^2 \end{bmatrix}.$$

The eigenvalues of $P(t)$ are then $|u(t)|^2 = u_1(t)^2 + u_2(t)^2$ and 0. Now

$$\dot{x} = -P(t)\cdot x$$

becomes $\dot{x} = -\langle u(t), x\rangle \cdot u(t)$ where $\langle\ \cdot\ \rangle$ denotes the canonical inner product on R^2. Thus the condition

$$\int_{t_0}^{t} |P(s)w|\, ds = \int_{t_0}^{t} |\langle u(s), w\rangle|\, |u(s)|\, ds \geqq a(t-t_0)+b$$

for fixed unit vectors w requires that both $|\langle u(s), w\rangle|$ and $|u(s)|$ "not get too small for too long". Thus $u(s)$ must change direction uniformly so that its inner product with any fixed direction w does not converge too quickly to zero, and also $u(s)$ itself must not converge too quickly to zero. To further illustrate this, consider the following explicit examples.

(d) Let $e_1 = (1, 0)$ and $e_2 = (0, 1)$. Define vectors $u(t)$ and $u'(t)$ to alternate between e_1 and e_2 according to the following formulas.

(i) $u(t) = e_1$ if $t\in[2n, 2n+1)$,
$u(t) = e_2$ if $t\in[2n+1, 2n+2)$,

(ii) $u'(t) = e_1$ if $t\in[0, 1]$,
$= e_2$ if $t\in[1, 2)$,
$= e_1$ if $t\in[2, 4)$,
$= e_2$ if $t\in[4, 5)$,
$= e_1$ if $t\in[5, 8)$,
$\vdots$
$= e_1$ if $t\in[k, k+n)$,
$= e_2$ if $t\in[k+n, k+n+1)$,
$= e_1$ if $t\in[k+n+1, (k+n+1)+(n+1))$,
$\vdots$

Now $\dot{x} = -u(t)u(t)^T x$ is u.a.s., because

$$|\langle u(s), w\rangle|\, |\langle u(s)\rangle| = |\langle u(s), w\rangle| \geqq \frac{\sqrt{2}}{2} \quad \text{either for all } s\in[2n, 2n+1)$$

or for all $s\in[2n+1, 2n+2)$, where $n = 0, 1, 2, 3, \cdots$.

But $\dot{x} = -u'(t)u'(t)^T x$ is not u.a.s., because u' spends longer and longer in the e_1 direction. Solutions with initial conditions on the y-axis must wait longer and longer before they can go to zero. It is clear that

$$\int_{t_0}^{t} |\langle u'(s), e_2\rangle| |u'(s)|\, ds = \int_{t_0}^{t} |\langle u'(s), e_2\rangle|\, ds$$

equals zero for longer and longer intervals and can dominate no linear function with positive slope. However, the above integral does go to infinity as $t \to \infty$, and we shall see in § 4 that this implies $x = -u'(t)u'(t)^T x$ is asymptotically stable.

(e) Consider the following final example. Let $u(t) = (1, 1/\sqrt{t})$. Then

$$u(t)u(t)^T = \begin{bmatrix} 1 & \dfrac{1}{\sqrt{t}} \\ \dfrac{1}{\sqrt{t}} & \dfrac{1}{t} \end{bmatrix},$$

and $|\langle u(s), w\rangle|\, |u(s)| = |w_1 + (1/\sqrt{t})w_2|\, |\sqrt{1+1/t}|$. Thus for $w = (0, 1)$, we require that

$$\int_{t_0}^{t} \frac{1}{\sqrt{s}} \sqrt{\left(1+\frac{1}{s}\right)}\, ds \leqq \int_{t_0}^{t} 2\frac{1}{\sqrt{s}}\, ds = 2\sqrt{s}\Big|_{t_0}^{t}$$

dominate a linear function; but this is false.

It is easy to confirm that if $u(t) = (1, t^\alpha)$ where $\alpha < 0$, then $\dot{x} = -u(t)u(t)^T x$ is not u.a.s. We shall see in § 4 that such equations are not even a.s.

We close this section by noting that the comments made in (c), (d) and (e) clearly hold for the general $n \times n$ case.

3. Uniform asymptotic stability. If $P(t)$ is symmetric positive semi-definite, then there is a symmetric $u(t)$ such that $P(t) = u(t)^2 = u(t)u(t)^T$. (See Reed and Simon [10, p. 196].) We will usually assume $P(t)$ is in this form. As a special case we consider $P(t) = u(t)u(t)^T$ with $u(t)$ an $n \times k$ matrix with $k \leqq n$. In this case, $u(t)u(t)^T$ can have at most rank k. In general, $u(t)$ is $n \times n$ but not necessarily of full rank. In fact, the rank of $u(t)u(t)^T$ may change with t. We do assume $u(t)$ is piecewise continuous and uniformly bounded.

Letting $V(x) = x_1^2 + x_2^2 + \cdots + x_n^2$, we see that $\dot{V}(x) = -x^T P(t)x \leqq 0$ for $\dot{x} = -P(t)x$. Thus the equation is easily seen to be uniformly stable. If $P(t)$ is constant or periodic, we have the well-known result of LaSalle by which if V is not constant on any solution of $\dot{x} = -P(t)x$, then asymptotic stability follows. (See LaSalle [5].) This result breaks down for general nonautonomous $P(t)$. This can be seen as a result of the lack of an invariance property for the ω-limit set. (See LaSalle [6].)

The following theorem gives a characterization of uniform asymptotic stability for $\dot{x} = -P(t)x$. The statement of the theorem is followed by a key Lemma and some remarks. Proofs are deferred until § 5. In reading the following material, the reader may find the case $u : [0, \infty) \to R^2$ an illuminating example.

THEOREM 1. *Let $u : [0, \infty) \to R_k^n$ be a piecewise continuous and bounded function, where R_k^n denotes the space of real $n \times k$ matrices.* (We identify R_1^n and

R^n.) *Then the following are equivalent.*

1. $\dot{x} = -u(t)u(t)^T x$ *is uniformly asymptotically stable.*

2. *There are real numbers* $a > 0$ *and* b *such that if* $y \in R^n$ *is a fixed unit vector, then*

$$\int_{t_0}^{t} y^T u(s)u(s)^T y \, ds \geqq a(t-t_0)+b$$

for all $t \geqq t_0 \geqq 0$.

Equivalently, we may replace the above integral expression by

$$\int_{t_0}^{t} |u(s)u(s)^T y| \, ds \geqq a(t-t_0)+b$$

or by

$$\int_{t_0}^{t} |u(s)^T y| \, ds \geqq a(t-t_0)+b.$$

3. *There are real numbers* $a > 0$ *and* b *such that*

$$\lambda_i\left[\int_{t_0}^{t} u(s)u(s)^T ds\right] \geqq a(t-t_0)+b \quad \text{for } i = 1, 2, \cdots, n,$$

where λ_i *denotes the* i*-th eigenvalue of the* $n \times n$ *matrix*

$$\int_{t_0}^{t} u(s)u(s)^T ds.$$

4. *Given* y *a unit vector in* R^n, *there is a conical neighborhood* C_y *for* y *and there are real numbers* $a_y > 0$ *and* b_y *such that*

$$\int_{[t_0,t]-\Omega_y} |u(s)|^2 \, ds \geqq a_y(t-t_0)+b_y \quad \text{for all } t \geqq t_0 \geqq 0,$$

where $\Omega_y = \{t \in [0, \infty) | u(t)^\perp \cap C_y \neq 0\}$, $u(t)^\perp$ = *orthogonal complement of* $u(t)$ = *kernel* $(u(t)^T)$, *and "conical neighborhood* C_y *of* y" *is defined as below.*

Part 4 is more technical than the others and helps to bridge the gap between parts 1 and 2 in the proof. It says, intuitively, that $u(t)$ is bounded away from each direction for a sufficient part of time over any reasonably long period of time. However, it is formulated to say that $u(t)^\perp$ is bounded away from any unit direction, which is actually more to the point.

Let ∂S_r denote a sphere of radius r about 0 and S_r a ball of radius r about 0. Thus $\partial S_r = \{x \in R^n | \, |x| = r\}$ and $S_r = \{x \in R^n | \, |x| \leqq r\}$. By a "conical neighborhood C_y^α for y" we mean that α is an open subset of the unit sphere $\partial S_1 \subseteq R^n$, $y/|y| \in \alpha$ or $-y/|y| \in \alpha$ if $y \neq 0$, and C_y^α is defined to be the union of all lines through 0 in R^n that intersect α. The width of C_y^α is defined to be the diameter of α. For simplicity, we sometimes omit the α and write C_y instead of C_y^α. (See Fig. 2.)

By $f : [0, \infty) \to R^1$ piecewise continuous, we mean that there is a decomposition of $[0, \infty)$ into half-open intervals, $[0, \infty) = \bigcup_{n=1}^{\infty} [a_n, a_{n+1})$ such that the restriction $u|(a_n, a_{n+1})$ is continuous for all n.

This completes the statement of Theorem 1.

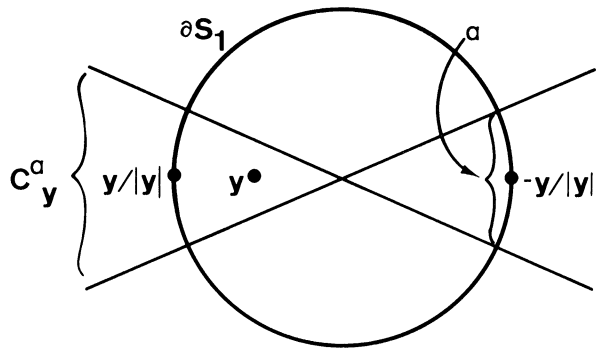


FIG. 2

The Theorem in § 1 asserts the equivalence of parts 1 and 2 of Theorem 1, except that only one of the three formulations of part 2 is given there. We will present an explicit proof later that each of these formulations implies the other two. In practice, it would seem that the equivalence of parts 1 and 2 would be the most useful implication of this theorem, as is illustrated in § 2. We should also note that the equivalence of part 2 and the eigenvalue condition (part 3) is not hard to show.

After the acceptance of this paper for publication, it was pointed out to the authors that Anderson in [1, p. 2.13], for the case that $u(t)$ is almost periodic, had established results from which the $2 \Rightarrow 1$ part of Theorem 1 could be derived.

The following key Lemma will be applicable to many cases besides those discussed in this paper. To indicate this, we present some corollaries after the statement of the Lemma, but first we need a definition.

DEFINITION. A function $\phi : [0, \infty) \to [0, \infty)$ is said to belong to class K, $\phi \in K$, if it is continuous, strictly increasing and $\phi(0) = 0$.

LEMMA. *Let $f(x, t) : S_\varepsilon \times [0, \infty) \to R^n$ be continuous in x and piecewise continuous in t with $f(0, t) = 0$ for all $t \geqq 0$, where $\varepsilon > 0$ is some fixed constant. Assume*

1. *there is $\phi_1 \in K$ such that*

$$|f(x, t) - f(y, t)| \leqq \phi_1(x - y) \quad \text{for all } x, y \in S_\varepsilon, \quad t \geqq 0,$$

2. *there are real numbers $a > 0$ and b and $\phi_2 \in K$ such that*

$$\int_{t_0}^{t} |f(x, s)|\, ds \geqq \phi_2(|x|)[a(t - t_0) + b]$$

for all fixed $x \in S_\varepsilon$ and $t \geqq t_0 \geqq 0$,

3. *there is a continuous differentiable function $V : S_\varepsilon \times [0, \infty) \to [0, \infty)$ and $\phi_3 \in K$ such that $\phi_3(|x|) \geqq V(x, t) \geqq 0$ and $\dot{V}(x, t) \leqq 0$ for all $t \geqq 0$ and $x \in S_\varepsilon$ where*

$$\dot{V}(x, t) = \frac{\partial V}{\partial t}(x, t) + \nabla V(x, t) \cdot f(x, t),$$

4. *there is a $\phi_4 \in K$ such that $-\dot{V}(x, t) \geqq |f(x, t)|^2 \cdot \phi_4(|x|)$ for all $x \in S_\varepsilon$, $t \geqq 0$,*
5. *the solution $x = 0$ of the equation $\dot{x} = f(x, t)$ is uniformly stable.*

Then the solution $x = 0$ is uniformly asymptotically stable.

Remarks. 1. Condition 1 is satisfied if $f(x, t) = A(t)x$ and $|A(t)| \leqq M$ for some constant M, all t. It is also satisfied if f is differentiable in x and its derivative with respect to x is bounded uniformly in t.

2. Intuitively, something like condition 2 seems necessary for u.a.s. However, it probably is not necessary as written.

3. Since Lyapunov function converse theorems for uniform asymptotic stability exist, condition 3 is very natural. (See Hale [4, Chap. X].)

4. We know from Krasovskii's theorem that if $\dot{x} = A(t)x$ is u.a.s., then a quadratic Lyapunov function exists (Narendra and Taylor, [9, p. 62]). In this case, if $|A(t)|$ is uniformly bounded, it is easy to see that we can choose ϕ_3 to make condition 4 hold. Thus, for $f(x, t)$ linear and V quadratic, condition 4 is necessary for u.a.s.

5. If there is a $\phi \in K$ such that $V(x, t) \geqq \phi(|x|)$ for all x and t, then uniform stability (condition 5) follows.

DEFINITION. $A \geqq B$ means $A - B$ is positive semi-definite.

COROLLARY 1. *If $f(x, t) = -P(t)x$ where $P(t)$ is a symmetric positive definite uniformly bounded matrix and if there are real numbers $a > 0$ and b such that*

$$\int_{t_0}^{t} |P(s)w| \, ds \geqq a(t - t_0) + b$$

for all $t \geqq t_0 \geqq 0$ and all fixed unit vectors w, then $\dot{x} = -P(t)x$ is u.a.s.

Proof. Applying the Lemma, conditions 1 and 2 are immediate. Letting $V(x) = |x|^2$, *we have* $\dot{V}(x, t) = -x^T P(t)x \leqq 0$ so conditions 3 and 5 are also easy. Condition 4 follows because $0 \leqq P(t) \leqq I$ implies $P(t)^2 \leqq P(t)$ for symmetric $P \geqq 0$. (We may as well assume $P(t) \leqq I$.) Thus $-\dot{V}(x, t) = x^T P(t)x \geqq x^T P(t)^2 x = |P(t)x|^2$. Q.E.D.

COROLLARY 2. *Suppose $\dot{x} = A(t)x$ is uniformly stable, $A(t)$ is uniformly bounded, and there are real numbers $a > 0$ and b such that*

$$\int_{t_0}^{t} |A(s)w| \, ds \geqq a(t - t_0) + b$$

for all $t \geqq t_0$ and all unit vectors w. Assume there is a positive definite $Q(t)$ uniformly bounded such that

$$-(Q(t)A(t) + A(t)^T Q(t) + \dot{Q}(t)) \geqq cA(t)^T A(t)$$

for all t where c is some positive constant. Then $\dot{x} = A(t)x$ is u.a.s.

Proof. Let $V(x, t) = x^T Q(t)x$. Then the result follows immediately from the Lemma. Q.E.D.

4. Asymptotic stability. We now consider the asymptotic (nonuniform) stability of (1). Theorem 2 provides sufficient conditions for asymptotic stability. The relation between Theorems 1 and 2 is discussed at the end of this section.

THEOREM 2. *Let $u : [0, \infty) \to R_k^n$ be piecewise continuous and bounded. Then*

$$\dot{x} = -u(t)u(t)^T x$$

is asymptotically stable if there are n linearly independent unit vectors $y_1, \cdots, y_n$

with closed disjoint conical neighborhoods $C_{y_1}, C_{y_2}, \cdots, C_{y_n}$ *such that*

$$\int_{\Lambda_i} |u(s)|^2 \, ds = \infty$$

for $i = 1, 2, \cdots, n$, *where* $\Lambda_i = \{s \in [0, \infty) | u(s) \cap C_{y_i} \neq 0\}$. *By* C_{y_i} *closed, we mean that* C_{y_i} *is the closure of an open conical neighborhood.*

Since Theorem 2 gives only sufficient conditions for asymptotic stability, we present the following as a step in the direction of obtaining necessary conditions.

PROPOSITION 3. *Let* $u : [0, \infty) \to R^n$ *be bounded piecewise differentiable with* $|u_1(t)|$ *bounded away from zero where* $u(t) = (u_1(t), \cdots, u_n(t))$. *If*

$$\int_0^\infty |\dot{u}(s)| \, ds < \infty,$$

then $\dot{x} = -u(t)u(t)^T x$ *is not asymptotically stable.*

Examples and Comments. 1. We now see that u' in § 2(d) yields an asymptotically stable system, even though not u.a.s.

2. The one-dimensional equation $\dot{x} = -(1/(1+t))x$ obeys

$$\int_0^\infty \frac{1}{1+t} \, dt = \ln (1+t)|_0^\infty = \infty$$

and so is a.s. but not u.a.s. Note that this example indicates why something like the conditions of the Lemma are required. With $V = x^2$, $\dot{V} = -2x^2/(1+t)$; with $V = (1+t)x^2$, $\dot{V} = -x^2$.

However, Proposition 3 shows that $\dot{x} = -u(t)u(t)^T x$ where $u(t) = (1, (1+t)^{-\alpha})$ with $\alpha > 0$ is not a.s. This is because

$$\dot{u}(t) = (0, -\alpha(1+t)^{-\alpha-1})$$

and

$$\int_0^\infty |\dot{u}(s)| \, ds = \alpha \int_0^\infty (1+s)^{-\alpha-1} \, ds = -(1+s)^{-\alpha}|_0^\infty = 1.$$

In particular, the above applies to $u(t) = (1, 1/(1+t))$, $u(t) = (1, 1/\sqrt{1+t})$, and $u(t) = (1, 1/(1+t)^2)$.

3. The condition of Theorem 2 is roughly similar to condition 4 of Theorem 1. (Note, however, the difference in the definitions of Ω_y and Λ_{y_i}.) The question of whether there are conditions implying asymptotic stability analogous to parts 2 and 3 of Theorem 1 is interesting. Letting $u(t) = (1, 1/\sqrt{1+t})$ as in example 2 above shows that

$$\int_0^\infty y^T u(t)u(t)^T y \, dt = \infty$$

does not imply $\dot{x} = -u(t)u(t)^T x$ is a.s. Also, it would be very useful to have a nonuniform version of the lemma for Theorem 1.

4. Proposition 3 suggests the following as a conjecture. Let $u:[0,\infty)\to R_k^n$ be bounded, piecewise differentiable with $k<n$ and $|u(s)|\neq 0$ for all s. If

$$\int_0^\infty \left|\frac{\dot{u}(s)}{u(s)}\right| ds<\infty,$$

then $\dot{x}=-u(t)u(t)x$ is not a.s.

5. Proofs. In this section we prove the Lemma and Theorem 1.

Proof of Lemma. 1. The hypotheses of the Lemma hold in a ball of radius ε about the origin. Since $\dot{x}=f(x,t)$ is uniformly stable, there is an $\varepsilon_1>0$ such that if $x(t)$ is a solution and $|x(t_0)|\leqq\varepsilon_1$, then $|x(t)|\leqq\varepsilon$ for all $t\geqq t_0$. Fix this ε_1.

2. All that is required to establish u.a.s. is to show that, given ε_2 with $0<\varepsilon_2<\varepsilon_1$, there exists a $T(\varepsilon_2)>0$ such that for all $t_0\geqq 0$, $|x(t_0)|<\varepsilon_1$ implies $|x(t_0+t')|<\varepsilon_2$ for some $t'\in[t_0, t_0+T(\varepsilon_2)]$.

3. Now we state two fundamental inequalities. Let $x(t)$ be a solution to $\dot{x}=f(x,t)$ such that $|x(t_0)|\leqq\varepsilon_1$ and $|x(s)|\geqq\varepsilon_2$ for $s\in[t_1,t_2]$ where $0<\varepsilon_2<\varepsilon_1$ and $0\leqq t_0\leqq t_1\leqq t_2$. Let $L(t_1,t_2)$ denote the arc length of $x(s)$ from t_1 to t_2. Then

$$L^2(t_1,t_2)\leqq\beta^2(t_2-t_1),\quad \text{where } \beta^2=\phi_3(\varepsilon_1)/\phi_4(\varepsilon_2), \tag{2}$$

$$\phi_2(\varepsilon_2)[a(t_2-t_1)+b]-\phi_1(\xi(t_1,t_2))[t_2-t_1]\leqq L(t_1,t_2), \tag{3}$$

where

$$\xi(t_1,t_2)=\max\{|x(s)-x(t_1)|;s\in[t_1,t_2]\}.$$

We postpone the proof of (2) and (3) until after completing the proof of the Lemma.

Define $\delta\equiv\phi_1^{-1}(\frac{1}{2}a\phi_2(\varepsilon_2))$ and

$$\gamma\equiv\left[\frac{2\beta}{a\phi_2(\varepsilon_2)}\right]^2+b^2.$$

Then, combining (2) and (3) into a single inequality eliminating $L(t_1,t_2)$, we see that $\xi(t_1,t_2)<\delta$ for $t_2-t_1>\gamma$ is a contradiction. Thus $t_2-t_1>\gamma$ implies $\xi(t_1,t_2)\geqq\delta$. Note that neither γ nor δ depends on t_0.

4. Now assume $|x(s)|\geqq\varepsilon_2$ for $s\in[t_0,t]$ and use (2) with t_1 replaced by t_0 and t_2 replaced by t. This yields $\beta\sqrt{t-t_0}\geqq L(t_0,t)$. Now clearly $L(t_1,t_2)\geqq\xi(t_1,t_2)$ for any $t_1\leqq t_2$. Also we have $\xi(t_1,t_2)\geqq\delta$ if $t_2-t_1\geqq\gamma$ by part 3 of the proof above.

Therefore, if $t-t_0=m\gamma$ for some positive integer m we have

$$\beta\sqrt{m\gamma}=\beta\sqrt{t-t_0}\geqq L(t_0,t)=\sum_{k=0}^{m-1}L(t_k,t_{k+1})\geqq\sum_{k=0}^{m-1}\delta=m\cdot\delta,$$

where $t_{k+1}-t_k=\gamma$ and $t_m=t$. This yields

$$\frac{\beta^2\gamma}{\delta^2}\geqq m$$

which puts an upper bound on the length of an interval $[t_0,t]$ with $|x(s)|\geqq\varepsilon_2$ for $s\in[t_0,t]$. Since this upper bound is independent of t_0, we have established the

existence of $T(\varepsilon_2)$. Specifically, we may define

$$T(\varepsilon_2) \equiv \frac{\beta^2\gamma^2}{\delta^2} + \gamma.$$

5. We now prove (2). Using the integral expression for arc length and the Cauchy–Schwarz inequality, we obtain

$$L(t_1, t_2) = \int_{t_1}^{t_2} |\dot{x}(s)|\, ds$$

$$= \int_{t_1}^{t_2} |f(x(s), s)|\, ds \leqq \sqrt{\int_{t_1}^{t_2} |f(x(s), s)|^2\, ds}\sqrt{t_2 - t_1}.$$

But, by hypothesis 4 in the statement of the Lemma, we have

$$\phi_4(\varepsilon_2) \int_{t_1}^{t_2} |f(x(s), s)|^2\, ds \leqq -\int_{t_1}^{t_2} \dot{V}(x(s), s)\, ds$$

$$= V(x(t_1), t_1) - V(x(t_2), t_2) \leqq V(x(t_1), t_1)$$

$$\leqq \phi_3(|x(t_1)|) \leqq \phi_3(\varepsilon_1).$$

Now (2) follows by combining the above two inequalities.

6. We now prove (3). Consider

$$\int_{t_1}^{t_2} |f(x(t_1), s) - f(x(s), s)|\, ds \geqq \int_{t_1}^{t_2} |f(x(t_1), s)|\, ds - \int_{t_1}^{t_2} |f(x(s), s)|\, ds$$

$$\geqq \phi_2(|x(t_1)|)[a(t_2 - t_1) + b] - \int_{t_1}^{t_2} |f(x(s), s)|\, ds$$

$$\geqq \phi_2(\varepsilon_2)[a(t_2 - t_1) + b] - L(t_1, t_2)$$

and also

$$\int_{t_1}^{t_2} |f(x(t_1), s) - f(x(s), s)|\, ds \leqq \int_{t_1}^{t_2} \phi_1(|x(t_1) - x(s)|)\, ds$$

$$\leqq \int_{t_1}^{t_2} \phi_1(\xi(t_1, t_2))\, ds = \phi_1(\xi(t_1, t_2))[t_2 - t_1].$$

Now (3) follows by combining the above two inequalities.

This completes the proof of the Lemma.

Proof of Theorem 1. For simplicity assume $|u(s)| \leqq 1$ for all s. We shall show the equivalence of the four parts of the theorem by proving in succession that 2 implies 1, 4 implies 2 and 1 implies 4 and that 2 and 3 imply each other.

(i) $2 \Rightarrow 1$. This will follow from Corollary 1 to the Lemma, once we show that the three formulations of condition 2 are equivalent.

Claim. The following are equivalent:

(a)

$$\int_{t_0}^{t} y^T u(s) u^T(s) y \, ds \geqq a(t-t_0)+b \quad \text{for some constants } a>0,\ b \text{ and all unit vectors } y.$$

(b)

$$\int_{t_0}^{t} |u(s)u^T(s)y| \, ds \geqq a'(t-t_0)+b' \quad \text{for some constants } a'>0,\ b' \text{ and all unit vectors } y.$$

(c)

$$\int |u(s)^T y| \, ds \geqq a''(t-t_0)+b'' \quad \text{for some constants } a''>0,\ b'' \text{ and all unit vectors } y.$$

Proof of Claim. First observe that $y^T u u^T y = |u^T y|^2$. Also $|y|=1$ and $|u(s)| \leqq 1$ implies $|y^T u(s) u(s)^T y| \leqq |u(s) u(s)^T y| \leqq |u(s)^T y|$. Thus (a)$\Rightarrow$(b), (a)$\Rightarrow$(c), and (b)$\Rightarrow$(c) follow at once. We need only show (c)$\Rightarrow$(a). This follows because if $\int_{t_0}^{t} |f(s)| \, ds \geqq a(t-t_0)+b$, then

$$a^2(t-t_0)^2 + ab(t-t_0) \leqq \left(\int_{t_0}^{t} |f| \, ds\right)^2 \leqq \left(\int_{t_0}^{t} f^2 \, ds\right)(t-t_0)$$

by the Cauchy–Schwarz inequality. Q.E.D.

(ii) 4$\Rightarrow$2. Assume condition 4 holds. We then have the conical open cover for R^n, $\{C_y | y \in R^n,\ y \neq 0\}$, with associated $\{\Omega_y\}$ as given by condition 4.

The C_y cover ∂S_1, the unit sphere. Choose a finite subcover $C_{y_1}, \cdots, C_{y_m}$. Fix $y \in \partial S_1$. Then $y \in C_{y_{i_0}}$ for some i_0. Then there is an $\varepsilon > 0$ such that $|u(s)^T y|^2 = y^T u(s) u(s)^T y \geqq \varepsilon |u(s)|^2$ for all $s \in [0,\infty) - \Omega_{y_{i_0}}$. This is because, for such an s, $u(s)^{\perp} = \ker(u(s)^T)$ is bounded away from y. (See Fig. 3.)

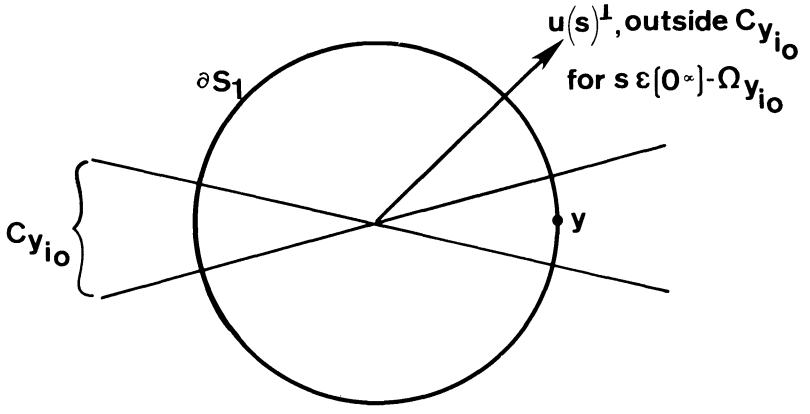


FIG. 3

Thus

$$\int_{t_0}^{t} y^T u(s)u(s)^T y\,ds \geqq \int_{[t_0,t]-\Omega_{y_{i_0}}} y^T u(s)u(s)^T y\,ds$$

$$\geqq \varepsilon \int_{[t_0,t]-\Omega_{y_{i_0}}} |u(s)|^2\,ds \geqq \varepsilon[a_{y_{i_0}}(t-t_0)+b_{y_{i_0}}].$$

This inequality is valid for all unit vectors in some small neighborhood of y. By compactness of ∂S_1, we conclude

$$\int_{t_0}^{t} y^T u(s)u(s)^T y\,ds \geqq a(t-t_0)+b \quad \text{for some } a>0,\ b \text{ and all } t \geqq t_0 \geqq 0,$$
$$\text{all } y \in \partial S_1.$$

(iii) 1⇒4. Assume $\dot{x} = -u(t)u(t)^T x$ is u.a.s.

(a) Suppose condition 4 is false. Then there is some $w \in \partial S_1$ such that for every conical neighborhood C_w^α, the following holds:

Given any $N>0$ and $\varepsilon>0$, there are t_1 and t_2 such that $t_2 - t_1 \geqq N$ and

$$\int_{[t_1,\,t_2]-\Omega_w^\alpha} |u(s)|^2\,ds < \varepsilon,$$

where $\Omega_w^\alpha = \{t \in [0,\infty) | u(t)^\perp \cap C_w^\alpha \neq 0\}$.

(b) By u.a.s. of $\dot{x} = -u(t)u(t)^T x$, given $\varepsilon_1 > \varepsilon_2 > 0$, there is a $\gamma > 0$ such that if $x(t)$ is a solution and $x(t_o) \in S_{\varepsilon_1}$, then $x(t) \in S_{\varepsilon_2}$ if $t \geqq t_0 + \gamma$.

Let $\varepsilon_1 = 1$, $\varepsilon_2 = \frac{1}{2}$, and choose γ for these ε_1, ε_2.

(c) By (a), for any $\varepsilon > 0$ and conical neighborhood C_w^α, there are t_1 and t_2 such that $t_2 - t_1 = \gamma$ and

$$\int_{[t_1,t_2]-\Omega_w^\alpha} |u(s)|^2\,ds < \varepsilon.$$

(This w is the one fixed in (a).)

(d) Let $w^\perp$ denote an $n \times (n-1)$ matrix which consists of columns which are a basis for the orthogonal complement to w, an $(n-1)$-dimensional hyperplane. Define $v(t) = w^\perp \cdot ((w^\perp)^T \cdot u(t)) =$ "projection of $u(t)$ onto $w^\perp$". If $u(t)$ is $n \times k$, then $v(t)$ is $n \times k$ also. If $u(t)$ is "close to $w^\perp$," then $v(t)$ is "close to $u(t)$."

The equation $\dot{x} = -v(t)v(t)^T x$ has stationary solutions (any initial condition on the line through w). We shall show that $\dot{x} = -u(t)u(t)^T x$ is close enough to $\dot{x} = -v(t)v(t)^T x$ to have "almost stationary" solutions, at least to an extent sufficient to contradict u.a.s.

(e) If $\dot{x} = A(t)x$, $\dot{y} = [A(t)+B(t)]y$, and $x(t)$, $y(t)$ are respective solutions with $x(t_0) = y(t_0)$, then

$$y(t) = x(t) + \int_{t_0}^{t} X(t)X(s)^{-1}B(s)y(s)\,ds,$$

where $X(t)$ is a fundamental matrix solution for $\dot{x} = A(t)x$.

Let $A(t) = -u(t)u(t)^T$ and $B(t) = [-v(t)v(t)^T] - A(t)$. Define the constant function $x(t) = w$, and note that $x(t)$ is a constant solution for $\dot{x} = A(t)x$.

Then, for any initial t_0, we have solution $y(t)$ with $y(t_0) = w$, and

$$y(t) = w + \int_{t_0}^{t} X(t)X(s)^{-1}B(s)y(s)\,ds.$$

Now

$$\left|\int_{t_0}^{t} X(t)X(s)^{-1}B(s)y(s)\,ds\right|$$

$$\leqq \int_{t_0}^{t} |X(t)X(s)^{-1}|\,|B(s)|\,|y(s)|\,ds \leqq \int_{t_0}^{t} |B(s)|\,ds,$$

since $|X(t)X(s)^{-1}| \leqq |w| = 1$ and $|y(t)| \leqq |y(t_0)| = |w| = 1$.

(f) $$|B(t)| = |-v(t)v(t)^T - A(t)| \leqq |v(t)v(t)^T| + |u(t)u(t)^T|$$
$$\leqq |u(t)|^2 + |u(t)|^2 = 2|u(t)|^2.$$

Choose C_w^α of width less than $1/(8\gamma)$; i.e., if $z_1, z_2 \in C_w^\alpha$ are unit vectors, then $|z_1 - z_2| \leqq 1/(8\gamma)$. Then, for $t \in \Omega_w^\alpha$, $|B(t)| < 1/(8\gamma)$, because $u(t)^\perp \cap C_w^\alpha \neq 0$. (See Fig. 4.) In other words, since $u(t)^\perp$ is close to w, we may conclude that $v(t) \equiv$ "projection of $u(t)$ on $w^\perp$" is close to $u(t)$. Then it follows that $v(t)v(t)^T$ is close to $u(t)u(t)^T$.

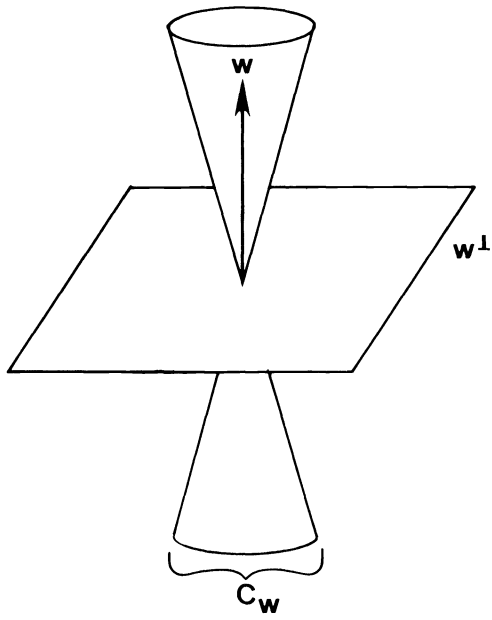


FIG. 4a

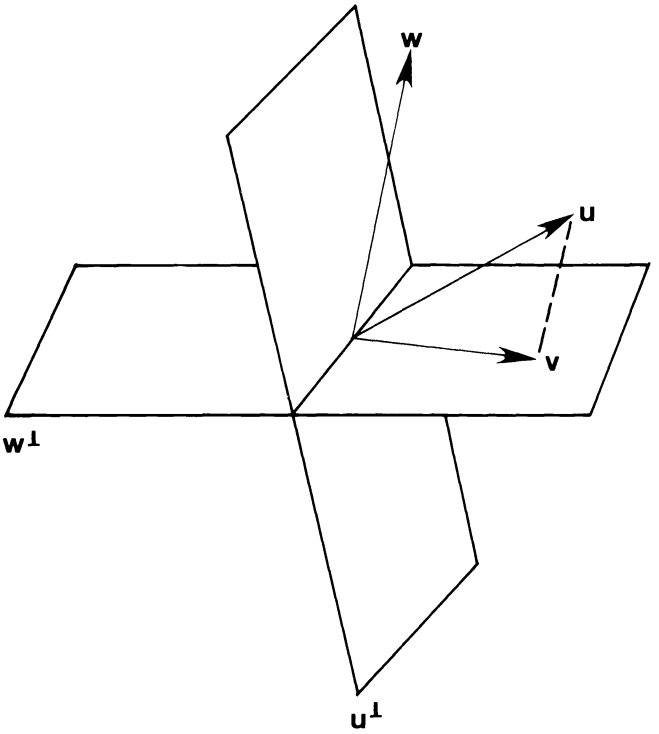


FIG. 4b. *The rank of u equals* 1

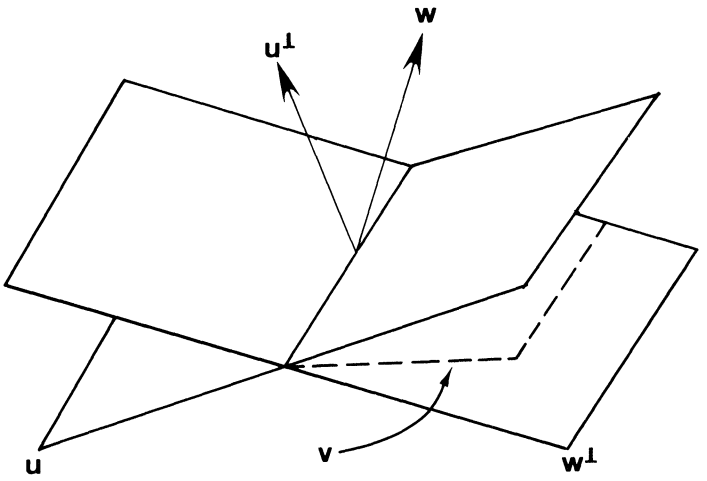


FIG. 4c. *The rank of u equals* 2

By (c) we can choose t_1 and t_2 with $t_2 - t_1 = \gamma$ and

$$\int_{[t_1,t_2]-\Omega_w^\alpha} |u(s)|^2\, ds < \frac{1}{16}.$$

Now, letting $t_0 = t_1$ in (e), we have

$$\begin{aligned}\left|\int_{t_1}^{t_2} X(t)X(s)^{-1}B(s)y(s)\, ds\right| &\leqq \int_{t_1}^{t_2} |B(s)|\, ds \\ &= \int_{[t_1,t_2]\cap\Omega_w^\alpha} |B(s)|\, ds + \int_{[t_1,t_2]-\Omega_w^\alpha} |B(s)|\, ds \\ &\leqq \int_{[t_1,t_2]\cap\Omega_w^\alpha} \frac{1}{8\gamma}\, ds + 2\int_{[t_1,t_2]-\Omega_w^\alpha} |u(s)|^2\, ds \\ &\leqq \frac{t_2-t_1}{8\gamma} + 2\cdot\frac{1}{16} = \frac{\gamma}{8\gamma} + \frac{1}{8} = \frac{1}{4}.\end{aligned}$$

Thus

$$\begin{aligned}|y(t_2)| &\geqq \left|w + \int_{t_1}^{t_2} X(t)X(s)^{-1}B(s)y(s)\, ds\right| \\ &\geqq |w| - \left|\int_{t_1}^{t_2} X(t)X(s)^{-1}B(s)y(s)\, ds\right| \geqq 1 - \frac{1}{4} = \frac{3}{4} > \frac{1}{2}.\end{aligned}$$

This contradicts the choice of γ, which was based on u.a.s. of $\dot{x} = -u(t)u(t)^T x$. In particular, $|y(t_2)|$ should be less than $\frac{1}{2}$.

(iv) $2 \Leftrightarrow 3$. The smallest eigenvalue of $A = \int_{t_0}^{t} u(s)u(s)^T\, ds$ is equal to

$$\inf_{|y|=1} \{y^T A y\} = \inf_{|y|=1} \left(\int_{t_0}^{t} y^T u(s)u(s)^T y\, ds\right).$$

The equivalence of 2 and 3 is now obvious. Q.E.D.

6. Proofs of Theorem 2 and Proposition 3. In this section we prove Theorem 2 and Proposition 3. First, however, we sketch the proof of Theorem 2 and discuss briefly its relation to the Lemma for Theorem 1.

Outline of proof for Theorem 2. Let $x(t)$ be a solution. We want $x(t) \to 0$ as $t \to \infty$. To get a contradiction, suppose the length of $x(t)$ is bounded away from 0. Then there are two possibilities.

(i) $x(t)$ is eventually bounded away from some $y_i^\perp = \{x \in R^n | \langle x, y_i\rangle = 0\} =$ an $(n-1)$-dimensional "hyperplane."

(ii) $x(t)$ gets close to each $y_i^\perp$, $i = 1, 2, \cdots n$, repeatedly as $t \to \infty$.

If the first possibility occurs, it is easy to show that we get a contradiction. If the second holds, then we argue as follows. First $x(t)$ must repeatedly spend a certain minimal amount of time away from all the $y_i^\perp$. For this time we may relate the decrease in $\dot{V}(x(t))$ to the increase in arc length of $x(t)$. We conclude that $\dot{V}(x(t))$ decreases by a certain fixed increment as $x(t)$ "travels the circuit" to each of the $y_i^\perp$. Since $\dot{V}(x(t))$ is bounded below, this also leads to a contradiction.

It is reasonable to suggest that the Lemma in § 3 for the special case $f(x, t) = -u(t)u(t)^T x$ could be established by a proof analogous to that of Theorem 2. However, it does not appear that this proof would work for the general case, and also this proof is not as simple as the one given for the Lemma. Since the proof of the lemma does not seem to be adaptable to the nonuniform case, we have chosen not to attempt a unified proof of the two results.

Proof of Theorem 2. For simplicity, assume $|u(s)| \leqq 1$ for all s and denote C_{y_i} by C_i.

(a) We have defined $y_i^\perp = \{x \in R^n | \langle x, y_i \rangle = 0\}$ to be an $(n-1)$-dimensional subspace of R^n. We extend the definition of "conical neighborhood" by defining a conical neighborhood D_i of $y_i^\perp$ by

$$D_i = [C_{y_i}]^\perp = \{x \in R^n | \langle x, y \rangle = 0 \text{ for some } y \in C_{y_i}\}.$$

Then it follows that D_i is the union of all lines in R^n intersecting a neighborhood β of $y_i^\perp \cap \partial S_1$ in ∂S_1. (∂S_1 is the unit $(n-1)$-dimensional sphere in R^n; $y_i^\perp \cap \partial S_1$ is an $(n-2)$-dimensional "subsphere" (a "great circle"); β is an open subset of ∂S_1 that contains $y_i^\perp \cap \partial S_1$). In fact, if $C_i = C_{y_i}^\alpha$ where α is an open neighborhood of y_i in ∂S_1, then $\beta = \alpha^\perp \cap \partial S_1 = \{x \in \partial S_1 | \langle x, y \rangle = 0 \text{ for some } y \in \alpha\}$. (See Fig. 5).

It is clear that if $u(t) \cap C_i \neq 0$, then $u(t)^\perp \in D_i$. Since the y_i are linearly independent, we may further assume that $D_1 \cap D_2 \cap \cdots \cap D_n = 0$.

(b) Now if $t \in \Lambda_i$, then $u(t) \cap C_i \neq 0$ and therefore $u(t)^\perp \in D_i$. Expand C_i and D_i slightly to closed conical neighborhoods C_i^* and D_i^* so that interior $(C_i^*) \supseteq C_i$, interior $(D_i^*) \supseteq D_i$, $C_i^* \cap C_j^* = 0$ if $i \neq j$, and $D_1^* \cap \cdots \cap D_n^* = 0$. Do not change Λ_i. Then if $t \in \Lambda_i$, we have $u(t)^\perp \in D_i^*$ and bounded away from the boundary of D_i^*. Therefore, we may conclude that there is an $\varepsilon > 0$ such that if $t \in \Lambda_i$ and $x \notin D_i^*$, then $|u(t)^T x| \geqq \varepsilon |x|\,|u(t)|$. This is because, for $t \in \Lambda_i$ and $x \notin D_i^*$, x is bounded away from $u(t)^\perp = \ker (u(t)^T)$.

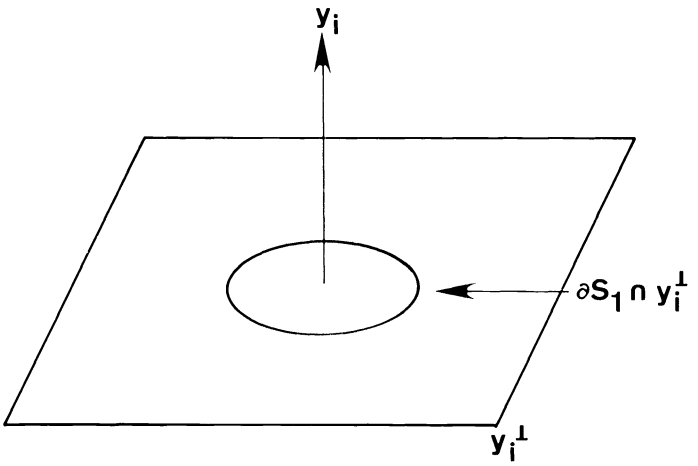


FIG. 5a

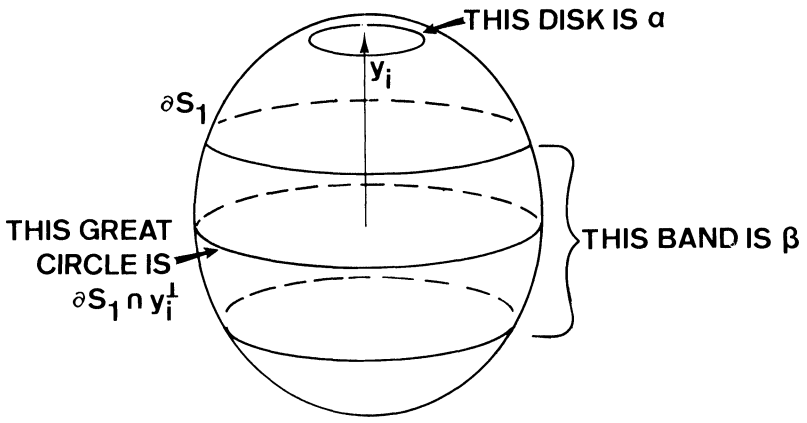


FIG. 5b

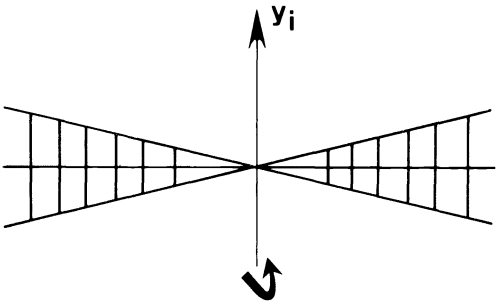


FIG. 5c. *The surface of revolution of the shaded part of the figure equals* $D_i = C_i^\perp$

(c) Thus if $x(t)$ is a solution that is eventually not in some D_i^* (say for $t \geqq t_0$), we have

$$-\int_{t_0}^{\infty} \dot{V}(x(s))\, ds \geqq \int_{\Lambda_i} |u(s)^T x(s)|^2\, ds \geqq \int_{\Lambda_i} \varepsilon^2 |u(s)|^2 |x(s)|^2\, ds$$

which is unbounded if $|x(s)| \geqq \alpha > 0$ for some α. This would be a contradiction, so $|x(s)| \to 0$.

(d) Suppose $x(t)$ is a solution which enters each D_i^*, $i = 1, \cdots, n$, repeatedly as $t \to \infty$. Suppose $|x(s)| \geqq \alpha > 0$.

Now letting $D = D_1^* \cup D_2^* \cup \cdots \cup D_n^*$, we conclude $x(t)$ must spend a minimal amount of time in $R^n - D$ when it travels to all of the D_i^*, $i = 1, \cdots, n$. This is because, in going to each of the D_i^*, $x(t)$ must cover a minimal distance in $R^n - D$. Since $|\dot{x}(t)|$ is bounded above, this implies $x(t)$ must spend at least a fixed amount of time in $R^n - D$. Thus we have a minimal distance δ and a minimal time γ.

Without loss of generality we have the following. If $x(t)$ travels to all the D_i^* as $t \in [c, d]$, then there is $[a, b] \leqq [c, d]$ with $b - a \geqq \gamma$ such that $x(t) \in R^n - D$ for all $t \in [a, b]$ and the arc length of $x(t)$ from a *to* b is at least δ.

Note that when $x(t) \in R^n - D$, we have $|u(t)^T x(t)| \geqq \varepsilon |u(t)^T|\,|x(t)| \geqq \varepsilon\alpha |u(t)|$

(e) Let $\Omega = \{t \in [0, \infty) | x(t) \in R^n - D\}$. Then,

$$\int_\Omega |u(s)^T x(s)|^2\, ds \geqq \varepsilon^2 \alpha^2 \int_\Omega |u(s)|^2\, ds.$$

Thus

$$\int_\Omega |u(s)|^2\, ds \leqq \frac{1}{\varepsilon^2 \alpha^2} \int_\Omega |u(s)^T x(s)|^2\, ds.$$

(f) If $[a, b] \subseteq \Omega$, then $L(a, b) \equiv$ arc length of $x(s)$ from a to $b =$

$$\int_a^b |\dot{x}(s)|\, ds = \int_a^b |u(s)u(s)^T x(s)|\, ds \leqq \int_a^b |u(s)|\,|u(s)^T x(s)|\, ds.$$

Applying the Cauchy–Schwarz inequality, we get

$$\begin{aligned}[L(a, b)]^2 &\leqq \int_a^b |u(s)|^2\, ds \int_a^b |u(s)^T x(s)|^2\, ds \\ &\leqq \frac{1}{\varepsilon^2 \alpha^2} \int_a^b |u(s)^T x(s)|^2\, ds \int_a^b |u(s)^T x(s)|^2\, ds\end{aligned}$$

by (e).

We conclude that if $x(s)$ enters each D_i^* as s ranges over values in $[c, d]$, then there is $[a, b] \subseteq [c, d] \cap \Omega \subseteq [c, d]$ such that

$$\int_{[c,d]} |u(s)^T x(s)|^2\, ds \geqq \int_{[a,b]} |u(s)^T x(s)|^2\, ds \geqq \varepsilon\alpha\delta.$$

(g) By assumption, $x(t)$ enters each D_i^* repeatedly. Therefore $[0, \infty) = \bigcup_{i=1}^{\infty} [c_i, d_i)$ where $x(t)$ enters each D_i^* as $t \in [c_i, d_i)$, and there are $[a_i, b_i) \subseteq [c_i, d_i) \cap \Omega \supseteq [c_i, d_i)$ as above.

Thus

$$\begin{aligned}-\int_0^\infty \dot{V}(x(s))\, ds &= \int_0^\infty |u(s)^T x(s)|^2\, ds = \int_{\cup [c_i, d_i)} |u(s)^T x(s)|^2\, ds \\ &= \sum_{i=1}^{\infty} \int_{[c_i, d_i)} |u(s)^T x(s)|^2\, ds \geqq \sum_{i=1}^{\infty} \int_{[a_i, b_i)} |u(s)^T x(s)|^2\, ds \\ &\geqq \sum_{i=1}^{\infty} \varepsilon\alpha\delta = \infty.\end{aligned}$$

Therefore, $|x(s)| \geqq \alpha > 0$ is false. Q.E.D.

Proof of Proposition 3. (a) We will use the technique of putting $u(t)u(t)^T$ into "L-diagonal form" as described by Cesari [2, p. 39]. We will find piecewise differentiable $P(t)$ such that

(i) $P(t)^{-1} u(t)u(t)^T P(t) = \Lambda(t) = \text{diag}\,(|u(t)|^2, 0, 0, \cdots, 0)$,

(ii) $|P(t)|$ and $|P(t)^{-1}|$ are bounded above, and

(iii) $$\int_0^\infty |P(t)^{-1} \dot{P}(t)| < \infty.$$

Then $\dot{x} = -u(t)u(t)^T x$ is asymptotically stable if and only if $\dot{x} = \Lambda(t)x + P(t)^{-1}\dot{P}(t)x$ is. It is easy to confirm that this second system is not asymptotically stable.

(b) Let

$$P(t) = \begin{bmatrix} u_1 & -u_2 & -u_3 & \cdot & \cdot & \cdot & -u_n \\ u_2 & u_1 & 0 & & & & \\ u_3 & 0 & u_1 & & & & \\ \cdot & 0 & 0 & u_1 & & & \\ \cdot & \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & \cdot & & & \cdot & \\ u_n & 0 & 0 & & & & u_1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & -u_2 & \cdot & \cdot & \cdot & -u_n \\ u_2 & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ u_n & & 0 & & & \end{bmatrix} + u_1 I.$$

Condition (i) is easy to check. The columns of P are eigenvectors for uu^T. Since $\det(P) = (-1)^{n-1}u_1^{n-2}|u|^2$, P^{-1} exists.

Condition (ii) follows from the fact that $|u(t)|$ is bounded above and $|u_1(t)|$ is bounded below.

Condition (iii) follows because $|P^{-1}(s)| \leqq k$ implies

$$\int_0^\infty |P^{-1}\dot{P}|\, ds \leqq \int_0^\infty |P^{-1}|\,|\dot{P}|\, ds \leqq k \int_0^\infty |\dot{P}(s)|\, ds$$

$$\leqq k \int_0^\infty |\dot{u}(s)|\, ds < \infty. \qquad \text{Q.ED.}$$

Acknowledgment. We would like to express our appreciation to Professor J. P. La Salle for his many helpful comments and especially for his suggestions on rewriting the proof of the Lemma to improve its readability.

Note. A proof of the converse of the lemma, in the sense that u.a.s. and 1 imply 2, has been discovered by the first named author.

REFERENCES

[1] B. ANDERSON, *Multivariate adaptive identification*, preprint, Dept. of Electrical Engineering, Univ. of Newcastle, New South Wales, Australia, 1974.

[2] L. CESARI, *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, 3rd ed., Springer-Verlag, Berlin, 1971.

[3] W. HAHN, *The Stability of Motion*, Springer-Verlag, Berlin, 1967.

[4] J. K. HALE, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[5] J. P. LASALLE, *Asymptotic stability criteria*, Proceedings of the Symposia in Applied Mathematics, Hydrodynamic Instability, vol. 13, American Mathematical Society, Providence, R.I., 1962, pp. 299–307.

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

[7] P. M. LION, *Rapid identification of linear and nonlinear systems*, AIAA J., 5 (1967), pp. 1835–1842.

[8] K. S. NARENDRA AND L. E. MCBRIDE, *Multiparameter self-optimizing systems using correlation techniques*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 31–38.

[9] K. S. NARENDRA AND J. H. TAYLOR, *Frequency Domain Criteria for Absolute Stability*, Academic Press, New York, 1973.

[10] M. REED AND B. SIMON, *Methods of Mathematical Physics 1: Functional Analysis*, Academic Press, New York, 1972.

[11] M. M. SONDHI AND D. MITRA, *New results on the performance of a well-known class of adaptive filters*, preprint, Bell Laboratories, Murray Hill, N.J., 1975.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

PROJECTION ON A CONE, PENALTY FUNCTIONALS AND DUALITY THEORY FOR PROBLEMS WITH INEQUALITY CONSTRAINTS IN HILBERT SPACE*

ANDRZEJ P. WIERZBICKI AND STANISLAW KURCYUSZ†

Abstract. Each element p of a real Hilbert space H can be uniquely decomposed into two orthogonal components, $p = p^D + p^{-D^*}$ where $p^D \in D$ is the projection of p on a closed convex cone D and p^{-D^*} is the projection of p on the minus dual cone $-D^*$. Hence, if the cone D generates a partial order in H, then the positive part p^D and the negative part p^{-D^*} of each $p \in H$ can be distinguished. For a general optimization problem: minimize $Q(y)$ over $Y_p = \{y \in E : p - P(y) \in D \subset H\}$, where $Q: E \to R$, $P: E \to H$, E is Banach, H is Hilbert: the violation of the constraint can be determined by $(p - P(y))^{-D^*}$. Hence a generalized penalty functional and an augmented Lagrange functional can be defined for this problem. The paper presents a short review of known penalty techniques, some properties of the projection on a cone, basic properties of penalty functionals for a general optimization problem and duality theory for nonconvex problems in infinite-dimensional spaces.

Properties of minimizing sequences in constrained optimization are discussed and the convergence of increased and shifted penalty techniques is studied in detail. Conditions of stability of the optimization problem, implying convergence conditions, are discussed in the closing section.

1. Introduction. R. Courant in [3] suggested that in order to solve the problem

$$\min_{y \in Y_0} Q(y);\ Y_0 = \{y \in R^n : P(y) = 0 \in R^m\}; \qquad Q: R^n \to R, \quad P: R^n \to R^m \tag{1.1}$$

a penalty function can be minimized

$$\Phi_0(y, \zeta) = Q(y) + \tfrac{1}{2}\zeta \|P(y)\|^2; \qquad \Phi_0: R^n \times R_+ \to R \tag{1.2}$$

for a sequence of parameters $\{\zeta_n\}$, $\zeta_n \to \infty$. This idea was later generalized—see Fiacco and McCormick [7]—for problems with inequality constraints

$$\min_{y \in Y_1} Q(y);\ Y_1 = \{y \in R^n : P(y) \leqq 0 \in R^m\}, \tag{1.3}$$

where the partial in R^m is generated by the positive orthant. The exterior penalty function for the problem (1.3) has the form

$$\Phi_1(y, \zeta) = Q(y) + \tfrac{1}{2}\zeta \sum_{i=1}^{m} P_i(y) \max(0, P_i(y)). \tag{1.4}$$

It is also possible [7] to define interior penalty functions for the problem (1.3), but these are not investigated in this paper.

The problem of minimizing a penalty function for large penalty coefficients ζ is badly conditioned numerically, since the spectral radius of the Hessian matrix of a penalty function increases with ζ. To overcome this difficulty, two equivalent

* Received by the editors July 16, 1974, and in revised form February 24, 1976. This work was supported by the National Science Foundation Grant GF-37298 to the Institute of Automatic Control, Technical University of Warsaw, Warsaw, Poland and the Center for Control Sciences, University of Minnesota, Minneapolis, Minnesota.

† Institute of Automatic Control, Technical University of Warsaw, Warsaw, Poland.

approaches have been developed independently. M. D. J. Powell in [20] introduced the shifted penalty function for the problem (1.1):

$$\Psi_0(y, \zeta, \nu) = Q(y) + \tfrac{1}{2}\zeta\|P(y) - \nu\|^2; \qquad \Psi_0 : R^n \times R_+ \times R^m \to R \tag{1.5}$$

which is minimized in respect to y for a sufficiently large ζ and a sequence of penalty shifts $\{\nu_n\}$. J. Szymanowski, A. Wierzbicki and others (see [25], [15]) investigated shifted penalty functions for problems with inequality constraints (1.3). M. R. Hestenes in [8] introduced the augmented Lagrange function for the problem (1.1):

$$\Lambda_0(\zeta, \eta, y) = Q(y) + \langle \eta, P(y)\rangle + \tfrac{1}{2}\zeta\|P(y)\|^2; \qquad \Lambda_0 : R_+ \times R^{m+n} \to R, \tag{1.6}$$

where the additional term $\frac{1}{2}\zeta\|P(y)\|^2$ "convexifies" the usual Lagrange function. R. T. Rockafellar in [40] introduced the augmented Lagrange function for the problem (1.3) and developed a duality theory for nonconvex problems. It should be noted that the minimization of the functions $\Psi_0(\cdot, \zeta, \nu)$ and $\Lambda_0(\zeta, \eta, \cdot)$ are equivalent, since

$$\Lambda_0(\zeta, -\zeta\nu, y) \equiv \Psi(y, \zeta, \nu) - \tfrac{1}{2}\zeta\|v\|^2, \tag{1.7}$$

but the augmented Lagrange function has useful properties, particularly in duality theory.

A penalty functional approach in infinite-dimensional problems has been applied by A. V. Balakrishnan, [2]. For the optimal control problem

$$\min Q(x, u) = \int_{t_0}^{t_1} f_0(x, u, t)\, dt + h(x(t_1)); \qquad \dot{x} = f(x, u, t); \quad x(t_0) = x_0 \tag{1.8}$$

a penalty functional has the form

$$\Phi_0\left(x, u, \frac{1}{\varepsilon}\right) = Q(x, u) + \frac{1}{2\varepsilon}\|\dot{x}(\cdot) - f(x(\cdot), u(\cdot), \cdot)\|^2, \tag{1.9}$$

where the norm is in $L^2(t_0, t_1)$; additional constraints can also be taken into consideration. The functional (1.9) results in the so-called ε-technique and in a computational approach to the maximum principle. The method of multipliers based on a functional similar to (1.9) in application to optimal and variational problems was discussed in the works of Rupp [42], [43].

A more abstract approach was used by Levitin and Poliak [13] for an optimization problem:

$$\min_{y \in A_0} Q(y); A_0 = \{y \in E : K(y) \leqq 0\}; \qquad Q : E \to R, \quad K : E \to R_+, \tag{1.10}$$

where E is a topological space or, more specifically, a Banach space. A general form of the penalty functional is then

$$\Psi(y, \zeta) = Q(y) + \zeta K(y). \tag{1.11}$$

One of the authors of this paper observed in [26] that for a problem with operator inequality constraints

$$\min_{y \in Y_p} Q(y); Y_p = \{y \in E : p - P(y) \in D \subset H\}; \qquad Q : E \to R, \quad P : E \to H, \tag{1.12}$$

where D is a positive cone in the Hilbert space H, the functional K can be defined by

$$K(y)=\tfrac{1}{2}\|(P(y)-p)^{D^*}\|^2, \tag{1.13}$$

where D^* is the dual cone and $(\cdot)^{D^*}$ is the projection on this cone. This approach has been developed in order to solve optimal control problems with state space constraints and, particularly, optimal control problems with delays and final complete state constraints. The corresponding shifted penalty techniques have been applied successfully to solve various optimal control problems [27]. However, the projection on a cone has many useful properties, which make it possible to develop a generalized theory of penalty functionals, augmented Lagrange functional and duality for nonconvex problems. The aim of this paper is to present an outline of this theory.

Beside the references cited above, a number of works have been devoted to the study of penalty functionals for various extremal problems. See, for instance, [34], [35], [36], [41]. The shifted penalty technique (often called the method of multipliers) has been recently investigated by numerous authors. Besides two important papers [15], [31], a good review of related problems along with a rather complete list of references is available by Bertsekas [32].

Part I. Fundamentals.

2. Projection on a cone and its properties. Let H be a Hilbert space, D a nonempty, convex closed set in H.

LEMMA 2.1 (see [5], [28]). *For any $p \in H$ there exists a unique element $p^D \in D$ satisfying*

$$\|p^D - p\| = \min_{d \in D} \|d - p\|. \tag{2.1}$$

The lemma holds also if H is a complete strictly normed space (if $\|x+y\| = \|x\|+\|y\|$ implies $x = \alpha y$, $\alpha \in R$). The element p^D is called the projection of p onto D, the mapping $(\cdot)^D$, the projection onto D.

Projections on linear subspaces play a fundamental role in functional analysis; but projection on more general convex sets and, in particular, on convex cones have been investigated relatively recently. A basic result, stated in Theorem 2.4 in this section, was announced by J. J. Moreau [18] in 1962. E. H. Zarantonello [29] used the projection on a cone to develop the spectral theory for a class of nonlinear operators. The application to penalty functional techniques have been introduced in [26]. The properties of a projection on a cone are presented here from the point of view of this application.

Throughout the paper, D is assumed to be a nonempty, closed convex cone in H with vertex at the origin, that is, $\alpha D + \beta D \subset D$ for $\alpha, \beta \geqq 0$. Recall that the dual cone D^* is defined by $D^* = \{d^* \in H : \langle d^*, d\rangle \geqq 0 \forall d \in D\}$. D^* is a closed convex cone and $(D^*)^* = D$.

LEMMA 2.2. *For any $p \in H$, $\bar{p} \in D$, the equality $\bar{p} = p^D$ holds iff*

$$\text{(i)} \quad \bar{p} - p \in D^*, \tag{2.2i}$$

$$\text{(ii)} \quad \langle \bar{p}, \bar{p} - p\rangle = 0. \tag{2.2ii}$$

Proof. Let $\bar{p}$ satisfy (i) and (ii). Then $\|d-p\|^2=\|d-\bar{p}\|^2+2\langle d,\bar{p}-p\rangle+\|\bar{p}-p\|^2\geqq\|\bar{p}-p\|^2$ for any $d\in D$. Since p^D is determined uniquely, $\bar{p}=p^D$. Conversely, if $\bar{p}=p^D$ and not (i), then there exists $d\in D$ such that $\langle p^D-p,d\rangle<0$ and, for some $\varepsilon>0$, also $\varepsilon\langle p^D-p,d\rangle+\varepsilon^2\|d\|^2<0$. Hence $\|p^D+\varepsilon d-p\|^2<\|p^D-p\|^2$; since D is a convex cone, $p^D+\varepsilon d\in D$ and (2.1) cannot be satisfied. If not (ii), then $\langle p^D,p^D-p\rangle>0$ in virtue of (i). There is an $\varepsilon_1>0$ such that inequality $-\varepsilon\langle p^D,p^D-p\rangle+\varepsilon^2\|p^D\|^2<0$ holds for all $\varepsilon\in(0,\varepsilon_1)$. Hence $\|(1-\varepsilon)p^D-p\|^2<\|p^D-p\|^2$; since D is a cone, $(1-\varepsilon)p^D\in D$ for $\varepsilon\in(0,1)$ and (2.1) cannot again be satisfied.

LEMMA 2.3. *For any $p\in H$ the following holds*:

(2.3i) (i) $p^{D^*}=p+(-p)^D$,

(2.3ii) (ii) $(-p)^D=-p^{-D}$,

(2.3iii) (iii) $(\lambda p)^D=\lambda p^D\quad\forall\lambda\geqq 0$.

(iv) *For any $\bar{p}\in D$ the equality $\bar{p}=p^D$ holds iff*

$$\|\bar{p}\|=\min_{d\in D^*+p}\|d\|. \tag{2.3iv}$$

Proof. (i) We have $(p+(-p)^D)-p=(-p)^D\in D=(D^*)^*$ and $\langle p+(-p)^D,(p+(-p)^D)-p\rangle=\langle(-p)^D-(-p),(-p)\rangle=0$ by (2.2ii). Hence $\bar{p}=p+(-p)^D$ satisfies conditions (i), (ii) of Lemma (2.2) with D changed to D^*. Part (ii) is proven similarly. To prove (iii) observe that $\|\lambda p^D-\lambda p\|=\lambda\|p^D-p\|=\lambda\min_{d\in D}\|d-p\|=\min_{\bar{d}\in\lambda D=D}\|\bar{d}-\lambda p\|$ for $\lambda>0$ since then $\lambda D=D$. If $\lambda=0$, (iii) is obvious. Part (iv) follows from (i), since $\min_{d\in D^*+p}\|d\|=\min_{\bar{d}\in D^*}\|\bar{d}+p\|=\|(-p)^{D^*}+p\|=\|p^D\|$.

The following statement, announced first by J. Moreau [18] in 1962 in a slightly different formulation, is a generalization of the classical decomposition theorem for Hilbert space [5], [28]: if $D=T$ is a closed subspace of H, then each $p\in H$ can be uniquely represented by $p=p^T+p^{T^\perp}$ where p^T, $p^{T^\perp}$ are projections on the orthogonal subspaces T, $T^\perp$ respectively. Note that any subspace is a cone and $T^\perp=T^*=-T^*$.

THEOREM 2.4. Decomposition theorem. *Any element $p\in H$ can be represented in the form*

$$p=p^D+p^{-D^*} \tag{2.4i}$$

with

$$\langle p^D,p^{-D^*}\rangle=0;\qquad\|p\|^2=\|p^D\|^2+\|p^{-D^*}\|^2. \tag{2.4ii}$$

This decomposition is unique: the relations $p=p_1+p_2$, $p_1\in D$, $p_2\in-D^*$, $\langle p_1,p_2\rangle=0$ imply $p_1=p^D$, $p_2=p^{-D^*}$. This decomposition is also norm-minimal: the relations $p=p_1+p_2$, $p_1\in D$, $p_2\in-D^*$ imply $\|p_1\|\geqq\|p^D\|$, $\|p_2\|\geqq\|p^{-D^*}\|$.

Proof. The theorem follows directly from Lemmas 2.2 and 2.3. The minimality property holds by (2.3iv): $p_1=p-p_2\in D^*+p$; hence $\|p^D\|\leqq\|p_1\|$.

The decomposition theorem is fundamental for determining constraint violation in the problem (1.12) and thus defining a penalty functional. But the projection on a cone has further useful properties.

LEMMA 2.5. *The projection on a convex closed cone has the following properties*:

(2.5i) (i) $\|p^D\| \leqq \|p\|, \qquad p \in H,$

(2.5ii) (ii) $\|p_1^D - p_2^D\| \leqq \|p_1 - p_2\|, \qquad p_1, p_2 \in H,$

(2.5iii) (iii) $\|(p_1 + p_2 - d^*)^D\| \leqq \|p_D^1 + p_D^2\|, \qquad p_1, p_2 \in H, \quad d^* \in D^*.$

Proof. Part (i) follows from Theorem 2.4. Part (ii) follows from the inequalities:

$$\begin{aligned}\|p_1^D - p_2^D\|^2 &= \langle p_1^D - p_2^D, p_1 - p_2 - p_1^{-D^*} + p_2^{-D^*}\rangle \\ &= \langle p_1^D - p_2^D, p_1 - p_2\rangle + \langle p_1^D, p_2^{-D^*}\rangle - \langle p_2^D, p_1^{-D^*}\rangle \\ &\leqq \langle p_1^D - p_2^D, p_1 - p_2\rangle \leqq \|p_1^D - p_2^D\| \cdot \|p_1 - p_2\|;\end{aligned}$$

these estimations are based on Lemma 2.2. To prove (iii) we apply (2.3iv):

$$\|(p_1 + p_2 - d^*)^D\| = \min_{d \in D^* + p_1 + p_2 - d^*} \|d\| \leqq \|p_1^D + p_2^D\|,$$

since

$$p_1^D + p_2^D = (p_1^D - p_1 + p_2^D - p_2 + d^*) + p_1 + p_2 - d^* \in D^* + p_1 + p_2 - d^*$$

by (2.2i).

Corollary 2.6. *The functional* $\|(\cdot)^D\|$ *is convex.*

Proof. Any subadditive ((2.5iii) with $d^* = 0$)) and positively-homogeneous (2.3iii) functional is convex.

More important are the properties of the functional $\frac{1}{2}\|(\cdot)^D\|^2$. Anticipating the applications to penalty functionals, we shall state the following lemma in terms of D^* rather than D though the roles of both cones are fully symmetric.

LEMMA 2.7. *Let* $q(p) = \frac{1}{2}\|p^{D^*}\|^2$. *Then*

(i) $q(\lambda p_1 + (1-\lambda)p_2 - d) \leqq \lambda q(p_1) + (1-\lambda) q(p_2), \quad p_1, p_2 \in H, \quad d \in D,$ $\lambda \in (0, 1)$. *In particular, q is convex.*

(ii) *The functional q is Frechet-differentiable with the derivative*

(2.7ii) $$q_p(p) = p^{D^*}.$$

Proof. Part (i) follows from (2.5iii) with D replaced by D^*:

$$\begin{aligned}\|(\lambda p_1 + (1-\lambda)p_2 - d)^{D^*}\|^2 &\leqq \|(\lambda p_1)^{D^*} + ((1-\lambda)p_2)^{D^*}\|^2 = \|\lambda p_1^{D^*} + (1-\lambda)p_2^{D^*}\|^2 \\ &\leqq \lambda \|p_1^{D^*}\|^2 + (1-\lambda)\|p_2^{D^*}\|^2.\end{aligned}$$

The proof of (ii) is omitted; various proofs are given in [9], [26], [29].

An extensive treatment of the projection on convex sets is given in [29]. We close this section with examples.

Example 2.8. Let $H = R^n$, $D = \{p = (p^1, \cdots, p^n) \in R^n : p^i \geqq 0 \ \forall i\}$ (the positive orthant). Then $D^* = D$ and from Lemma 2.2 it follows that $p^D = (p_+^1, \cdots, p_+^n)$ where $p_+^i = \max(0, p^i)$, $p^i \in R$.

Example 2.9. Let $H = W_1^2[0, 3]$. This is the space of absolutely continuous real functions on $[0, 3]$ with square integrable derivatives and with the scalar product

(2.9) $$\langle p_1, p_2\rangle = p_1(0)p_2(0) + \int_0^3 \dot{p}_1(t)\dot{p}_2(t)\, dt.$$

Let $D=\{p\in H: p(t)\geqq 0 \forall t\in[0,3]\}$. One can verify that: $D^*=\{p\in H: \dot{p}$ is nonincreasing, $0\leqq\dot{p}(t)\leqq p(0)$ a.e.$\}$. Let $p(t)=-1$, $t\in[0,1]$, $p(t)=t-2$, $t\in[1,3]$. Then from Lemma 2.2 it follows that $p^{-D^*}(t)\equiv -1$ and $p^D(t)=0$, $t\in[0,1]$; $t-1$, $t\in[1,3]$.

Example 2.9 shows that the projection on a cone can in general have quite a complicated form. The projection is simple in the space of L^2 type.

Example 2.10. Let $\mathcal{H}$ be a separable Hilbert space, $\mathcal{D}\subset\mathcal{H}$ a closed convex cone and $(\Omega, \mathcal{M}, \mu)$ a measure space. Let $H=L^2(\Omega, \mathcal{M}, \mu; \mathcal{H})$; this is the space of equivalence classes of the Bochner square integrable function from Ω into $\mathcal{H}$, with the scalar product

$$\langle p_1, p_2\rangle=\int_\Omega (p_1(\omega), p_2(\omega))\mu(d\omega), \tag{2.10}$$

where $(\cdot,\cdot)$ is the scalar product in $\mathcal{H}$. Let D be the closed convex cone $D=\{p\in H: p(\omega)\in\mathcal{D}$ a.e.$\}$. Then $D^*=\{p\in H: p(\omega)\in\mathcal{D}^*$ a.e.$\}$ and $p^D(\omega)=(p(\omega))^{\mathcal{D}}$ a.e. If, in particular, $\mathcal{H}=R$, $\mathcal{D}=R_+$, then $p^D(\omega)=(p(\omega))_+$.

3. Penalty functionals for a general optimization problem. In this and the following sections, let E be a real Banach space, H a real Hilbert space, D a nonempty closed convex cone in H with vertex at zero, $P: E\to H$ an operator and $Q: E\to R$ a functional.

Problem 3.1.

$$\min_{y\in Y_p} Q(y);\ Y_p=\{y\in E: p-P(y)\in D\}. \tag{3.1}$$

This is a rather general optimization problem, which includes most of the problems of optimal control, nonlinear programming, etc. The assumption that H be a Hilbert space is not really restrictive, since it is the most natural setting for many infinite-dimensional optimization problems—for example, problems with constraints described by partial differential equations. On the other hand, the Hilbert space has a useful and strong mathematical structure; here, the most important feature of the space H is the notion of the projection on the cone D.

Observe that $p-P(y)\in D$ if and only if $(p-P(y))^{-D^*}=0$ (Theorem 2.4). Define the constraint violation functional

$$K(y)=\tfrac{1}{2}\|(P(y)-p)^{D^*}\|^2;\qquad K: E\to R_+. \tag{3.2}$$

Since $(p-P(y))^{-D^*}=-(P(y)-p)^{D^*}$ by (2.3ii), then the condition $p-P(y)\in D$ is equivalent to $K(y)=0$. Thus, the notion of the projection on a cone makes it possible to reformulate the general problem (3.1) with operator constraints into a simpler one with functional constraints.

Simplified problem 3.1′.

$$\min_{y\in A_0} Q(y);\ A_0=\{y\in E: K(y)\leqq 0\}. \tag{3.1′}$$

Moreover, the functional K preserves some properties of the operator P.

LEMMA 3.3. (i) *If P is D-convex, that is, $(1-\lambda)P(y_1)+\lambda P(y_2)-P((1-\lambda)y_1+\lambda y_2)\in D$ for all $\lambda\in[0,1]$, $y_1, y_2\in E$, then K is convex. If, in addition, P is continuous, then K is weakly lower semicontinuous.*

(ii) *If P is continuous, then K is also. If P is weakly continuous (in weak topologies of both E and H), then K is weakly lower semicontinuous.*

(iii) *If P is differentiable, then K is also and*

$$Ky(y) = P_y^*(y)(P(y)-p)^{D^*}. \tag{3.3iii}$$

Proof. (i) $P((1-\lambda)y_1+\lambda y_2)-p=(1-\lambda)(P(y_1)-p)+\lambda(P(y_2)-p)-d$ for some $d\in D$, if $\lambda\in[0,1]$, $y_1, y_2\in D$. Since $K(y)=q(P(y)-p)$ the convexity of K follows from (2.7i). Since q is continuous by (2.5ii), then K is continuous; being convex, K is weakly lower semicontinuous. Part (ii) is immediate and (iii) follows directly from (2.7ii).

Example 3.4. Let $H=R^n\times R^m$, $D=R^n_+\times\{0_m\}=\{p=(p^1,\cdots,p^n,p^{n+1},\cdots,p^{n+m})\in R^n\times R^m : p^i\geqq 0,\ 1\leqq i\leqq n,\ p^i=0,\ i\geqq n+1\}$. Then $D^*=R^n_+\times R^m$. Let $P_i: E\to R$, $i=1,\cdots,n+m$, $P=(p_1,\cdots,P_{n+m})$, and $p=(p',\cdots,p^{n+m})\in H$. Then

$$K(y)=\tfrac{1}{2}\sum_{i=1}^{n}(P_i(y)-p_i)_+^2+\tfrac{1}{2}\sum_{i=n+1}^{n+m}(P_i(y)-p_i)^2.$$

Infinite dimensional examples.

Example 3.5a. *Nonlinear operator with values in $L^2(0,1)$.* Let $H=L^2(0,1)$, $D=\{p\in L^2(0,1) : p(t)\geqq 0 \text{ a.e.}\}$. Let $P: E\to L^2(0,1) : p\in L^2(0,1)$. K is defined by

$$K(y)=\frac{1}{2}\int_0^1 (P(y)(t)-p(t))_+^2\,dt. \tag{3.5a'}$$

Examples 3.4, 3.5a are general and simple. The constraint violation functional (3.2) appears in various problems with complex structure. The examples below are described without details, which can be easily filled in by the reader.

Example 3.5b. *A controlled system with inequality constraints.* Suppose the constraints are:

$$\begin{aligned} &\dot{x}(t)=f(x(t),u(t),t) && \text{a.e. in } [t_0,t_1],\\ &x(t_0)=x_0, \qquad g(x(t_1))=0, &&\\ &h(x(t))\geqq 0 && \forall t\in[t_0,t_1]. \end{aligned}$$

Assume the customary hypothesis on f to guarantee for any $u(\cdot)\in L^\infty(t_0,t_1;R^r)$ the existence of a unique solution $x(u)(\cdot)$ to the initial value problem (the first two equations). Suppose $g: R^n\to R^m$, $h: R^n\to R^k$. Define $H=R^m\times L^2(t_0,t_1;R^k)$, $E=L^\infty(t_0,t_1;R^r)$. Thus $y=u(\cdot)$. Define $P: E\to H$ by

$$P(u)=(g(x(u)(t_1)),h(x(u)(\cdot))).$$

Define the cone

$$D=\{(\bar{g},\bar{h}(\cdot))\in H : \bar{g}=0,\ \bar{h}(t)\geqq 0 \text{ a.e. in } [t_0,t_1]\}.$$

Then the whole set of constraints can be written as

$$-P(u)\in D.$$

Functional K is here

$$K(y)=\frac{1}{2}\|g(x(y)(t_1))\|^2+\frac{1}{2}\int_{t_0}^{t/1}|h(x(y)(t))_+\|^2\,dt. \tag{3.5b'}$$

Example 3.5c. *A controlled system with delay.* For simplicity, consider the linear case:

$$\begin{aligned}
&\dot{x}(t)=A(t)x(t)+B(t)x(t-1)+C(t)u(t)z && \text{a.e. in } [t_0, t_1],\\
&x(t)=\varphi_0(t) && \forall t\in[t_0-1, t_0],\\
&x(t)=\varphi_1(t) && \forall t\in[t_1-1, t_1],
\end{aligned}$$

where A, B, C are measurable bounded matrices of suitable dimensions. Provided $u(\cdot)$ is square integrable and φ_0, e.g., continuous we have that $x(\cdot)|_{[t_0,t_1]}\in W_1^2(t_0, t_1; R^n)$—that is, x is absolutely continuous with an L^2 derivative. Similarly as above, the first two equations define $x(u)$ for each $u\in E=L^2(t_0, t_1; R^r)$; therefore the constraining operator P can be defined as

$$P: E\to H,\ P(u)=x(u)|_{[t_1-1,t_1]}-\varphi_1.$$

Generally, the Hilbert space H can be chosen in at least two ways: $H=L^2(t_1-1, t_1; R^n)$ or $H=W_1^2(t_1-1, t_1; R^n)$. The corresponding penalty terms $K(y)$ (setting $y=u$) would have the form

$$\frac{1}{2}\int_{t_1-1}^{t_1}|x(y)(t)-\varphi_1(t)|^2\,dt \tag{3.5c'}$$

or

$$\frac{1}{2}|x(y)(t_1)-\varphi_1(t_1)|^2+\frac{1}{2}\int_{t_1-1}^{t_1}|\dot{x}(y)(t)-\dot{\varphi}_1(t)|^2 dt. \tag{3.5c''}$$

For particular problems of this type, other spaces and norms could also be employed; for instance, if the state equations were

$$\begin{aligned}
\dot{x}_1(t)&=x_1(t)-x_1(t-1)+x_2(t),\\
\dot{x}_2(t)&=x_2(t)+x_2(t-1)+u(t),
\end{aligned}$$

one could readily use the constraint violation term of the form[1] (abbreviating $x(u)(t)$ to $x(t)$)

$$\begin{aligned}
&\frac{1}{2}|x(t_1)-\varphi_1(t_1)|^2+\frac{1}{2}|\dot{x}_1(t_1)-\dot{\varphi}_{11}(t_1)|^2\\
&\qquad+\frac{1}{2}\int_{t_1-1}^{t_1}\left(|\ddot{x}_1(t)-\ddot{\varphi}_{11}(t)|^2+|\dot{x}_2(t)-\dot{\varphi}_{12}(t)|^2\right)dt.
\end{aligned} \tag{3.5c'''}$$

Example 3.5d. *A problem described by partial differential equation* (the

[1] This is the square norm in the product Sobolev space $W_2^2(t_1-1, t_1)\times W_1^2(t_1-1, t_1)$.

model of a gas pipe-line system [27]).

$$\frac{\partial p_i}{\partial t} = -A_i \frac{\partial Q_i}{\partial x}, \qquad (t, x) \in \Omega = [0, T] \times [0, L_i], \quad i = 1, 2,$$
$$\frac{\partial p_i}{\partial x} = -B_i Q_i$$

(where p, Q are the gas pressure and flow). Initial and boundary conditions are

$$p_i(x, 0) = f_i(x), \qquad i = 1, 2,$$
$$p_1(0, t) = g(t),$$
$$Q_1(L_1, t) = u(t),$$
$$Q_2(0, t) = u(t),$$
$$Q_2(L_2, t) = h(t)$$

(u is the control). Additional constraints:

$$p_{i\min} \leqq p_i(x, t) \leqq p_{i\max}, \qquad i = 1, 2, \quad \forall x, t$$
$$F_{j\min} \leqq F_j(u(t), p_1(L_1, t), p_2(0, t)) \leqq F_{j\max}, \qquad j = 1, 2 \quad \forall t.$$

For any control $u \in E = C(0, T)$, the state equations along with boundary and initial conditions define the pressures $p_1(u), p_2(u)$ belonging to $L^2(\Omega)$. Denote briefly $F_j(u(t), p_1(L_1, t), p_2(0, t))$ by $F_j(u)$. Define the Hilbert space H to be

$$H = L^2(\Omega; R^4) \times L^2([0, T]; R^4)$$

and the operator $P: E \to H$ by:

$$P(u) = (p_1(u) - p_{1\max}, p_{1\min} - p_1(u), p_2(u) - p_{2\max}, p_{2\min} - p_2(u),$$
$$F_1(n) - F_{1\max}, F_{1\min} - F_1(u), F_2(u) - F_{2\max}, F_{2\min} - F_2(u)).$$

Define also the cone $D \subset H$ by

$$D = \{(\bar{p}, \bar{F}) \in H : \bar{p}(x, t) \geqq 0 \text{ a.e. in } \Omega, \bar{F}(t) \geqq 0 \text{ a.e. in } [0, T]\},$$

where the inequalities are taken in R^4. Then the set of constraints is equivalent to

$$-P(u) \in D$$

and consequently

$$K(u) = \frac{1}{2} \sum_{i=1}^{2} \int_{\Omega} ((p_i(u) - p_{i\max})_+^2 + (p_{i\min} - p_i(u))_+^2 \, dt\, dx$$
$$+ \frac{1}{2} \sum_{j=1}^{2} \int_0^T (F_j(u) - F_{j\max})_+^2 + (F_{j\min} - F_j(u))_+^2) \, dt. \tag{3.5d'}$$

Given the functional K it is routine to define the penalty functional for problem (3.1) by

$$\Phi(y, \zeta) = Q(y) + \zeta K(y) = Q(y) + \tfrac{1}{2}\zeta \| (P(y) - p)^{D^*} \|^2. \tag{3.6}$$

It is also possible to define a shifted penalty functional by substituting the constant

element p in (3.6) by a variable penalty shift $\nu \in H$:

$$\Psi(y, \zeta, \nu) = Q(y) + \tfrac{1}{2}\zeta \|(P(y) - \nu)^{D^*}\|^2; \qquad y \in E, \zeta \geqq 0, \quad \nu \in H. \tag{3.7}$$

Clearly, $\Phi(y, \zeta) = \Psi(y, \zeta, p)$.

As was shown before, this form of penalization is well justified by numerous examples arising in computational experience. Functional (3.6) with (3.5a′) has already been suggested by Levitin–Poljak [13]. Functionals (3.6) and (3.7) with (3.5b′, c′, c″, c‴, d′) have been effectively used for solving optimization problems of (3.1) type. See [27], [38], [39] for computational results.

The penalty functionals (3.6), (3.7) can also be used for optimal control problems reformulated in a manner different from that shown in the examples above, where state equations along with initial and/or boundary conditions have been excluded from the set of constraints and the optimization has been carried out in the space of controls u. Another approach, proposed by Balakrishnan (ε-technique) and Rupp consists in carrying out the optimization in the space of pairs $y = (u(\cdot), x(\cdot)) = (\text{control, state})$, and treating state equations as principal constraint; other constraints, e.g., endpoint conditions on x can be included in the definition of the space E of optimized trajectories. See Balakrishnan [2] for the use of (3.6) and Rupp [41], [42] for the use of (3.7) for optimal control problems. Some computational results are given in [42], [43] and [27].

It is also possible to include all constraints in the functional K; in Example 3.5d one could augment functional (3.5d′) by the term:

$$\frac{1}{2}\sum_{i=1}^{2}\int_{\Omega}\left|\frac{\partial p_i}{\partial t} - \frac{A_i}{B_i}\frac{\partial^2 p_i}{\partial x^2}\right|^2 dt\,dx + \frac{1}{2}\sum_{i=1}^{2}\int_0^{L_i}|p_i(x,0) - f_i(x)|^2\,dx$$

$$+\int_0^T (|Q_1(L_1, t) - u(t)|^2 + |Q_2(0, t) - u(t)|^2 + |Q_2(L_2, t) - h(t)|^2)\,dt$$

in order to avoid solving numerically partial differential equations.

The properties of Φ and Ψ are related to the Lagrange multiplier theory in optimization techniques. Recall the following.

DEFINITION 3.8. Let Q, P be both differentiable or Q be convex, P be D-convex. The functional $L : H \times E \to R$ defined by

$$L(\eta, y) = Q(y) + \langle \eta, P(y) - p \rangle \tag{3.8i}$$

is called the *normal Lagrange functional* for the problem (3.1). An element $\eta \in H$ is called a *normal Lagrange multiplier* for the problem (3.1) at a (optimal) point $\hat{y} \in Y_p$ if

$$\eta \in D^*, \qquad \langle \eta, P(\hat{y}) - p \rangle = 0 \tag{3.8ii}$$

and

$$L_y(\eta, \hat{y}) = 0 \tag{3.8iii}$$

for Q, P differentiable, or

$$L(\eta, \hat{y}) \leqq L(\eta, y) \quad \forall y \in E \tag{3.8iv}$$

in the convex case.

It is well known that rather severe additional assumptions are needed in order to ensure the existence of normal Lagrange multipliers at an optimal point for the problem (3.1)—see, for example, [1], [11].

LEMMA 3.9.[2] *Suppose $\bar{y}$ minimizes $\Psi(\cdot, \zeta, \nu)$ over E. Then $\bar{y}$ is a solution of the problem*

$$\min_{y \in Y_{\bar{p}}} Q(y);\ Y_{\bar{p}} = \{y \in E : \bar{p} - P(y) \in D\}, \tag{3.9i}$$

where

$$\bar{p} = (P(y) - \nu)^{D^*} + \nu \tag{3.9ii}$$

is a normal Lagrange multiplier for the problem (3.9i) *at $\bar{y} = Y_{\bar{p}}$.*

Proof. $\bar{p} - P(\bar{y}) = (P(\bar{y}) - \nu)^{D^*} - (P(\bar{y}) - \nu)$ by (2.3i); hence $\bar{y} \in Y_{\bar{p}}$. Moreover, for $y \in Y_{\bar{p}}$, $\bar{p} - P(y) \in D$ and

$$\begin{aligned}\|(P(y) - \nu)^{D^*}\|^2 &= \|(\nu - P(y))^D - (\nu - P(y))\|^2 \leqq \|\bar{p} - P(y) - (\nu - P(y)\|^2 \\ &= \|\bar{p} - \nu\|^2 = \|(P(\bar{y}) - \nu)^{D^*}\|^2\end{aligned}$$

by (2.3i) and the definition of projection. Since $\psi(\bar{y}, \zeta, \nu) \leqq \psi(y, \zeta, \nu)$ for all $y \in E$, then $Q(\bar{y}) \leqq Q(y)$ for $y \in Y_{\bar{p}}$. Thus $\bar{y}$ solves (3.9i). Clearly, $\bar{\eta} \in D^*$ and $\langle \bar{\eta}, P(\bar{y}) - \bar{p} \rangle = \zeta \langle -(P(\bar{y}) - \nu)^{D^*}, (\nu - P(\bar{y}))^D \rangle = \zeta \langle (\nu - P(\bar{y}))^{-D^*}, (\nu - P(\bar{y}))^D \rangle = 0$ by Theorem 2.4. If Q, P are differentiable, then $\psi_y(\bar{y}, \zeta, \nu) = 0$ and $L_y(\bar{\eta}, \bar{y}) = Q_y(\bar{y}) + P_y^*(\bar{y})\bar{\eta} = \psi_y(\bar{y}, \zeta, \nu) = 0$ by (3.3iii). Thus (3.8ii) and (3.8iii) are satisfied at $(\bar{\eta}, \bar{y})$. Now, let Q be convex, P be D-convex and suppose (3.8iv) is not satisfied, i.e., there exists $\tilde{y} \in E$ such that $Q(\tilde{y}) - Q(\bar{y}) + \langle \eta, P(\tilde{y}) - P(\bar{y}) \rangle = \alpha < 0$. Let $y_\varepsilon = \bar{y} + \varepsilon(\tilde{y} - \bar{y}) = (1 - \varepsilon)\bar{y} + \varepsilon\tilde{y}$ for $0 < \varepsilon < 1$. By convexity, $(1 - \varepsilon)Q(\bar{y}) + \varepsilon Q(\tilde{y}) \geqq Q(y_\varepsilon)$ and $P(y_\varepsilon) = (1 - \varepsilon)P(\bar{y}) + \varepsilon P(\tilde{y}) - d(\varepsilon)$, $d(\varepsilon) \in D$. Hence Lemma 2.7ii implies the following estimate:

$$\begin{aligned}\Psi(y_\varepsilon, \zeta, \nu) - \Psi(\bar{y}, \zeta, \nu) &= Q(y_\varepsilon) - Q(\bar{y}) + \tfrac{1}{2}\zeta(\|(P(y_\varepsilon) - \nu)^{D^*}\|^2 - \|(P(\bar{y}) - \nu)^{D^*}\|^2 \\ &\leqq \varepsilon(Q(\tilde{y}) - Q(\bar{y})) + \tfrac{1}{2}\zeta(\|(P(\bar{y}) + \varepsilon(P(\tilde{y}) - P(\bar{y})) - \nu)^{D^*}\|^2 \\ &\quad - \|(P(\bar{y}) - \nu)^{D^*}\|^2) \\ &= \varepsilon(Q(\tilde{y}) - Q(\bar{y})) + \zeta\langle (P(\bar{y}) - \nu)^{D^*}, \varepsilon(P(\tilde{y}) - P(\bar{y}))\rangle + o(\varepsilon) \\ &= \varepsilon(Q(\tilde{y}) - Q(\bar{y})) + \langle \bar{\eta}, P(\tilde{y}) - P(\bar{y})\rangle + o(\varepsilon) = \varepsilon\left(\alpha + \frac{o(\varepsilon)}{(\varepsilon)}\right).\end{aligned}$$

Thus $\Psi(y_\varepsilon, \zeta, \nu) - \Psi(\bar{y}, \zeta, \nu) < 0$ for small ε and $\bar{y}$ cannot minimize $\Psi(\cdot, \zeta, \nu)$ if it does not minimize $L(\bar{\eta}, \cdot)$. This proves (3.8iv).

Without substantially changing the proof, the lemma can be restated for the case of a local or constrained minimum.

LEMMA 3.9′. *Suppose $\bar{y}$ minimizes $\Psi(\cdot, \zeta, \nu)$ over a set $A \subset E$. Then $\bar{y}$ is a solution of the problem*

$$\min_{y \in Y_{\bar{p}} \cap A} Q(y);\ Y_{\bar{p}} = \{y \in E : \bar{p} - P(y) \in D\}, \tag{3.9′i}$$

[2] This is the Everett theorem for the penalty functional (3.7).

where $\bar{p}$ is defined by (3.9ii); *the element $\bar{\eta}$ defined by* (3.9iii) *satisfies* (3.8ii). *If A is open and Q, P are differentiable, then $L_y(\bar{\eta}, \bar{y}) = 0$; hence $\bar{\eta}$ is a normal Lagrange multiplier. If A, Q are convex and P is D-convex, then $L(\bar{\eta}, \bar{y}) \leqq L(\bar{\eta}, y)$ for all $y \in A$; hence $\bar{\eta}$ is also a normal Lagrange multiplier for the problem* (3.9'i) *at $\bar{y} \in Y_{\bar{p}} \cap A$.*

The Lemmas 3.9, 3.9' are fundamental for understanding penalty functional techniques. First, it is assumed that $\Psi(\cdot, \zeta, \nu)$ does have a minimum; conditions for the existence of minimal points of penalty functionals are investigated in the next sections. Secondly, when minimizing a penalty functional, one actually solves not the original problem (3.1), but a slightly modified (3.9i); observe that $\bar{p} = (P(\bar{y}) - p)^{D^*} + p$ for unshifted penalty functionals. The modified problem is a normal one, i.e., it has normal Lagrange multipliers. The original problem need not be normal. If it is possible to choose a sequence $\{\zeta_n, \nu_n\}$ such that $\bar{p}_n$ converges to p, then the original problem is approximated by a sequence of normal ones. Since $\bar{p}_n - p = \nu_n - p + (P(\bar{y}_n) - \nu_n)^{D^*} = \nu_n - p + (1/\zeta_n)\bar{\eta}_n$ one can expect a fast convergence of $\bar{p}_n$ to p when choosing suitable shifts ν_n and keeping ζ_n constant. But $\bar{p}_n$ can be equal to p only if the original problem is normal. If it is not, ζ_n must be increased in order to approximate p by $\bar{p}_n$. The suitable algorithms and their covergence are discussed in §§6 and 7.

4. Augmented Lagrangians and duality theory. An augmented Lagrange functional can be defined by adding to the shifted penalty functional (3.7) a term independent from y; hence these two functionals are equivalent when minimized in y. But the study of augmented Lagrangians results in an extensive duality theory for nonconvex problems. See [22], [30], [34], [35], [36] for the discussion of this theory for nonconvex problems with $H = R^n$. In infinite dimensions, the convex case was studied—e.g., in [10], [12], [21]. The authors are not aware of any presentation of duality theory for nonconvex, infinite-dimensional problems. Nevertheless, the presentation here is brief and confined to main points which allow a generalization of the extensive theory presented in [22]. Those proofs which are obvious modifications of the proofs in R^n given in [22] are omitted in the sequel.

DEFINITION 4.1. The augmented Lagrange functional for the problem (3.1) is defined by introducing an equivalence between the Lagrange multiplier η and the penalty shift ν and coefficient ζ

$$\eta = \zeta(p - \nu). \tag{4.1i}$$

Then the augmented Lagrangian is

$$\begin{aligned} \Lambda(\zeta, \nu, y) &= \bar{\Lambda}(\zeta, \eta, y) = \Psi(y, \zeta, \nu) - \tfrac{1}{2}\zeta\|p - \nu\|^2 \\ &= Q(y) + \tfrac{1}{2}\zeta\|(P(y) - \nu)^{D^*}\|^2 - \tfrac{1}{2}\zeta\|p - \nu\|^2. \end{aligned} \tag{4.1ii}$$

In the sequel, only the functional $\Lambda(\zeta, \nu, y)$ will be studied.

LEMMA 4.2. *The optimization problem* (3.1) *is equivalent to the primal problem*

$$\text{(P)} \quad \min_{y \in E} \Big(\sup_{(\zeta, \nu) \in R_+ \times H} \Lambda(\zeta, \nu, y) \Big) = \min_{y \in E} \bar{Q}(y, p), \tag{4.2i}$$

where

$$\bar{Q}(y,p) \overset{\text{def}}{=} \begin{cases} Q(y), & y \in Y_p, \\ +\infty, & y \notin Y_p. \end{cases} \tag{4.2ii}$$

Proof. It is sufficient to show that

$$\sup_{(\zeta,\nu)} (\zeta \|(P(y)-\nu)^{D^*}\|^2 - \zeta\|p-\nu\|^2) = \begin{cases} 0, & y \in Y_p, \\ +\infty, & y \notin Y_p. \end{cases} \tag{4.2iii}$$

To prove this, note that for $p - P(y) \in D$ the following inequality holds:

$$\begin{aligned} \|(P(y)-\nu)^{D^*}\| - \|p-\nu\| &= \|(P(y)-p+p-\nu)^{D^*}\| - \|p-\nu\| \\ &\leqq \|(p-\nu)^{D^*}\| - \|p-\nu\| \leqq 0 \end{aligned}$$

due to (2.5iii) with D changed to D^*. If $p - P(y) \notin D$, then $(P(y)-p)^{D^*} \neq 0$; take $(\zeta_n, \nu_n) = (n, p)$ to obtain

$$\zeta_n(\|(P(y)-\nu_n)^{D^*}\|^2 - \|p-\nu_n\|^2) = n\|(P(y)-p)^{D^*}\|^2 \xrightarrow[n\to\infty]{} \infty.$$

DEFINITION 4.3. The functional

$$\hat{\Lambda}(\zeta,\nu) \overset{\text{def}}{=} \inf_{y\in E} \Lambda(\zeta,\nu,y) = \inf_{y\in E} \psi(y,\zeta,\nu) - \tfrac{1}{2}\zeta\|p-\nu\|^2 \tag{4.3i}$$

is called the *dual functional.* The *dual problem* is defined by

$$\text{(D)} \quad \max_{(\zeta,\nu)\in R_+\times H} \left(\inf_{y\in E} \Lambda(\zeta,\nu,y) \right) = \max_{(\zeta,\nu)\in R_+\times H} \hat{\Lambda}(\zeta,\nu). \tag{4.3ii}$$

Observe that the optimization in both (4.2i) and (4.3ii) in respect to ν (or η, see (4.1i)) is unconstrained in the space H, whereas in the classical convex duality theory the optimization is performed in respect to $\eta \in D^*$.

DEFINITION 4.4. Consider a family of optimization problems (3.1) with the parameter p varying over H. The functional

$$\hat{Q}(p) \overset{\text{def}}{=} \inf_{y\in Y_p} Q(y) \tag{4.4i}$$

is called the *primal functional.* Clearly,

$$\hat{Q}(p) = \inf_{y\in E} \bar{Q}(y,p). \tag{4.4ii}$$

A crucial role in the generalization of Rockafellar's duality theory is played by the following representation lemma.

LEMMA 4.5. *The functionals* $\Lambda(\zeta,\nu,y)$ *and* $\hat{\Lambda}(\zeta,\nu)$ *satisfy the relations*

$$\text{(i)} \quad \Lambda(\zeta, p+\bar{\nu}, y) = \inf_{\bar{p}\in H} \left(\bar{Q}(y, p+\bar{p}) + \frac{\zeta}{2}\|\bar{p}\|^2 - \zeta\langle \bar{p}, \bar{\nu}\rangle \right), \tag{4.5i}$$

$$\text{(ii)} \quad \hat{\Lambda}(\zeta, p+\bar{\nu}) = \inf_{\bar{p}\in H} \left(\hat{Q}(p+\bar{p}) + \frac{\zeta}{2}\|\bar{p}\|^2 - \zeta\langle \bar{p}, \bar{\nu}\rangle \right). \tag{4.5ii}$$

(iii) *The functionals* $(\zeta,\eta) \mapsto \Lambda(\zeta, p+(1/\zeta)\eta, y)$ *and* $(\zeta,\eta) \mapsto$

$\hat{\Lambda}(\zeta, p+(1/\zeta)\eta)$ *are concave and weakly upper semicontinuous.*

(iv) *For* $\zeta > \sigma \geqq 0$, $\bar{\nu} \in H$,

$$\hat{\Lambda}(\zeta, p+\bar{\nu}) \geqq \max_{\bar{z}\in H} \left(\hat{\Lambda}(\sigma, p+\bar{z}) - \frac{\|\zeta\bar{\nu} - \sigma\bar{z}\|^2}{2(\zeta-\sigma)} \right). \tag{4.5iv}$$

Proof. Without loss of generality, let $p=0$ to simplify notation. By (2.3iv), $\|(P(y)-\bar{\nu})^{D^*}\| = \min_{\bar{p}\in D+P(y)} \|\bar{p}-\bar{\nu}\|$. Therefore

$$\Lambda(\zeta, \bar{\nu}, y) = Q(y) + \tfrac{1}{2}\zeta \min_{\bar{p}\in D+P(y)} \|\bar{p}-\bar{\nu}\|^2 - \tfrac{1}{2}\zeta\|\bar{\nu}\|^2$$

$$= \min_{p\in D+P(y)} (Q(y) + \tfrac{1}{2}\zeta\|\bar{p}| - \zeta\langle\bar{p}, \bar{\nu}\rangle) = \inf_{\bar{p}\in H} (\bar{Q}(y, p) + \tfrac{1}{2}\zeta\|\bar{p}\|^2 - \zeta\langle\bar{p}, \bar{\nu}\rangle).$$

Thus (i) holds; the point (ii) follows from (i). Part (iii) holds since both functions are biggest minorants of a family of affine functions; see [12]. To prove (iv), observe that

$$\hat{\Lambda}(\zeta; \bar{\nu}) = \inf_{\bar{p}\in H} (\hat{Q}(p) + \tfrac{1}{2}\sigma\|\bar{p}\|^2 - \sigma\langle\bar{p}, \bar{z}\rangle + \tfrac{1}{2}(\zeta-\sigma)\|\bar{p}\| + \langle\bar{p}, \sigma\bar{z} - \zeta\bar{\nu}\rangle)$$

$$\geqq \inf_{\bar{p}\in H} (\hat{Q}(p) + \tfrac{1}{2}\sigma\|\bar{p}\|^2 - \sigma\langle\bar{p}, \bar{z}\rangle) + \min_{\bar{p}\in H} (\tfrac{1}{2}(\zeta-\sigma)\|\bar{p}\|^2 + \langle\bar{p}, \sigma\bar{z} - \zeta\bar{\nu}\rangle)$$

$$= \hat{\Lambda}(\sigma, z) - \frac{\|\zeta\bar{\nu} - \sigma\bar{z}\|^2}{2(\zeta-\sigma)}.$$

COROLLARY 4.6. For *any* $\eta \in H$,

$$\lim_{\zeta\to\infty} \hat{\Lambda}\left(\zeta, p+\frac{\eta}{\zeta}\right) = \sup_{(\sigma,\bar{z})\in R_+ h} \hat{\Lambda}(\sigma, p+\bar{z}) = \sup (\mathrm{D}), \tag{4.6}$$

where sup (D) *denotes the supremum of the dual problem* (4.3ii).

Proof. For any $(\sigma, \bar{z}) \in R_+ \times H$ and any $\varepsilon > 0$ it is possible to choose ζ' such that

$$\hat{\Lambda}\left(\zeta, p+\frac{\eta}{\zeta}\right) \geqq \hat{\Lambda}(\sigma, p+\bar{z}) - \frac{\|\eta - \sigma\bar{z}\|^2}{2(\zeta-\sigma)}$$

$$\geqq \hat{\Lambda}(\sigma, p+\bar{z}) - \varepsilon, \qquad \zeta \geqq \zeta'.$$

Lemma 4.5 allows a straightforward generalization of several duality theorems given in [22]. To state these theorems, some further definitions are necessary.

DEFINITION 4.7. The primal functional (4.4i) for the problem (3.1) is called *quadratically bounded* or, equivalently, it is said that the problem (3.1) satisfies the quadratic growth condition if there exist $q, \zeta \in R$ such that

$$\hat{Q}(p+\bar{p}) \geqq q - \zeta\|\bar{p}\|^2 \quad \forall \bar{p} \in H. \tag{4.7}$$

THEOREM 4.8 [22]. *If the primal functional* (4.4i) *is quadratically bounded, then*

$$-\infty < \sup\,(\mathrm{D}) = \liminf_{\bar{p}\to 0} \hat{Q}(p+\bar{p}) \geqq \hat{Q}(p) = \inf\,(\mathrm{P}), \tag{4.8}$$

where sup (D) *and* inf (P) *denote the supremum of* (4.3ii) *and the infimum of* (4.2i), *respectively. If the primal functional is not quadratically bounded, then* sup (D) = $-\infty$.

The next definitions are related to the so-called stability of the problem (3.1) in the family of perturbed problems defining the primal functional $\hat{Q}(p+\bar{p})$. Actually, stability is a kind of continuity of the primal functional. The notion of stability was introduced in [21]; see also [10], [12], [6].

DEFINITION 4.9. The problem (3.1) is called *inf-stable* if the primal functional is lower semicontinuous at p, that is,

$$\liminf_{\bar{p}\to 0} \hat{Q}(p+\bar{p}) \geqq \hat{Q}(p). \tag{4.9}$$

DEFINITION 4.10. The problem (3.1) is called stable of degree 2, if there is a neighborhood $\mathcal{O}$ of zero, an element $\bar{\nu} \in H$ and a number $\zeta > 0$ such that

$$\hat{Q}(p+\bar{p}) \geqq \hat{Q}(p) + \bar{\zeta}\langle \bar{p}, \bar{\nu}\rangle - \tfrac{1}{2}\bar{\zeta}\|\bar{p}\|^2 \quad \forall \bar{p} \in \mathcal{O}. \tag{4.10}$$

Conditions guaranteeing stability shall be discussed in Part II, § 8 of this paper. The notion of stability is the basis for two following theorems. The theorems are stated and proven for the nonconvex finite-dimensional case in [22]; the first theorem is also stated and proven for the general convex case in [21]. Due to the Lemma 4.5, the proofs of the theorems remain valid also for the nonconvex infinite-dimensional case.

THEOREM 4.11. *Suppose the primal functional* (4.4i) *for the optimization problem* (3.1) *is quadratically bounded. The duality relation* inf (P) = sup (D), *that is,*

$$\inf_{y\in E}\sup_{(\zeta,\nu)\in R_+\times H} \Lambda(\zeta, \nu, y) = \sup_{(\zeta,\nu)\in R_+\times H}\inf_{y\in E} \Lambda(\zeta, \nu, y) \tag{4.11}$$

holds if and only if the problem (3.1) *is inf-stable.*

THEOREM 4.12.[3] *Suppose the primal functional* (4.4i) *for the optimization problem* (3.1) *is quadratically bounded. The duality relation* inf (P) = max (D), *that is,*

$$\inf_{y\in E}\sup_{(\zeta,\nu)\in R_+\times H} \Lambda(\zeta, \nu, y) = \max_{(\pi,\nu)\in R_+\times H}\inf_{y\in E} \Lambda(\zeta, \nu, y) \tag{4.12}$$

holds if and only if the problem (3.1) *is stable of degree* 2. *Moreover, a pair* $(\bar{\zeta}, \bar{\nu}) \in R_+ \times H$ *is an optimal solution to the dual problem* (4.3ii) *iff it satisfies* (4.10) *for some neighborhood* $\mathcal{O}$ *of zero. If* Q *and* P *are differentiable or* Q *is convex,* P *is* D*-convex, then* $\eta = -\bar{\zeta}\bar{\nu}$ *is a normal Lagrange multiplier for the problem* (3.1).

[3] Compare also [34], [35], [36].

COROLLARY 4.13. *Assume that $\hat{y}$ is a solution of the problem* (3.1) *and let the problem satisfy the quadratic growth condition. A necessary and sufficient condition for the existence of $(\hat{\zeta}, \hat{v}) \in R^+ \times H$ such that $\hat{y}$ minimizes the augmented Lagrangian $\Lambda(\hat{\zeta}, \hat{v}, \cdot)$ or, equivalently, the shifted penalty functional $\Psi(\cdot, \bar{\zeta}, \hat{v})$ is that the problem* (3.1) *be stable of degree* 2. *The set of all these pairs $(\hat{\zeta}, \hat{v})$ is identical with the set of all pairs $(\bar{\zeta}, \bar{v})$ satisfying* (4.10) *for some neighborhood $\mathcal{O}$ of zero.*

Part II: Algorithms and convergence.

5. Minimizing sequences in constrained optimization. Consider the original problem (3.1) in its equivalent functional-constrained form.

Problem 5.1.

$$\min_{y \in A_0} Q(y); \; A_0 = \{y \in E : K(y) \leqq 0\}, \tag{5.1i}$$

where

$$K(y) = \tfrac{1}{2} \| (P(y) - p)^{D^*} \|^2 \tag{5.1ii}$$

is the constraint violation functional for the operator constraint $p - P(y) \in D$ corresponding to the squared distance from $p - P(y)$ to the positive cone D. Recall that E is a Banach space, H is a Hilbert space; $Q, K : E \to R$, $P : E \to H$, D is a closed convex cone in H, D^* is the dual cone, $(\cdot)^{D^*}$ is the projection on D^*.

In a numerical method solving the problem (5.1i), a sequence of points $\{y_n\}_{n=1}^{\infty}$ is generated, with the aim to approximate a solution $\hat{y}$ of the problem.

DEFINITION 5.2 (Rockafellar [21]). A sequence $\{y_n\}_{n=1}^{\infty} \subset E$ is called an *asymptotically minimizing sequence* (ASMS) iff

$$\text{(i)} \quad \lim_{n \to \infty} Q(y_n) = \liminf_{\bar{p} \to 0} \hat{Q}(p + \bar{p}), \tag{5.2i}$$

$$\text{(ii)} \quad \lim_{n \to \infty} K(y_n) = 0, \tag{5.2ii}$$

where $\hat{Q}$ is the primal functional (4.4i).

DEFINITION 5.3. A sequence $\{y_n\}_{n=1}^{\infty} \subset E$ is called a *weakly approximately minimizing sequence* (WAMS) iff

$$\text{(i)} \quad \lim_{n \to \infty} Q(y_n) = \hat{Q}(p) = \inf \text{(P)}, \tag{5.3i}$$

$$\text{(ii)} \quad \lim_{n \to \infty} K(y_n) = 0, \tag{5.3ii}$$

where inf (P) is the infimum of the primal problem (4.2i).

According to Theorem 4.8, if the primal functional is quadratically bounded (4.7), then $\hat{Q}(p) \geqq \liminf_{\bar{p} \to 0} \hat{Q}(p + \bar{p}) = \sup$ (D), where sup (D) denotes the supremum of the dual problem (4.3ii). In this case, a WAMS gives a better approximation of the solution of the original problem than an ASMS. However, if the problem is inf-stable (4.9), ASMS and WAMS are equivalent.

DEFINITION 5.4 (Levitin–Poliak [13]). A sequence $\{y_n\}_{n=1}^{\infty} \subset E$ is called

approximately minimizing (AMS) iff

(5.4i) $$\text{(i)} \quad \lim_{n\to\infty} Q(y_n) = \hat{Q}(p),$$

(5.4ii) $$\text{(ii)} \quad \lim_{n\to\infty} \operatorname{dist}(y_n, A_0) = 0.$$

An AMS is the strongest type of approximating sequence. If Q, K are uniformly convex, each AMS is norm convergent to the (unique) solution y of (5.1i)—see [4], [13]. But it is usually easier to show that a sequence of points generated by an algorithm is ASMS or WAMS rather than AMS. Under certain regularity assumptions, it is possible to verify that a WAMS is AMS.

DEFINITION 5.5 (Levitin–Poliak [13]). The constraint functional K is called *correct* if, for any sequence $\{y_n\}_{n=1}^{\infty} \subset E$ (5.3ii) implies (5.4ii).

Several conditions of correctness of K are discussed in [13]. One of them is the following:

LEMMA 5.6 [13]. *Suppose $K(y) \geqq 0$ for all $y \in E$, K is Fréchet-differentiable in E, its derivative $K_y(y)$ satisfies $\|K_y(y)\|^2 \geqq \lambda K(y)$ for all $y \in E$ and for some $\lambda > 0$, and the mapping $k_y(\cdot)$ is Lipschitzian. Then K is correct.*

COROLLARY 5.7. *Let P be affine continuous and its linear part P_0 be a surjection. Then K defined by* (5.1ii) *is correct.*

Proof. By (3.3iii), $K_y(y) = P_0^*(P(y)-p)^{D^*}$ and $K_y(y)$ is Lipschitzian with Lipschitz constant equal to $\|P_0\|^2$ by (2.5ii). Moreover,

$$\|K_y(y)\|^2 = \|P_0^*(P(y)p)^{D^*}\|^2 \geqq \lambda_0 \|(P(y)-p)^{D^*}\|^2 = 2\lambda_0 K(y)$$

for some $\lambda_0 > 0$, since P_0^* is a normal injective operator; see [5; VI.6.2]. Hence all assumptions of Lemma 5.6 are satisfied.

We close this section with a theorem on convergence of penalty techniques, which is due to Rockafellar [22] in the case of $H = R^n$.

THEOREM 5.8. *Let a sequence $\{(\zeta_n, \bar{\nu}_n)\} \subset R_+ \times H$ be given, such that $\zeta_n \geqq \delta > 0$ for all n and*

(5.8i) $$\lim_{n\to\infty} \hat{\Lambda}\left(\zeta_n - \delta, p + \frac{\zeta_n}{\zeta_n - \delta}\bar{\nu}_n\right) = \sup(\mathrm{D}) < +\infty,$$

where $\hat{\Lambda}$ is defined by (4.3i) *and* sup (D) *is the supremum of* (4.3ii).

Suppose each $y_n \in E$ minimizes approximately $\Lambda(\zeta_n, p+\bar{\nu}_n, \cdot)$ (or, equivalently, $\Psi(\cdot, \zeta_n, p+\bar{\nu}_n)$ over E), that is,

(5.8ii) $$\Lambda(\zeta_n, p+\bar{\nu}_n, y_n) \leqq \hat{\Lambda}(\zeta_n, p+\bar{\nu}_n) + \alpha_n,$$

where $\alpha_n \to 0$. Then

(5.8iii) $$\text{(a)} \quad \|(P(y_n)-p)^{D^*}\| \xrightarrow[n\to\infty]{} 0; \qquad \|\bar{\nu}_n^D\| \xrightarrow[n\to\infty]{} 0$$

or, equivalently, dist $(p - P(y_n), D) \to 0$ *and* dist $(-\bar{\nu}_n, D^*) \to 0$.

(b) *If the sequence $\{\zeta_n \bar{\nu}_n\}$ is bounded, then $\{y_n\}$ is* ASMS *for the problem* (5.1i, ii).

(c) *If the sequence* $\{\zeta_n \bar{v}_n\}$ *is bounded and the problem* (5.1i, ii) *is inf-stable* (4.9), *then* $\{y_n\}$ *is* WAMS *for the problem.*

Proof. Without loss of generality, let $p = 0$ to simplify notation.

(a) Denote $\bar{p}_n = \bar{v}_n + (P(y_n) - \bar{v}_n)$. According to (2.3iv) and (4.5i) the following estimate holds:

$$\Lambda(\zeta_n, \bar{v}_n, y_n) = Q(y_n) + \tfrac{1}{2}\zeta_n \|\bar{p}_n\|^2 - \zeta_n \langle \bar{p}_n, \bar{v}_n \rangle$$

$$= Q(y_n) + \tfrac{1}{2}(\zeta_n - \delta)\|\bar{p}_n\|^2 - (\zeta_n - \delta)\left\langle \bar{p}_n, \frac{\zeta_n}{\zeta_n - \delta}\bar{v}_n \right\rangle + \tfrac{1}{2}\delta\|\bar{p}_n\|^2$$

$$\geqq \hat{\Lambda}\left(\zeta_n - \delta, \frac{\zeta_n}{\zeta_n - \delta}\bar{v}_n\right) + \tfrac{1}{2}\delta\|\bar{p}_n\|^2.$$

This and (5.8ii) imply

$$\sup(\mathrm{D}) - \hat{\Lambda}\left(\zeta_n - \delta, \frac{\zeta_n}{\zeta_n - \delta}\bar{v}_n\right) + \alpha_n \geqq \tfrac{1}{2}\delta\|\bar{p}_n\|^2;$$

hence $p_n \to 0$ by (5.8i). Moreover, $\|P(y_n)^{D^*}\| = \|(\bar{v}_n + P(y_n) - \bar{v}_n)^{D^*}\| = \|(\bar{v}_n + (P(y_n) - \bar{v}_n)^{-D} + (P(y_n) - \bar{v}_n)^{D^*})^{D^*}\| \leqq \|(\bar{v}_n + (P(y_n) - \bar{v}_n)^{D^*})^{D^*}\| = \|\bar{p}_n^{D^*}\|$ by Theorem 2.4 and Lemma 2.5(i), (iii) (all statements of § 2 are obviously valid after interchanging D and D^*). Similarly, $\|\bar{v}_n^D\| = \|(\bar{p}_n - (P(y_n) - \bar{v}_n)^{D^*})^D\| \leqq \|\bar{p}_n^D\| \leqq \|p_n\|$ by (2.5i, iii). Thus, $\|P(y_n)^{D^*}\| \leqq \|\bar{p}_n\| \to 0$, $\|\bar{v}_n^D\| \leqq \|\bar{p}_n\| \to 0$.

(b) $\{\zeta_n \bar{v}_n\}$ being bounded, $-\zeta_n \langle \bar{v}_n, \bar{p}_n \rangle$ converges to zero. Since $\Lambda(\zeta_n, \bar{v}_n, y_n) = Q(y_n) + \frac{1}{2}\zeta_n \|\bar{p}_k\| - \zeta_n \langle p_n, v_n \rangle$ we have

$$\limsup_{n\to\infty} \tfrac{1}{2}\zeta_n \|\bar{p}_n\|^2 \leqq \lim_{n\to\infty} \Lambda(\zeta_n, \bar{v}_n, y_n) - \liminf_{n\to\infty} Q(y_n)$$

$$\leqq \sup(\mathrm{D}) - \liminf_{n\to\infty} Q(y_n).$$

But $\liminf_{n\to\infty} Q(y_n) \geqq \liminf_{n\to\infty} \hat{Q}(P(y_n)^{D^*}) \geqq \liminf_{\bar{p}\to 0} \hat{Q}(\bar{p})$. Hence $\limsup_{n\to\infty} \frac{1}{2}\zeta_n \|\bar{p}\|^2 \leqq \sup(\mathrm{D}) - \liminf_{\bar{p}\to 0} \hat{Q}(\bar{p}) = 0$ according to Theorem 3.8, which can be applied here since $\sup(\mathrm{D}) > -\infty$. Thus $\lim_{n\to\infty} \frac{1}{2}\zeta_n \|\bar{p}_n\|^2 = 0$ and $\lim_{n\to\infty} Q(y_n) = \lim_{n\to\infty} \Lambda(\zeta_n, \bar{v}_n, y_n) = \sup(\mathrm{D})$ so that $\{y_n\}$ is an ASMS.

(c) is an obvious consequence of (b).

Comment 5.9. If the assumptions (5.8i), (5.8ii) are satisfied, then—as it is shown in the proof above—$\lim_{n\to\infty} \Lambda(\zeta_n, p + \bar{v}_n\, y_n) = \sup(\mathrm{D})$; hence also $\lim_{n\to\infty} \hat{\Lambda}(\zeta_n, p + \bar{v}_n) = \sup(\mathrm{D})$. But $\lim_{n\to\infty} \hat{\Lambda}(\zeta_n, p + \bar{v}_n) = \sup(\mathrm{D})$ does not necessarily imply (5.8i). Thus the assumption (5.8i) is somewhat stronger than a typical dual approximation. But in some applications—for example, in the case of increased penalty algorithms—the assumption (5.8i) is easy to check and Theorem 5.8 implies quite powerful convergence results.

6. Increased penalty techniques. The problem (5.1) can be solved numerically by the following:

ALGORITHM 6.1 (pure increased). Given $\alpha > 0$, $\varepsilon > 0$, $\zeta_0 > 0$, $k > 1$ define

$$\zeta_n = k^n \zeta_0. \tag{6.1i}$$

For each ζ_n take $\bar{\nu}_n = p$ and minimize approximately the penalty functional

$$\Phi(y, \zeta_n) = \Psi(y, \zeta_n, p) = Q(y) + \tfrac{1}{2}\zeta_n \|(P(y) - p)^{D^*}\|^2 \tag{6.1ii}$$

in order to determine y_n such that

$$\Phi(y_n, \zeta_n) \leqq \inf_{y \in E} \Phi(y, \zeta_n) + \alpha_n, \tag{6.1iii}$$

where $\alpha_n \to 0$. If $\alpha_n \leqq \alpha$ and $\|(P(y_n) - p)^{D^*}\| \leqq \varepsilon$, stop.

It is known that the sequence $\{y_n\}$ generated by Algorithm 6.1 is a WAMS under mild assumptions concerning Q and K—see [4], [13]. The numerical effectiveness of the algorithm can be slightly improved by guessing a Lagrange multiplier $\eta \in D^*$ or penalty shift $\nu_0 = p - (1/\zeta)\eta = p_0 - \bar{\nu}_0$.

ALGORITHM 6.2 (increased-shifted). Given $\alpha > 0$, $\varepsilon > 0$, $\zeta_0 > 0$, $\bar{\nu}_0 \in -D^*$, $k > 1$ define

$$\zeta_n = k^n \zeta_0; \quad \nu_n = p + \bar{\nu}_n, \quad \bar{\nu}_n = k^{-n} \bar{\nu}_0. \tag{6.2i}$$

For each (ζ_n, ν_n) minimize approximately the penalty functional

$$\Psi(y_1, \zeta_n, \nu_n) = Q(y) + \tfrac{1}{2}\zeta_n \|(P(y) - \nu_n)^{D^*}\|^2. \tag{6.2ii}$$

in order to determine y_n such that

$$\Psi(y_n, \zeta_n, \nu_n) \leqq \inf_{y \in E} \Psi(y, \zeta_n, \nu_n) + \alpha_n, \tag{6.2iii}$$

where $\alpha_n \to 0$. If $\alpha_n \leqq \alpha$ and $\|(P(y_n) - \nu_n)^{D^*} + \nu_n - p\| \leqq \varepsilon$, then stop.

Observe that the stopping test in Algorithm 6.2 is slightly different than in Algorithm 6.1. This stopping test is based on Lemma 3.9: y_n solves (approximately) the problem $\min_{y \in Yp_n} Q(y)$ where $p_n = (P(y_n) - \nu_n)^{D^*} + \nu_n$; hence we require that $\|p_n - p\| \leqq \varepsilon$.

Theorem 5.8 implies the following lemma.

LEMMA 6.3. *Suppose that* $\sup(D) < +\infty$ *and a sequence* $\{y_n\}_{n=1}^{\infty}$ *generated by the Algorithms* 6.2 *or* 6.1 *is given. Then*:

(i) $\{y_n\}_{n=1}^{\infty}$ *is* ASMS *for the Problem* 5.1;

(ii) *if the problem is inf-stable, then it is* WAMS.

Proof. Assume $\{y_n\}_{n=1}^{\infty}$ is a sequence generated by Algorithm 6.2, Algorithm 6.1 being a special case of Algorithm 6.2 with $\bar{\nu}_0 = 0$. We have $\zeta_n \geqq \delta - \zeta_o > 0$ and

$$\hat{\Lambda}\left(\zeta_n - \delta, p + \frac{\zeta_n}{\zeta_n - \delta}\bar{\nu}_n\right) = \hat{\Lambda}\left(\zeta_n - \delta, p + \frac{1}{\zeta_n - \delta}\zeta_0 \bar{\nu}_0\right).$$

By Corollary 4.6, the assumption (5.8i) is satisfied; so are other assumptions of Theorem 5.8.

The assumptions of Lemma 6.3 are rather weak and the lemma indicates the strength of penalty functional techniques. However, the increased penalty technique becomes ill-conditioned numerically as $\zeta_n \to \infty$; hence, one should avoid increasing ζ_n if it is possible. This possibility is discussed in the next section. However, the increased penalty technique can also be used as a theoretical tool to derive necessary conditions of optimality. A. V. Balakrishnan [2] was first to

investigate this approach and to derive the maximum principle from his ε-technique. The theorem below is an abstract model of his reasoning.

THEOREM 6.4. *Suppose* $\sup(D) < +\infty$, *the Problem* 5.1 *is inf-stable and there exists* $\zeta \geqq 0$ *such that* $\Phi(\cdot, \zeta)$ *has a minimal point over* E *for each* $\zeta \geqq \bar{\zeta}$. *Let a sequence* $\zeta_n \geqq \bar{\zeta}$, $\zeta_n \to \infty$ *be given, and* $\{y_n\}_{n=1}^{\infty}$ *be a sequence satisfying* (6.1iii) *with* $\alpha_n = 0$. *Then*:

(i) $\{y_n\}$ *is* WAMS *for the Problem* 5.1.

(ii) *Denote*

$$p_n = (P(y_n) - p)^{D^*} + p. \tag{6.4i}$$

Then $p_n \to 0$ *and* y_n *solves the problem*

$$\min_{y \in Y_{p_n}} Q(y); \qquad Y_{p_n} = \{y \in E : p_n - P(y) \in D\}. \tag{6.4ii}$$

(iii) *Denote*

$$\eta_n = \zeta_n (P(y_n) - p)^{D^*}. \tag{6.4iii}$$

If Q, P *are differentiable or* Q *is convex,* P *is* D*-convex, then* η_n *is a normal Lagrange multiplier for the problem* (6.4ii) *at* y_n.

(iv) *If* Q, K *are* (*weakly*) *lower semicontinuous, then each* (*weak*) *accumulation point of* $\{y_n\}$ *is a solution to the Problem* 5.1.

(v) *Let* Q, P *be continuously differentiable or* Q, P *be continuous,* Q *be convex,* P *be* D*-convex. If the sequence* $\{y_n\}$ *converges in norm to a point* $\hat{y}$ *being thus a solution to Problem* 5.1, *then each weak accumulation point of the sequence* $\{\eta_n\}$ *is a normal Lagrange multiplier for the Problem* 5.1 *at* $\hat{y}$.

Proof. Points (i), (ii), (iii) follow directly from Lemmas 6.3, 3.9; points (iv) and (v) are obvious.

This theorem relates the apparently crude Algorithm 6.1 to rather delicate aspects of optimization theory. The original problem (not necessarily a normal one) is approximated by a sequence of normal problems.

COROLLARY 6.5. *Consider a family of problems* (5.1) *for various* $p \in \mathcal{P} = \{p \in H: Y_p \neq \emptyset\}$. *Denote by* $\mathcal{P}_1$ *the set of all* $p \in \mathcal{P}$ *such that the corresponding problems satisfy all the assumptions of Theorem* 6.4. *Denote by* $\mathcal{P}_1$ *the set of all* $p \in \mathcal{P}$ *such that the corresponding problems possess a normal Lagrange multiplier at some solution. Then* $\mathcal{P}_2$ *is dense in* $\mathcal{P}_1$.

Under moderate assumptions—see § 8—$\mathcal{P}_1 = \mathcal{P}$ and $\mathcal{P}_2$ is dense in $\mathcal{P}$. Thus, the existence of normal Lagrange multipliers is a metrically typical property. In other words, the normal problems are rich enough for computational purposes—just as rational numbers are rich enough for computations on the real axis.

7. Shifted penalty techniques. The goal of a shifted penalty technique is to approximate, if possible, the saddle point of the augmented Lagrangian $\Lambda(\zeta, \nu, y) = Q(y) + \frac{1}{2}\zeta \|(P(y) - \nu)^{D^*}\|^2 - \frac{1}{2}\zeta \|p - \nu\|^2$, without increasing the penalty coefficient ζ towards infinity. The shifted penalty techniques are usually more effective computationally than the increased ones.

A natural algorithm for shifted penalty techniques is the following:

ALGORITHM 7.1. (saddle-point seeking). Given $\alpha, \beta, \gamma > 0$, $\zeta_0 > 0$, $\nu_0 \in$

H, $y_0 \in E$ determine a sequence of (ζ_n, ν_n, y_n) by computing the gradient

$$\Lambda_\zeta(\zeta, \nu, y) = \tfrac{1}{2}(\|(P(y)-\nu)^{D^*}\|^2 - \|p-\nu\|^2), \tag{7.1i}$$

$$\Lambda_\nu(\zeta, \nu, y) = \zeta(p-\nu-(P(y)-\nu)^{D^*}), \tag{7.1ii}$$

$$\Lambda_y(\zeta, \nu, y) = Q_y(y) + \zeta P_y^*(y)(P(y)-\nu)^{D^*} \tag{7.1iii}$$

and by applying a saddle-point seeking gradient algorithm. If $|\Lambda_\zeta(\zeta_n, \nu_n, y_n)| \leqq \gamma$, $\|\Lambda_\nu(\zeta_n, \nu_n, y_n)\| \leqq \beta$, $\|\Lambda_y(\zeta_n, \nu_n, y_n)\| \leqq \alpha$, stop.

However, there are only a few saddle-point seeking algorithms known and they are not very reliable computationally. Most of them are based on the following:

ALGORITHM 7.2 (dual gradient). Given $\alpha, \beta, \gamma > 0$, $\zeta_0 > 0$, $\nu_0 \in H$, $y_0 \in E$ determine a sequence of (ζ_n, ν_n, y_n) by minimizing approximately $\Lambda(\zeta_n, \nu_n, \cdot)$ or, equivalently, $\Psi(\cdot, \zeta_n, \nu_n)$ over E—hence satisfying the condition (6.2iii)—and by choosing a step-size coefficient τ_n in the relations

$$\zeta_{n+1} = \zeta_n + \tfrac{1}{2}\tau_n(\|(P(y_n)-\nu_n)^{D^*}\|^2 - \|p-\nu_n\|^2), \tag{7.2i}$$

$$\nu_{n+1} = \nu_n + \tau_n\zeta_n(p-\nu_n-(P(y_n)-\nu_n)^{D^*}). \tag{7.2ii}$$

If $|\Lambda_\zeta(\zeta_n, \nu_n, y_n)| \leqq \gamma$, $\|\Lambda_\nu(\zeta_n, \nu_n, y_n)\| \leqq \beta$, $\alpha_n \leqq \alpha$, stop.

But the augmented Lagrangian $\Lambda(\zeta, \nu, y)$ has saddle points at many pairs of (ζ, ν) if the problem is stable of degree 2 and ζ is sufficiently large (see Theorem 4.12). In such cases, one may apply only the part (7.2ii) of the algorithm. The choice $\tau_n = 1/\zeta_n = 1/\zeta$ is particularly useful.

ALGORITHM 7.3 (pure shifted). Given a sufficiently large $\zeta > 0$ and $\beta > 0$, $\nu_0 = p$ determine a sequence $\{(y_n, \nu_n)\}$ by

$$y_n = \arg\inf_{y \in E} \Psi(y, \zeta, \nu_n), \tag{7.3i}$$

$$\nu_{n+1} = p - (P(y_n)-\nu_n)^{D^*}. \tag{7.3ii}$$

If $\|\nu_{n+1}-\nu_n\| \leqq \beta$, stop.

Here it is assumed for simplicity that Ψ is minimized precisely (7.3i). The stopping test results from the following consideration. Denote $p_n = \nu_n + (P(y_n)-\nu_n)^{D^*}$. Then each y_n minimizes $Q(y)$ over $Y_{p_n} - \{y \in E : p_n - P(y) \in D\}$—see Lemma 3.9. Denote $\tilde{p}_n = (P(y_{n-1})-\nu_{n-1})^{D^*}$; hence

$$\begin{aligned}\|\nu_{n+1}-\nu_n\| &= \|p-p_n\| = \|(P(y_{n-1})-\nu_{n-1})^{D^*} - (P(y_n)-\nu_n)^{D^*}\| \\ &= \|P(y_n)-p+(P(y_n)-p+\tilde{p}_n)^{-D}\| \geqq \|(P(y_n)-p)^{D^*}\|\end{aligned}$$

by Lemma 2.3. Thus if the stopping test is satisfied, then $\|(P(y_n)-p)^{D^*}\| \leqq \|p_n-p\| = \|\nu_{n+1}-\nu_n\| \leqq \beta$ whereby $y_n \in Y_{p_n}$.

Algorithm 7.3 is a generalization of the penalty shift algorithm given by Powell [20] in case of equality constraints in R^n and by Szymanowski and Wierzbicki [25] for inequality constraints. This algorithm can be further improved by introducing an automatic choice of the penalty coefficient ζ:

ALGORITHM 7.4. (shifted-increased). Given $\zeta_0 > 0$, $k > 1$, $\delta \in (0, 1)$, $c_0 > 0$

and $\beta > 0$, $\nu_0 = p$, determine a sequence $\{(y_n, \zeta_n, \nu_n)\}$ by

$$y_n = \arg\inf_{y \in E} \Psi(y, \zeta_n, \nu_n), \tag{7.4i}$$

$$p_n = \nu_n + (P(y_n) - \nu_n)^{D^*}. \tag{7.4ii}$$

If $\|p_n - p\| > c_n$, set

$$\nu_{n+1} = p - (P(y_n) - \nu_n)^{D^*}, \quad \zeta_{n+1} = \zeta_n, \quad c_{n+1} = \delta c_n. \tag{7.4iii}$$

If $\|p_n - p\| > c$, set

$$\nu_{n+1} = p - \frac{1}{k}(P(y_n) - \nu_n)^{D^*}, \quad \zeta_{n+1} = k\zeta_n, \quad c_{n+1} = c_n. \tag{7.4iv}$$

If $\|p_n - p\| \leqq \beta$ stop.

This algorithm is actually a combination of Algorithms 7.3 and 6.2. It is also one of the most powerful penalty algorithms, most effective for solving various static and dynamic optimization problems [27]. If the problem is only inf-stable, the algorithm is convergent by a modification of Lemma 6.3. If the problem has a higher order of stability, the penalty increase part (7.4iv) of the algorithm is applied only as many times as it is necessary to secure the convergence.

DEFINITION 7.5. Assume that Q, P are differentiable or Q is convex, P is D-convex. The Problem 5.1 will be called L-*stable* (locally in a nonempty open set $A \subset E$) if there exists a neighborhood $U_p \subset H$ of zero such that the problems $\min_{y \in Y_{p+\bar{p}} \cap A} Q(y)$, $Y_{p+\bar{p}} \cap A = \{y \in A : p + \bar{p} - P(y) \in D\}$ have solutions and unique normal Lagrange multipliers $\eta(p + \bar{p})$ for each $\bar{p} \in U_p$ and the mapping $\bar{p} \mapsto \eta(p + \bar{p})$ is Lipschitz continuous in U_p.

Similarly—see (4.10)—one can define local stability of degree 2 in A.

The following theorem was given first in [25] for $H = R^n$ and in [26] for the general case. The theorem is valid also when $A = E$ but in numerical implementations local minima are usually of interest; moreover, the conditions guaranteeing local L-stability are somewhat simpler.

THEOREM 7.6. *Suppose there exists* $\zeta' > 0$ *such that the functional* $\Psi(\cdot, \zeta, p + \bar{\nu})$ *attains its minimum over* A *for any* $\zeta \geqq \zeta'$ *and each* $\bar{\nu}$ *in a neighborhood* $U_\nu \subset H$ *of zero. If the Problem* 5.1 *is locally* L-*stable in* A, *then*:

(i) *there exists* $\zeta'' \geqq \zeta'$ *such that for any* $\zeta \geqq \zeta''$ *there is* $\bar{\nu}_\zeta \in U_\nu$ *such that any point minimizing* $\Psi(\cdot, \zeta, p + \bar{\nu}_\zeta)$ *is a local solution to Problem* 5.1 *in* A;

(ii) *the Problem* 5.1 *is locally stable of degree* 2 *in* A;

(iii) *if* $\zeta \geqq \zeta''$, *then Algorithm* 7.3 *has the following properties*:

$$\{\bar{\nu}_n\} \subset U_\nu, \qquad \bar{\nu}_n \xrightarrow[n \to \infty]{} \bar{\nu}_\zeta, \tag{7.6i}$$

$$\{\bar{p}_n\} \subset U_p, \quad \bar{p}_n \xrightarrow[n \to \infty]{} 0, \quad \bar{p}_n = \bar{\nu}_n + (P(y_n) - p - \bar{\nu}_n)^{D^*}, \tag{7.6ii}$$

$$\{y_n\} \text{ is a WAMS for Problem } 5.1 \text{ in } A. \tag{7.6iii}$$

(iv) *Given any* $\delta \in (0, 1)$ *denote by* R_η *the Lipschitz constant of the multiplier mapping* $\bar{p} \mapsto \eta(p + \bar{p})$; *then there exists* $\zeta_\delta = \max(\zeta'', 1 + \delta/\delta R_\eta)$ *such that* $\zeta \geqq \zeta_\delta$

implies that the Algorithms 7.4 *and* 7.3 *are equivalent and*

$$\|\bar{p}_{n+1}\| \leqq \delta \|\bar{p}_n\|, \tag{7.6iv}$$

$$\|\bar{\nu}_{n+1} - \bar{\nu}_\zeta\| \leqq \delta \|\bar{\nu}_n - \bar{\nu}_\zeta\|. \tag{7.6v}$$

The proof of the theorem shall be based on a contraction mapping and is different from the proof given in [26], but before proving the theorem, we need the following:

LEMMA 7.7. *Let* $y(\bar{\nu})$ *denote an arbitrary element minimizing* $\Psi(\cdot, \zeta, p+\bar{\nu})$ *over* A *for each* $\bar{\nu} \in U_\nu$. *Define the mapping* $T: U_\nu \to H$ *by* $T(\nu) = -(P(y(\bar{\nu})) - p - \bar{\nu})^{D^*}$. *Under the assumptions of Theorem* 7.6, *for each* $\delta \in (0, 1)$ *there exists a* ζ_δ *such that* $\zeta \geqq \zeta_\delta$ *implies*

(i) $T: B(\zeta^{-3/4}) \to \beta(\zeta^{-3/4})$ *where* $B(r)$ *is a closed ball of radius* r *with center at zero*;

(ii) $\|T(\nu') - T(\nu'')\| \leqq \delta \|\nu' - \nu''\|$ *for* $\nu', \nu'' \in B(\zeta^{-3/4})$.

Proof of the lemma. Without loss of generality, assume $p = 0$. Let $\bar{p}(\bar{\nu}) = \bar{\nu} - T(\bar{\nu})$. By Lemma 3.9, $y(\bar{\nu})$ solves the problem $\inf_{y \in A \cap Y_{\bar{p}(\bar{\nu})}} Q(y)$ and $-\zeta T(\bar{\nu})$ is a normal Lagrange multiplier for this problem at $y(\bar{\nu})$. Also, $\hat{\Lambda}(\zeta, \bar{\nu}) = \Lambda(\zeta, \bar{\nu}, \bar{y}(\bar{\nu}))$. Similarly as in the proof of Theorem 5.8, the following estimate holds:

$$\sup (\mathrm{D}) - \hat{\Lambda}\left(\zeta - 2, \frac{\zeta}{\zeta - 2}\bar{\nu}\right) \geqq \|\bar{p}(\bar{\nu})\|^2. \tag{7.7i}$$

Let $\varepsilon > 0$ be such that $B(\varepsilon^{1/2}) \subset U_p$. By Corollary 4.6, there is a $\zeta_\varepsilon \geqq 2$ such that

$$\hat{\Lambda}(\zeta_\varepsilon - 2, 0) \geqq \sup (D) - \frac{\varepsilon}{2}. \tag{7.7ii}$$

By (4.5iv),

$$\hat{\Lambda}\left(\zeta - 2, \frac{\zeta}{\zeta - 2}\bar{\nu}\right) \geqq \hat{\Lambda}(\zeta_\varepsilon - 2, 0) - \frac{\|\zeta\bar{\nu}\|^2}{2(\zeta - \zeta_\varepsilon)} \tag{7.7iii}$$

$$\geqq \sup (\mathrm{D}) - \frac{\varepsilon}{2} - \frac{\sqrt{\zeta}}{\zeta - \zeta_\varepsilon} \cdot \frac{1}{2}\|^{3/4}\bar{n}\|^2.$$

Take $\bar{\zeta}$ such that $\sqrt{\zeta}/(\zeta - \zeta_\varepsilon) \leqq \varepsilon$ for $\zeta \geqq \bar{\zeta}$. Since $\sup (\mathrm{D})$ is bounded by (7.7i), combine (7.7i) and (7.7iii) to obtain

$$\|\bar{p}(\bar{\nu})\|^2 \leqq \varepsilon \quad \text{for} \quad \zeta \geqq \bar{\zeta}, \quad \bar{\nu} \in B(\zeta^{-3/4}). \tag{7.7iv}$$

Hence, $\bar{p}(\bar{\nu}) \in B(\varepsilon^{1/2}) \subset U_p$. Therefore, by the L-stability assumption,

$$\zeta T(\bar{\nu}) = -\eta(\bar{p}(\bar{\nu})). \tag{7.7v}$$

Let $\bar{\nu}', \bar{\nu}'' \in B(\zeta^{-3/4})$ and let R_η denote the Lipschitz constant of the map $\bar{p} \mapsto \eta(\bar{p})$

$$\begin{aligned} \zeta\|T(\bar{\nu}') - T(\bar{\nu}'')\| &= \|\eta(\bar{p}(\bar{\nu}')) - \eta(\bar{p}(\bar{\nu}''))\| \\ &\leqq R_\eta \|\bar{p}(\bar{\nu}') - \bar{p}(\bar{\nu}'')\| \\ &\leqq R_\eta \|T(\bar{\nu}') - T(\bar{\nu}'')\| + R_\eta \|\bar{\nu}' - \bar{\nu}''\| \end{aligned} \tag{7.7vi}$$

and, for $\zeta \geq \tilde{\zeta} = \max(\bar{\zeta}, R_\eta)$:

$$\|T(\bar{\nu}') - T(\bar{\nu}'')\| \leq \frac{R_\eta}{\zeta - R_\eta} \|\nu' - \nu''\|. \tag{7.7vii}$$

The relations (7.7v), (7.7vii) imply

$$\left\| T(\bar{\nu}) + \frac{1}{\zeta}\eta(\bar{p}(0)) \right| \leq \frac{R_\eta}{\zeta - R_\eta}\|\bar{\nu}\| \leq \zeta^{-3/4}\frac{R_\eta}{\zeta - R_\eta}, \qquad \bar{\nu} \in B(\zeta^{-3/4}), \tag{7.7viii}$$

and

$$\begin{aligned} \|T(\bar{\nu})\| &\leq \frac{1}{\zeta}\|\eta(\bar{p}(0))\| + \zeta^{-3/4}\frac{R_2}{\zeta - R_\eta} \\ &= \zeta^{-3/4}\left(\frac{\|\eta(p(0)\|}{\zeta^{1/4}} + \frac{R_\eta}{\zeta - R_\eta}\right). \end{aligned} \tag{7.7ix}$$

Therefore there exists $\bar{\bar{\zeta}} \geq \tilde{\zeta}$ such that $\zeta \geq \bar{\bar{\zeta}}$ implies $\|T(\bar{\nu}\| \leq \zeta^{-3/4}$ if $\|\nu\| \leq \zeta^{-3/4}$ and the point (i) of the lemma is proved. Take $\zeta_\delta = \max(\bar{\bar{\zeta}}(1+\delta/\delta)R_\eta)$; then the point (ii) of the lemma follows from (7.7vii).

Proof of Theorem 7.6. Again, assume $p = 0$. (i) Choose an arbitrary $\delta \in (0, 1)$ and take $\zeta'' = \zeta_\delta$. By the contraction mapping theorem, there is a unique element $\bar{\nu}_\zeta \in B(\zeta^{-3/4}) \subset U_\nu$ such that $\bar{p}(\nu_\zeta) = \bar{\nu}_\zeta - T(\bar{\nu}_\zeta) = 0$. By Lemma 3.9′, $\bar{y}(\bar{\nu}_\zeta)$ is a local solution to Problem 5.1 in A. Moreover, $\zeta\bar{\nu}_\zeta = -\eta(0)$ is the Lagrange multiplier for the problem.

(ii) Since $\sup(D) > -\infty$—see the proof of Lemma 7.7—Problem 5.1 in A satisfies the quadratic growth condition (4.7). Local stability of degree 2 in A follows from (i) and Corollary 4.13.

(iii) Observe that Algorithm 7.3 is a fixed-point algorithm $\bar{\nu}_{n+1} = T(\bar{\nu}_n)$. Hence (7.6i), (7.6ii) follow from Lemma 7.7; Condition (7.6iii) is implied by (7.5ii), Lemma 3.9′ and the fact that an optimization problem is inf-stable if it is stable of degree 2. Part (iv) is a direct consequence of Lemma 7.7.

In the case of $H = R^n$ the convergence of this algorithm has been investigated by many authors, starting from [25] up to the most complete discussion in [31]. In [15] also other penalties different from the square norm were considered. In the convex case the assumptions can be essentially relaxed, as shown in [24]. For further references see the survey paper [32].

8. Conditions of stability and convergence conditions. Convergence conditions in the previous section were stated in terms of various stability assumptions. The aim of this section is to discuss the assumptions and to achieve more explicit convergence conditions.

The simplest assumption is the quadratic growth condition (4.7). It is obviously satisfied, if Q is bounded from below. By Theorem 4.8, it is equivalent to the condition $\sup(D) > -\infty$. The latter holds iff there is a pair $(\zeta, \nu) \in R_+ \times H$ such that $\hat{\Lambda}(\zeta, \nu) = \inf_{y \in E} \Lambda(\zeta, \nu, y) > -\infty$.

Now consider the notion of inf-stability (4.9) and the question of the existence of points minimizing $\Lambda(\zeta, \nu, \cdot)$ or, equivalently, $\Psi(\cdot, \zeta, \nu)$ over E (this assumption was used in Theorems 6.4, 7.6).

THEOREM 8.1. *Let E be reflexive, Q be weakly lower semicontinuous and P continuous in weak topologies of E and H.*[4] *Suppose there exist ε and $\delta_0 > 0$ and an element $\bar{p}_0 \in \operatorname{int} D$ such that the sets*

$$C_\delta = \{y \in E : Q(y) \leqq \varepsilon, p + \delta\bar{p}_0 - P(y) \in D\} \tag{8.1}$$

are bounded and nonempty for $0 \leqq \delta \leqq \delta_0$. Then:

(i) *Problem* 5.1 *is inf-stable.*

If, in addition, $Q(y) \geqq \beta > -\infty$ for all $y \in E$, then:

(ii) *there exists $\bar{\zeta} \geqq 0$ and a neighborhood U_ν of zero in H such that for every $\zeta \geqq \bar{\zeta}$ and each $\bar{\nu} \in U_\nu$ satisfying $\tilde{Y}_\nu = \{y \in E : Q(y) \leqq \varepsilon, p + \bar{\nu} - P(y) \in D\} \neq \varnothing$, there is a point $y_{\zeta\bar{\nu}}$ minimizing $\Psi(\cdot, \zeta, p + \bar{\nu})$ over E.*

Proof. Assume $p = 0$ without loss of generality. Denote $S_\delta = \{y \in E : \delta\bar{p}_0 - P(y) \in D\}$ and $\tilde{Q}(\delta) = \inf_{y \in S_\delta} Q(y)$. We prove first that $\tilde{Q}(\cdot)$ is right continuous at zero. Observe that $\tilde{Q}(\cdot)$ is nonincreasing for $\delta \geqq 0$ and $\tilde{Q}(\delta) \leqq \varepsilon$ for $\delta \in [0, \delta_0]$. Thus, if $\tilde{Q}(\bar{\delta}) = \varepsilon$ for some $\bar{\delta} > 0$, then $\tilde{Q}(\delta) = \varepsilon$ for $\delta \in [0, \bar{\delta}]$ and $\tilde{Q}(\cdot)$ is right continuous at zero. Assume therefore that $\tilde{Q}(\delta) < \varepsilon$ for all $\delta > 0$. By its monotonicity, it is sufficient to show that

$$\tilde{Q}(0) \leqq \liminf_{n \to \infty} \tilde{Q}(\delta_n) \tag{8.1iii}$$

for any sequence $\{\delta_n\} \subset (0, \delta_0]$ convergent to zero. For each n choose $y_n \in S_{\delta_n}$ satisfying

$$Q(y_n) \leqq \tilde{Q}(\delta_n) + \min\left(\frac{1}{n}, \varepsilon - \tilde{Q}(\delta_n)\right). \tag{8.1iv}$$

Then $Q(y_n) \leqq \varepsilon$, $y_n \in C_{\delta_n} \subset C_{\delta_0}$ and $\{y_n\}$ contains a subsequence $\{y_{n_k}\}$ weakly convergent to some $\bar{y}$. By the weak continuity of P,

$$-P(\bar{y}) = \text{w-}\lim_{k \to \infty} (-P(y_{n_k})) = \text{w-}\lim_{k \to \infty} (\delta_{n_k}\bar{p}_0 - P(y_{n_k})) \in D \tag{8.1v}$$

since D is weakly closed. Thus $\bar{y} \in S_0 (= Y_p \text{ for } p = 0)$ and $Q(\bar{y}) \geqq \tilde{Q}(0)$. On the other hand,

$$Q(\bar{y}) \leqq \liminf_{K \to \infty} Q(y_{n_k}) \leqq \liminf_{k \to \infty} \left(\tilde{Q}(\delta_{n_k}) + \frac{1}{n_k}\right) \leqq \liminf_{k \to \infty} \tilde{Q}(\delta_{n_k}).$$

Hence

$$\tilde{Q}(0) \leqq \liminf_{k \to \infty} \tilde{Q}(\delta_{n_k}). \tag{8.1vi}$$

Since $\{y_n\}$ is a sum of its weakly convergent subsequences, (8.1iii) holds and $\tilde{Q}(\cdot)$ is right continuous at zero.

Take any sequence $\bar{p}_n \to 0$. Since $\bar{p}_0 \in \operatorname{int} D$, there exists a sequence $\{\delta_n\} \subset [0, \delta_0]$, $\delta_n \to 0$ such that $\delta_n\bar{p}_0 - \bar{p}_n \in D$ for sufficiently large n. The relation

[4] If $H = R^n$, it is sufficient to assume that P is coordinate-wise weakly lower semicontinuous.

$\bar{p}_n - P(y) \in D$ implies then $\delta_n \bar{p}_0 - P(y) = \delta_n \bar{p}_0 - \bar{p}_n + \bar{p}_n - P(y) \in D$. Hence, if $y \in Y_{p_n}$, then $y \in S_{\delta_n}$ for large n and $\hat{Q}(\bar{p}_n) \geqq \tilde{Q}(\delta_n)$. Therefore

$$\liminf_{n\to\infty} \hat{Q}(\bar{p}_n) \geqq \lim_{n\to\infty} \tilde{Q}(\delta_n) = \tilde{Q}(0) = \hat{Q}(0) \tag{8.1vii}$$

and (8.1i) holds.

To prove (8.1ii), let $r > 0$ be the radius of a ball $B(r)$ with center at zero such that $\delta_0 \bar{p}_0 + B(r) \subset D$. Take $U_\nu = B(r/2)$. If $\bar{\nu} \in U_\nu$, then $\tilde{Y}_{\bar{\nu}} \subset C_{\delta_0}$ and is therefore weakly compact. For any $\zeta \geqq 0$ and any $\bar{\nu} \in U_\nu$ such that $\tilde{Y}_{\bar{\nu}}$ is nonempty, the following relation holds:

$$-\infty < \alpha = \inf_{y\in E} \Psi(y, \zeta, \bar{\nu}) \leqq \inf_{y\in Y_{\bar{\nu}}} \Psi(y, \zeta, \bar{\nu}) = \inf_{y\in Y_{\bar{\nu}}} Q(y) \leqq \varepsilon \tag{8.1viii}$$

since $\|(P(y) - \bar{\nu})^{D^*}\|^2 = 2K_{\bar{\nu}}(y) = 0$ for $y \in \tilde{Y}_{\bar{\nu}}$. If $\alpha = \varepsilon$, take $y_{\zeta\bar{\nu}}$ to be any point minimizing the weakly lower semicontinuous functional Q in the weakly compact set $\tilde{Y}_{\bar{\nu}}$. If $\alpha < \varepsilon$, take a sequence $\{y_n\}$ satisfying

$$\Psi(y_n, \zeta, \bar{\nu}) \leqq \alpha + \min\left(\frac{1}{n}, \varepsilon - \alpha\right) \leqq \varepsilon. \tag{8.1ix}$$

Since $K_{\bar{\nu}}(y) = \frac{1}{2}\|(P(y) - \bar{\nu})^{D^*}\|^2 \geqq 0$, then $Q(y_n) \leqq \varepsilon$ and

$$\begin{aligned} K_{\bar{\nu}}(y_n) &= \frac{1}{\zeta}(\Psi(y_n, \zeta, \bar{\nu}) - Q(y_n)) \leqq \frac{1}{\zeta}(\varepsilon - \beta); \\ \|(P(y_n) - \bar{\nu})^{D^*}\| &\leqq \sqrt{\frac{2}{\zeta}(\varepsilon - \beta)}. \end{aligned} \tag{8.1x}$$

By (2.5ii),

$$\|(P(y_n))^{D^*}\| \leqq \|(P(y_n) - \bar{\nu})^{D^*}\| + \|\bar{\nu}^{D^*}\| \leqq \sqrt{\frac{2}{\zeta}(\varepsilon - \beta)} + \frac{1}{2}r. \tag{8.1xi}$$

Hence, for sufficiently large $\bar{\zeta}$ and $\zeta \geqq \bar{\zeta}$, $\|(P(y_n))^{D^*}\| \leqq r$. But $(P(y_n))^{D^*} - P(y_n) \in D$ which reads $y_n \in Y_{(P(y_n))^{D^*}}$. Since $-(P(y_n))^{D^*} \in B(r)$, then $\{y_n\} \subset C_{\delta_0}$ and contains a weakly convergent subsequence, the limit of which can be taken for $y_{\zeta\bar{\nu}}$.

The proof of part (i) of Theorem 8.1 is classical; the conclusion 8.1(i) seems to be widely known. The criteria for inf-stability in the infinite-dimensional convex case were studied in [21], [10], [12]. See also [6], [25] for the nonconvex in R^n, and [33], [38].

Observe that if there exists an element $\bar{y} \in C_0$ with $p - P(\bar{y}) \in \operatorname{int} D$, then of course $\tilde{Y}_{\bar{\nu}} \neq \varnothing$ for any $\bar{\nu}$ in a neighborhood of zero. However, many optimization problems are posed with positive cones of empty interior. In that case, a similar theorem can be formulated.

THEOREM 8.2. *Let E be reflexive, Q be weakly lower semicontinuous and P continuous in weak topologies of E and H. Suppose there exist numbers ε and $\delta_0 > 0$ such that the sets*

$$\tilde{C}_\delta = \{y \in E : Q(y) \leqq \varepsilon, K(y) \leqq \delta\} \tag{8.2}$$

are bounded and nonempty for each $\delta \in [0, \delta_0]$. *Then*:

(i) *Problem* 5.1 *is in-stable. If, in addition,* $Q(y) \geqq \beta > -\infty$ *for all* $y \in E$, *then*:

(ii) *there exists* $\bar{\zeta} \geqq 0$ *and a neighborhood* U_ν *of zero in* H *such that for each* $\zeta \geqq \bar{\zeta}$ *and each* $\bar{\nu} \in U_\nu$ *satisfying* $\tilde{Y}_{\bar{\nu}} = \{y \in E : Q(y) \leqq \varepsilon,\ p + \bar{\nu} - P(y) \in D\} \neq \varnothing$, *there is a point* $y_{\zeta\bar{\nu}}$ *minimizing* $\Psi(\cdot, \zeta, p + \bar{\nu})$ *over* E.

The proof principally follows that of Theorem 8.1 with C_δ changed to $\tilde{C}_\delta$; it is therefore omitted.

Example 8.3. Suppose Q is weakly lower semicontinuous with bounded level sets and P weakly continuous, as in the theorem. Then $K(y) = \frac{1}{2}\|(P(y) - p)^{D^*}\|^2$ is weakly lower semicontinuous, and there is an $\varepsilon = \varepsilon(p)$ such that the sets $\tilde{C}_\delta = \{y \in E : Q(y) \leqq \varepsilon,\ K(y) \leqq \delta\}$ are bounded and nonempty whenever Y_p is. Hence, if Q, P are additionally either convex or differentiable, then by Theorem 8.2 and Corollary 6.5, the set $\mathscr{P}_2$ of all p such that the corresponding problems are normal is dense in the set $\mathscr{P}_1$ of all p such that the sets Y_p are nonempty (thus in the set of all well-posed problems of the type (5.1) satisfying the above moderate assumptions).

The questions of stability of order 2 and L-stability are more delicate and require much stronger assumptions. We are only able to discuss here the case of a finite number of inequality constraints. In the sequel it is assumed that $H = H_o \times R^n$, $D = \{0_{H_0}\} \times R^n_+$, $P = (P_o, \cdots, P_n)$, $P_0 : E \to H_0$, $P_i : E \to R$, $i = 1, \cdots, n$. The elements $p, \eta \in H$ are understood as $(n+1)$-tuples $(p_0, p_1, \cdots, p_n)$, $(\eta_0, \eta_1, \cdots, \eta_n)$. Recall the following.

DEFINITION 8.4. For the problem $\min_{y \in Y_0} Q(y)$, $Y_0 = \{y \in E : -P(y) \in D\}$ it is said that the second order sufficiency condition holds at a point $\hat{y} \in Y_0$, iff:

(i) Q, P are twice continuously Frechet differentiable in a neighborhood of $\hat{y}$,

(ii) there exists a normal Lagrange multiplier $\hat{\eta}$ for the problem at $\hat{y}$ and the range $\mathrm{im} P_{0y}(\hat{y})$ is closed in H_0,

(iii) the second derivative $L_{yy}(\hat{\eta}, \hat{y})$ of $L(\hat{\eta}, y) = Q(y) + \langle \hat{\eta}, P(y) \rangle$ satisfies the relation $\langle L_{yy}(\hat{\eta}, \hat{y})\bar{y}, \bar{y} \rangle \geqq \delta \|\bar{y}\|^2$ for some $\delta > 0$ and all $\bar{y}$ such that $P_{iy}(\hat{y})\bar{y} = 0$, $i \in J \cup \{0\}$, $P_{iy}(\hat{y})\bar{y} \geqq 0$, $i \in \partial J$ where

$$J = \{i : 1 \leqq i \leqq n,\ P_i(\hat{y}) = 0,\ \hat{\eta}_i \neq 0\}, \tag{8.4iv}$$

$$\partial J = \{i : 1 \leqq i \leqq n,\ P_i(\hat{y}) = 0,\ \hat{\eta}_i = 0\}. \tag{8.4v}$$

For the same problem, it is said that strict complementarity holds, iff $\partial J = \varnothing$.

It is proven easily (compare [37, p. 307] for the case of equality constraints only) that these conditions are indeed sufficient for $\hat{y}$ to be a local (not necessarily isolated) minimum. They are similar in formulation to second order sufficiency conditions for nonlinear programs in $H = R^n$ [7] but may also be easily translated into the language of the calculus of variations: the main condition is equivalent to strict positivity of the second variation of Lagrange functional.

The following theorem was first proved by Rockafellar in the case $E = R^m$, $H = R^n$.

THEOREM 8.5. *Suppose the problem* $\min_{y \in Y_0} Q(y)$, $Y_0 = \{y \in E : -P(y) \in D\}$ *satisfies the quadratic growth condition. Suppose that the second order sufficiency condition holds at* $\hat{y}$. *Assume that either* $E = R^m$ *or strict complementarity holds at* $\hat{y}$. *Then the problem is locally stable of degree* 2 *in a neighborhood of* $\hat{y}$. *If, in addition,*

for any neighborhood A of $\hat{y}$ there is a neighborhood U_p of zero in H such that $\hat{Q}(p) = \inf_{y \in Y_p} Q(y) = \inf_{y \in A \cap Y_p} Q(y)$ for all $p \in U_p$, then the problem is stable of degree 2.

The proof is omitted since in the case $E = R^m$ it is essentially given in [22] and in the general case requires only some modification based on the closed range theorem [28].

Somewhat stronger conditions are required for the L-stability. The following theorem was proved in [25] for $E = R^m$, $H = R^n$.

THEOREM 8.6. *Suppose E is Hilbert, and let the second order sufficiency conditions with strict complementarity hold at $\hat{y} \in Y_0 = \{y \in E : -P(y) \in D\}$ for the problem* $\min_{y \in Y_0} Q(y)$. *Denote $H_J = \{p \in H : p_i = 0, i \notin J \cup \{0\}\}$ and let P_J be the restriction of P to H_J. Assume $P_{Jy}(\hat{y})$ is surjective. Then the problem is locally* L-*stable in a neighborhood of $\hat{y}$.*

Proof. (a) Let A_1 denote a neighborhood of $\hat{y}$ such that $P_i(y) < -\varepsilon < 0$ for $i \notin J \cup \{0\}$ and $P_{Jy}(y)$ is surjective whenever $y \in A_1$. Let U_1 be a neighborhood of zero in H such that $|p_i| < \varepsilon$ for $i \notin J \cup \{0\}$ whenever $p \in U_1$. Denote by p_J, η_J the elements of the Hilbert space H_J. Observe that $y(p)$ is a local solution to the problem: minimize $Q(y)$ over $Y_p \cap A_1$ (where $p \in U_1$), and $\eta(p)$ is a normal Lagrange multiplier for this problem at $y(p)$ if and only if $y(p) = y(p_J)$ is also a local solution to the problem:

$$\min_{y \in Y_{pJ} \cap A_1} Q(y); \quad Y_{pJ} = \{y \in E : p_J - P_J(y) \in H_J \cap D\} \tag{8.6i}$$

while $\eta(p) \in H_J$, $\eta(p) = \eta_J(p)$ is a normal Lagrange multiplier to the problem (8.6i) at $y(p_J)$:

$$Q_y(y(p_J)) + P^*_{Jy}(y(p_J))\eta_J(p_J) = 0, \tag{8.6ii}$$

$$\eta_i(p_J) \geqq 0, \quad i \in J; \quad \eta_i(p_J)(P_i(y(p_J)) - p_i) = 0, \quad i \in J. \tag{8.6iii}$$

This is readily verified since $p_i - P_i(y) > 0$ for $i \notin J \cup \{0\}$ whenever $y \in A_1$, $p \in U_1$.

It is therefore sufficient to investigate the local L-stability of the problem (8.6i) at $p_J = 0$, i.e., to establish the existence and Lipschitz-continuity of the mapping $U_{0J} \ni p_y \mapsto \eta_J(p_J)$ on some neighborhood U_{0J} of zero in H_J. Having this done, one can set $U_p = U_1 \cap (U_{0J} \times H_J^{\perp})$ in Definition 7.5 to obtain the local L-stability of the original problem as claimed.

Notice first that since $P_{Jy}(y)$ is surjective for $y \in A_1$ then [11] any $y(p_J)$, being a local solution to (8.6i), must satisfy (8.6ii), (8.6iii) for a uniquely determined $\eta_J(p_J)$.

(b) Observe that, according to the closed range theorem [28], since the range of $P_{Jy}(\hat{y})$ is closed, then $P^*_{Jy}(\hat{y})$ maps H_J onto $\operatorname{Im} P^*_{Jy}(\hat{y}) = (\ker P_{Jy}(\hat{y}))^{\perp}$. Let π denote the projection operator in E onto $\ker P_{Jy}(\hat{y})$. According to the LaxMilgram theorem [28], the condition $\langle L_{yy}(\hat{\eta}, \hat{y})\bar{y}, \bar{y}\rangle \geqq \delta\|\bar{y}\|^2$, $\delta > 0$, $\bar{y} \in \ker P_{Jy}(\hat{y})$ implies that $\pi \circ L_{yy}(\hat{\eta}, \hat{y})$ maps $\ker P_{Jy}(\hat{y})$ onto itself. Consider the operator

$$M = \begin{bmatrix} L_{yy}(\hat{\eta}, \hat{y}) & P^*_{Jy}(\hat{y}) \\ P_{Jy}(\hat{y}) & 0 \end{bmatrix} : E \times H_J \to E \times H_J. \tag{8.6iv}$$

We shall show that the operator M is surjective, that is, for every $(z, s) \in E \times H_J$

there exists a pair (y, η_J) satisfying $M \circ (y, \eta_J) = (z, s)$. Indeed, since $P_{Jy}(\hat{y})$ is onto, there exists y' such that

$$P_{Jy}(\hat{y})y' = s. \tag{8.6v}$$

Let $w_1 = \pi z - \pi \circ L_{yy}(\hat{\eta}, \hat{y})\bar{y}$. Since $w \in \ker P_{Jy}(\hat{y})$, there is y'' such that

$$w_1 = \pi \circ L_{yy}(\hat{\eta}, \hat{y})y'', \qquad y'' \in \ker P_{Jy}(\hat{y}). \tag{8.6vi}$$

Let $w_2 = (I - \pi)(z - L_{yy}(\hat{\eta}, \hat{y})(y' + y''))$. Since $w_2 \in (\ker P_{Jy}(\hat{y}))^{\perp}$, there exists η_J such that

$$w_2 = P^*_{Jy}(\hat{y})\eta_J. \tag{8.6vii}$$

Combine (8.6v), (8.6vi), (8.6vii) to obtain

$$M(y, \eta_J) = \begin{bmatrix} L_{yy}(\hat{\eta}, \hat{y})y + P^*_{Jy}(\hat{y})\eta_J \\ P_{Jy}(\hat{y})_y \end{bmatrix} = \begin{bmatrix} z \\ s \end{bmatrix}; \qquad y = y' + y''. \tag{8.6viii}$$

Hence the operator M is surjective; since it is selfadjoint, it is also invertible. By the implicit function theorem, the equations

$$\begin{aligned} Q_y(y) + P^*_{Jy}(y)\eta_J &= 0, \\ P_J(y) - p_J &= 0 \end{aligned} \tag{8.6ix}$$

determine (y, η_J) as a differentiable function of p_J in a neighborhood U_{2J} of zero in H_J (observe that these equations are satisfied by $(\hat{y}, \hat{\eta}_J)$ for $p_J = 0$, the derivative M of the left-hand side is invertible at $(\hat{y}, \hat{\eta}_J)$ and this derivative is continuous as a function of (y, η_J)).

(c) Denote (y, η_J) satisfying (8.6ix) by $y = f(p_J)$, $\eta_J = g(p_J)$. It remains to prove that, for p_J in some U_{0J}, $y = f(p_J)$ are uniquely defined local solutions to the problems (8.6i) and $\eta_J = g(p_J)$ are uniquely defined normal Lagrange multipliers for these problems (being differentiable, $g(\cdot)$ is Lipschitz continuous; hence the conditions of local L-stability are all satisfied).

Take U_{2J} smaller, if necessary, to obtain $\eta_i = g_i(p_J) > 0$, $i \in J$ for $p_J \in U_{2J}$. Choose neighborhoods $A \subset A_1$ of $\hat{y}$ and $U_{0J} \subset U_{2J}$ of zero satisfying $P_J(A) = U_{0J}$, $f(U_{0J}) = A$ and such that the inequality $\langle L_{yy}(g(p_J), f(p_J))\bar{y} \rangle \geqq \frac{1}{2}\delta \|\bar{y}\|^2$ holds for $\bar{y} \in \ker P_{Jy}(f(p_J))$ whenever $p_J \in U_{0J}$ (the possibility of such a choice follows from the assumption of continuous second order differentiability). Then the second order sufficiency condition with strong complementarity holds at $y = f(p_J)$, $p_J \in U_{0J}$ so that $f(p_J)$ is a local solution to (8.6i) in A (not necessarily isolated). It remains to show that for fixed $p_J = \tilde{p}_J \in U_{oJ}$, $f(p_J)$ is the unique local solution to (8.6i) in A. Let $y(\tilde{p}_J)$ be any local solution to that problem; then (8.6ii) and (8.6iii) are satisfied with $p_J = \tilde{p}_J$ for some $\eta_J(p_J) = \hat{\eta}_J$. If $\tilde{\eta}_i > 0$, $i \in J$, then by (8.6iii), $P_J(y(\tilde{p}_J)) = \tilde{p}_J$ so that (8.6ix) is satisfied by $y = y(\tilde{p}_y)$, $\eta_j = \tilde{\eta}_y$ and $p_J = \tilde{p}_J$. In this case $y(\tilde{p}_J) = f(\tilde{p}_J)$, $\hat{\eta}_J = g(\tilde{p}_J)$ and we are done. But $\tilde{\eta}_i$, $i \in J$, must be positive. Indeed, $y = y(\tilde{p}_J)$ and $\eta_J = \tilde{\eta}_J$ always satisfy (8.6ix) with $p_J = P(y(\tilde{p}_J)) \in U_{oJ}$. Therefore $\tilde{\eta}_J = g(p_J)$ and $\tilde{\eta}_i > 0$, $i \in J$.

Therefore $\eta_J(p_J) = g(p_J)$ is a unique Lagrange multiplier for the problem (8.6i) in A; the mapping $U_{0J} \ni p_J \mapsto (p_J)$ is continuously Fréchet differentiable, in particular, Lipschitz-continuous on sufficiently small U_{0J}.

9. Conclusions. The properties of the projection on a positive cone make it possible to define penalty functionals and to develop a duality theory for infinite-dimensional problems with operator constraints. Two groups of results are of fundamental importance. The Theorems 4.11, 4.12 and the Corollary 4.13 explain the relations between dual methods and penalty functional techniques. Under appropriate stability assumptions, the solution of the original problem inf (P) can be approximated by a sequence obtained by subsequent unconstrained minimizations of the shifted penalty functional $\Psi(\cdot, \zeta_n, \nu_n)$—or, equivalently, of the augmented Lagrangian $\Lambda(\zeta_n, \nu_n, \cdot)$— if the sequence $\{(\zeta_n, \nu_n)\}$ is maximizing the dual problem. The saddle point (inf (P) = max (D)) can be achieved at a bounded pair $(\hat{\zeta}, \hat{\nu})$ iff the original problem is stable of degree 2; in this case, a sufficiently large $\hat{\zeta}$ can be kept constant and a penalty shift algorithm of changing ν_n can be applied. If the original problem is inf-stable, but not stable of degree 2, an approximate saddle point (inf (P) = sup (D)) can only be achieved at $\zeta \to \infty$; hence a penalty increase algorithm with $\zeta_n \to \infty$ must be applied, whereas ν_n can be kept constant or changed. Thus the penalty techniques are actually dual methods [44].

The Lemmas 3.9, 3.9′ explain other aspects of penalty techniques—or dual methods. If the penalty functional $\Psi(\cdot, \zeta, \nu)$—or $\Lambda(\zeta, \nu, \cdot)$—does possess a minimum at $\bar{y}$, then the point $\bar{y}$ is actually a solution of an optimization problem which differs from the original one only in the constraining value $\bar{p}$ (Everett theorem). Thus, penalty techniques are two-level algorithms with coordination of constraining values. The applications of dual methods to two-level coordination algorithms are known; but they are not the only possible coordination methods. In addition to typical saddle-point algorithms, two types of coordination methods can be considered. The increased penalty technique coordinates the violation of constraints by an increase of penalty coefficients only, and is convergent under rather weak assumptions inf-stability. No normality assumptions are required, although the original optimization problem is approximated by a sequence of normal ones. But the increased penalty technique is badly conditioned numerically. Hence, if the problem satisfies stronger assumptions, normality, stability of degree 2, L-stability and the shifted penalty technique is convergent, the latter gives better numerical results. The shifted penalty technique coordinates the violation of constraints by penalty shifts or, equivalently, by changing Lagrange multipliers. Most universal is the shifted-increased Algorithm 7.4 which switches to penalty increase if penalty shifts fail to provide for a given convergence rate.

The conditions for convergence of penalty techniques are related to various degrees of stability of optimization problems. The inf-stability follows from the weak continuity of a minimized functional and constraining operator, and from the boundedness of some level sets related to these functions. The stability of degree 2 and L-stability require much stronger assumptions, as second order sufficiency conditions, etc.

Acknowledgment. The authors are indebted to R. T. Rockafellar for a useful discussion of the problem and for most valuable, yet unpublished references. They also acknowledge gratefully the referees' remarks, which helped to improve the manuscript and complete the list of references.

REFERENCES

[1] K. J. ARROW, L. HURWICZ AND H. UZAWA, *Studies in Linear and Nonlinear Programming*, Stanford University Press, Stanford, Calif., 1958.

[2] A. V. BALAKRISHNAN, *A computational approach to the maximum principle*, J. Comput. System Sci., 5 (1971), pp. 163–191.

[3] R. COURANT, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), p. 1–23.

[4] J. W. DANIEL, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[5] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Interscience, New York, 1958.

[6] J. P. EVANS AND F. J. GOULD, *Stability in nonlinear programming*, Operations Res., 18 (1970), pp. 107–119.

[7] A. FIACCO AND G. MCCORMICK, *Sequential Unconstrained Nonlinear Programming*, John Wiley, New York, 1968.

[8] M. R. HESTENES, *Multiplier and gradient methods*, Computing Methods in Optimization Problems—2, L. A. Zadeh, L. W. Neustadt and A. V. Balakrishnan, eds., Academic Press, New York, 1969, pp. 143–164.

[9] R. B. HOLMES, *A Course on Optimization and Best Approximation*, Springer-Verlag, Berlin, 1972.

[10] J. L. JOLY AND P. J. LAURENT, *Stability and duality in convex minimization problems*, Rev. Française Informat. Recherche Operationelle, R2 (1971), pp. 3–42.

[11] S. KURCYUSZ, *On the existence of Lagrange multipliers for infinite dimensional extremal problems*, Bull. Polish Acad. Sci., XXII (1974), No. 9.

[12] P. J. LAURENT, *Approximation et Optimisation*, Hermann, Paris, 1972.

[13] E. S. LEVITIN AND B. T. POLIAK, *Convergence of minimizing sequences in conditional extremum problems*, Soviet Math. Dokl., 7 (1967), pp. 764–767.

[14] D. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[15] O. L. MANGASARIAN, *Unconstrained Lagrangians in nonlinear programming*, this Journal, 13 (1975), pp. 772–791.

[16] A. MIELE, E. E. CRAGG, R. R. IVER AND A. V. LEVY, *Use of the augmented penalty function in mathematical programming problems, Part I*, J. Optimization Theory Appl., 8 (1971), pp. 115–130.

[17] ———, *Use of the augmented penalty function in mathematical programming problems, Part II*, Ibid., 8 (1971), pp. 131–153.

[18] J. J. MOREAU, *Décomposition orthogonale d'un espace hilbertien selon deux cones mutuellement polaires*, C.R. Acad. Sci. Paris, Sèr. A–B, 225 (1962), pp. 238–240.

[19] ———, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

[20] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.

[21] R. T. ROCKAFELLAR, *Duality and stability in extremum problems involving convex functions*, Pacific J. Math., 21 (1967), pp. 167–187.

[22] ———, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.

[23] ———, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Programming, 5 (1973), pp. 354–373.

[24] ———, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optimization Theory Appl., 12 (1973), pp. 555–562.

[25] A. P. WIERZBICKI, *A penalty function shifting method in constrained static optimization and its convergence properties*, Archiwum Automatyki i Telemechaniki, 16 (1971), pp. 395–416.

[26] A. P. WIERZBICKI AND A. HATKO, *Computational methods in Hilbert space for optimal control problems with delays*, Proc. of 5th IFIP Conf. on Optimization Techniques, Rome, Springer-Verlag, Berlin, 1973.

[27] A. P. WIERZBICKI, A. JANIAK, S. KURCYUSZ, A. LEWANDOWSKI AND T. ROGOWSKI, *Penalty functional methods and their applications in optimal control problems*, Institute of Automatics Tech. Rep., Technical University of Warsaw, Poland, 1973. (In Polish.)

[28] K. YOSIDA, *Functional Analysis*. Springer-Verlag, Berlin, 1966.
[29] E. H. ZARANTONELLO, *Projections on convex sets in Hilbert space and spectral theory*, Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York–London, 1971.
[30] M. BELLMORE, H. J. GREENBERG AND J. J. JARVIS, *Generalized penalty-function concepts in mathematical optimization*, Operations Res., 18 (1970), pp. 229–252.
[31] D. P. BERTSEKAS, *Combined primal-dual and penalty methods for constrained minimization*, this Journal, 13 (1975), pp. 521–544.
[32] ———, *Multiplier methods: a survey*, Automatica, 12 (1976), pp. 133–145.
[33] S. DOLECKI, *Bounded controlling sequences, inf-stability and certain penalty procedures*, Appl. Math. and Optimization, to appear.
[34] F. J. GOULD, *Extensions of Lagrange multipliers in nonlinear programming*, SIAM J. Appl. Math., 17 (1969), pp. 1280–1297.
[35] ———, *Nonlinear pricing: applications to concave programming*, Operations Res., 19 (1971), pp. 1026–1035.
[36] H. J. GREENBERG, *The generalized penalty function surrogate model*, Ibid., 21 (1973), pp. 162–178.
[37] A. D. IOFFE AND W. M. TIKHOMIROV, *Theory of Extremal Problems*, Nauka, Moscow, 1974. (In Russian.)
[38] S. KURCYUSZ, *Some remarks on generalized Lagrangians*, Proc. 7th IFIP Conf. on Optimization, Nice, France, September, 1975, Springer-Verlag, Berlin.
[39] K. M. PRZYLUSKI, *Application of the shifted penalty method to dynamic optimization of delay processes*, M.Sc. thesis, Institute of Automatic Control, Technical University of Warsaw, Warsaw, Poland, 1974. (In Polish.)
[40] R. T. ROCKAFELLAR, *A talk given at the 7th International Symposium on Mathematical Programming, The Hague, 1970*, Proc. 4th Conf. on Probability, Brasov, Romania, 1971.
[41] J. D. ROODE, *Generalized Lagrangian functions and mathematical programming*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969.
[42] R. D. RUPP, *A method for solving a quadratic optimal control problem*, J. Optimization Theory Appl., 9 (1972), pp. 238–250.
[43] ———, *A nonlinear optimal control minimization technique*, Trans. Amer. Math. Soc., 178 (1973), pp. 357–381.
[44] W. I. ZANGWILL, *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, N. J., 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

PROPER EFFICIENT POINTS FOR MAXIMIZATIONS WITH RESPECT TO CONES*

J. BORWEIN†

Abstract. Proper efficient points (Pareto maxima) are defined in tangent cone terms and are characterized by the existence of equivalent real-valued maximization problems.

1. Introduction. Suppose that X and Y are (locally) convex (topological vector) spaces over R and that $S \subset Y$ is a nontrivial closed convex cone which induces a partial ordering $\leqq_s$. The vector maximization problem for f mapping X into Y and $A \subset X$,

$$\max f(x) \quad \text{subject to } x \in A \quad \text{(VMP)},$$

is the problem of finding all efficient points $\bar{x} \in A$: $\bar{x}$ is said to be efficient (Pareto optimal) if $x \in A$ and

$$f(x) \geqq_s f(\bar{x}), \quad f(x) \neq f(\bar{x}) \quad \text{implies that } \bar{x} \notin A.$$

Geffrion [1] has studied this problem in finite dimensions with the coordinate ordering and has suggested a restriction to "proper" efficient points which allows for a reasonable characterization. Kuhn and Tucker [4] have also used the term but their notion requires differentiability and appears too broad for satisfactory analysis (see below).

This paper proposes a general notion of properness which is defined in terms of tangent cones as developed by Varaiya [8], Guignard [2], Zlobec [9] and others and which coincides with Geffrion's definition in the central case.

2. Preliminaries. Throughout the paper all spaces are assumed Hausdorff and convex and "$\leqq_s$" is the partial order induced by S.

DEFINITION 1. Suppose $C \subset X$ and $\bar{x} \in \bar{C}$. The *tangent cone* to C at $\bar{x}$ is defined to be the set of limits of the form $h = \lim t_n(x_n - \bar{x})$ with $\{t_n\}$ a sequence of nonnegative real numbers and $\{x_n\} \subset C$ a sequence with limit $\bar{x}$. It is denoted $T(C, \bar{x})$.

When X is metrizable, $T(C, \bar{x})$ is closed. It is always a nonempty cone containing 0, but need not be closed in general spaces unless defined in terms of nets which leads to other embarrassments. The closed convex hull of $T(C, \bar{x})$ is called the *pseudo-tangent* cone and is denoted by $P(C, \bar{x})$. Various properties of pseudo-tangent cones can be found in [2], [8], [9] and a forthcoming paper of the author.

DEFINITION 2. A point $\bar{x}$ will be said to be a *proper efficient point of* (VMP) if it is efficient and

$$\bar{T}(f(A) - S, f(\bar{x})) \cap S = 0. \tag{1}$$

* Received by the editors March 5, 1975, and in final revised form April 14, 1975.

† Department of Mathematics, Dalhousie University, Halifax, Nova Scotia, Canada. This research was supported in part by National Research Council Grants A7751 and A7675.

DEFINITION 3. When $f(x)=(f_1(x),\cdots,f_p(x))$ maps R^k into R^p, Geffrion defines $\bar{x}$ to be properly efficient with respect to the coordinate ordering if it is efficient and if there is some real $M>0$, such that for each i one has

$$\frac{f_i(x)-f_i(\bar{x})}{f_j(\bar{x})-f_j(x)}\leqq M \tag{2}$$

holding for some j with $f_j(x)<f_j(\bar{x})$, whenever $x\in A$ and $f_i(x)>f_i(\bar{x})$.

It is a simple matter to verify that in this later framework, (1) is a weaker requirement on (VMP) than (2) and, in fact, that when f is continuous (1) implies the local efficiency of $\bar{x}$ with respect to the coordinate ordering.

PROPOSITION 1. *Suppose $\bar{x}$ is Geffrion proper efficient for f over A. Then $\bar{x}$ satisfies*

$$T(f(A)-R^{n+},f(\bar{x}))\cap R^{n+}=0.$$

Proof. Suppose $k\neq 0\in R^{n+}\cap T(f(A)-R^{n+},f(\bar{x}))$. Without loss of generality one may assume that $k_1>1$, $k_i\geqq 0$, $i=2,\cdots,n$. Let

$$t_n(f(x_n)-r_n^+-f(\bar{x}))\to k,$$

where $r_n^+\in R_n^+$, $t_n\geqq 0$ and $f(x_n)-r_n^+\to f(\bar{x})$ with $x_n\in A$. By choosing a subsequence one can assume that

$$\tilde{I}=\{i|f_i(x_n)<f_i(\bar{x})\}$$

is constant for all n (and nonempty since $\bar{x}$ is Pareto efficient). Set $M>0$. Then for $n\geqq n_0$,

$$f_1(x_n)-f_1(\bar{x})\geqq t_{n/2}^{-1},$$
$$f_i(x_n)-f_i(\bar{x})\geqq -t_n^{-1}/2M.$$

Then for all $i\in\tilde{I}$, one has

$$0<f_i(\bar{x})-f_i(x_n)\leqq t_n^{-1}/2M,$$

and for $n\geqq n_0$,

$$\frac{f_1(x_n)-f_1(\bar{x})}{f_i(\bar{x})-f_i(x_n)}\geqq\frac{t_{n/2}^{-1}}{t_n^{-1}/2M}=M,$$

which contradicts Geffrion's definition. □

DEFINITION 4. Suppose X' is the topological dual of X. The *dual cone K^+ of a convex cone $K\subset X$* is defined by

$$K^+=\{x'\in X'|x'(x)\geqq 0,\ \forall x\in K\},$$

while the *dual cone $(K')^+$ of a convex cone $K'\subset X'$* is defined by

$$(K')^+=\{x\in X|x'(x)\geqq 0,\ \forall x\in K'\}.$$

It follows from these definitions that (i) K^+ is weakly* closed; (ii) $(K^+)^+=\bar{K}$; (iii) $(\bar{K}_1\cap\bar{K}_2)^+=\mathrm{cl}\,(K_1^+ + K_2^+)$ (with closure in the weak* topology). K^+ is well-defined even if K is not a convex cone. In this case $(K^+)^+$ is the closed convex hull of K (denoted $[\bar{K}]$).

These facts all hold in convex spaces. Proofs can be found for normed spaces in [7].

3. Geometric motivation. The main aim of Definition 2 is to provide a notion of properness which can be applied when the cone S is not the orthant ordering in R^n and is not even polyhedral. Consider (VMP):

$$\max_{x\in A} f(x) = \max_{z\in f(A)} z = \max_{z\in f(A)-S=E} z.$$

This last equivalence is introduced so that, in the case that f is concave with respect to S and A is convex, the optimization in the image space is still a concave problem. Definition 2 says that $\bar{x}$ is proper when, with $\bar{z}=f(\bar{x})$, one has

$$T(E,\bar{z})\cap S=0.$$

In general, then, the concept of properness is an attempt to remove those efficient points which can be approached in directions which point into S. In the case that $S=R^{n+}$, this can be done by considering the components separately; in more general orderings a more technical notion of direction must be introduced. This is done herein with tangent cones. Consider the following examples.

Example 1. Let $X=R^3$, $S=\{x|x=(x_1,x_2,x_3),\ x_3\geqq 0,\ x_1^2+x_2^2\leqq x_3^2\}$. Let $A=\{x|\|x\|\leqq 1\}$ and let $f=I$. The efficient points for (VMP) are $\{x|\|x\|=1,\ x\in S\}$. Since $E=A-S$ is convex, $T(E,f(\bar{x}))$ is the smallest closed convex cone containing E with vertex at $f(\bar{x})=\bar{x}$. It is easily seen that for those x with $\|x\|=1$, $x\in S$ and $x_1^2+x_2^2=x_3^2$ (or $x_3=\frac{1}{\sqrt{2}}$, $x_1^2+x_2^2=\frac{1}{2}$), this cone has a boundary ray in common with S; while for any other efficient x this cannot happen. In this case the efficient improper points form the relative boundary of the efficient points on A.

Example 2. Let $X=R^2$, $S=\{x|x_2\geqq 0, x_1\geqq x_2\}$. Let $A=\{(x_1,x_2)|x_1^2+x_2^2\leqq 1, x_1\geqq 0, x_2\geqq 0\}$ and let $f=I$. The efficient points are those x on the arc in A for which $x_1^2+x_2^2=1$. Again $T(f(A)-S,f(x))$ is the smallest closed convex cone containing $A-S$ at $f(x)$. This only intersects $S|\{0\}$ when $x=f(x)$ lies at the upper endpoint of the arc. There is, therefore, only one improper point $(0,1)$.

Example 3. Consider X,S,f as in the previous example, and let $A_1=A\cap\{(x_1,x_2)|x_1\leqq\frac{1}{2}$ or $x_2=0\}$. The efficient points are now

$$\{(1,0)\}\cup\{(x_1,x_2)|x_1\leqq\tfrac{1}{2}, x_1^2+x_2^2=1\}\cup\{(\tfrac{1}{2},x_2)|x_2\geqq 0, x_2\leqq\tfrac{1}{2}\}.$$

The problem is no longer convex, and $T(A-S,f(x))$ is easily calculated. Only $(\frac{1}{2},0)$ and $(0,1)$ are improper.

We see that properness gives us a criterion for excluding some efficient points (those which can be "approached from within S") for which, as will be shown, equivalent real maximizations fail to exist.

4. Some cone separation theorems. It is necessary to establish two abstract separation theorems for convex cones before proving general multiplier theorems for (VMP).

PROPOSITION 2. *Suppose N, S are closed convex cones in X and that $N\cap S=0$. Suppose that the dual cone S^+ has nonempty interior in some topology τ which gives*

X as the dual of X'. Then there is some $s^+ \in (S^+)^0$ *with* $-s^+ \in N^+$ *and*

$$s^+(s) > 0 \quad \forall s \in S/\{0\}. \tag{3}$$

In fact this last condition is equivalent to $s^+ \in (S^+)^0$.

Proof. Using property (iii) under Definition 4 one sees that

$$\{0\}^+ = X' = (N \cap S)^+ = \tau - \mathrm{cl}\,(N^+ + S^+),$$

since τ is a topology of the dual pair (X', X). Let $s' \in (S^+)^0$. There is then some net $-s'_\alpha = n^+_\alpha + s^+_\alpha$ with $n^+_\alpha \in N^+$, $s^+_\alpha \in S^+$ and $n^+_\alpha + s^+_\alpha$ tending (τ) to $-s'$. Since $-s'$ is a τ-interior point for $-S^+$, it follows that for $\alpha \geqq \alpha_0$,

$$-s'_\alpha = n^+_\alpha + s^+_\alpha \in -(S^+)^0.$$

Thus $n^+_\alpha = -(s'_\alpha + s^+_\alpha) \in -(S^+)^0 - S^+ \subset -(S^+)^0$. It follows that $n^+_\alpha \in N^+$ and satisfies (3). Conversely, if s^+ exists satisfying (3) and $(S^+)^0 \neq \varnothing$, then

$$S \cap \{x | s^+(x) \leqq 0\} = 0,$$

and one can apply the previous argument to the two sets S and $\{x|s^+(x) \leqq 0\}$ to derive that some $n^+ \in N^+ = \{x|s^+(x) \leqq 0\}^+$ is also in $(-S^+)^0$. But $N^+ = \bigcup_{\lambda \leqq 0} \lambda s^+$ and, since $0 \notin (S^+)^0$, $n^+ = \lambda s^+$, $\lambda < 0$, which implies that $s^+ \in (S^+)^0$. □

In particular, the theorem holds for any cone S in R^n which is pointed $(S \cap -S = 0)$, since this means $S^+ \subset R^n$ has nonempty interior. In the case that $(S^+)^0$ cannot be guaranteed nonempty, one can still prove the existence of s^+ satisfying (3) if one requires that S have a compact base B.

PROPOSITION 3. *Suppose N, S are closed convex cones in X such that* $N \cap S = 0$. *Suppose that* $S \cap -S = 0$ *and that S is locally compact (has a compact neighborhood base in the relative topology on S). Then one can find* $s^+ \in -N^+$ *satisfying* (3).

Proof. The local compactness condition on S guarantees by [3, (2.4)] that one can find a compact convex subset B of S, such that $0 \notin B$, with $S = \bigcup_{\lambda \geqq 0} \lambda B$.

It follows that B and N can be strictly separated [5] and that there is some $s^+ \in X'$ with

$$s^+ \in -N^+ \quad \text{and} \quad s^+(x) > 0 \quad \forall x \in B.$$

It follows immediately that $s^+(s) > 0\ \forall s \in S/\{0\}$. □

Remark. It is easy to see that in a locally convex space a pointed cone S is locally compact exactly when it has a compact generating base. That is: $S = \bigcup_{\lambda \geqq 0} \lambda B$ where B is compact, convex and $0 \notin B$.

5. Equivalences. One can now derive the basic characterization of proper efficient points.

THEOREM 1. *Suppose that* $\bar{x}$ *is optimal for*

$$\max s^+ f(x) \quad \textit{subject to } x \in A \quad (P(s^+))$$

and that s^+ *satisfies* (3). *Then* $\bar{x}$ *is a proper efficient point.*

Proof. Suppose $h \in T(f(A) - S, f(\bar{x}))$. Then

$$h_n = t_n(f(x_n) - s_n - f(\bar{x})) \to h$$

with $t_n \geqq 0$, $f(x_n) - s_n \to f(\bar{x})$, $x_n \in A$, $s_n \in S$. For each n, $s^+ f(x_n) \leqq s^+ f(\bar{x})$ since $\bar{x}$ is optimal for $(P(s^+))$ and so

$$\lim t_n(s^+(f(x_n) - s_n) - s^+ f(\bar{x})) \leqq 0.$$

It follows that $s^+(h) \leqq 0 \; \forall h \in \bar{T}(f(A) - S, f(\bar{x}))$. Were h to belong to $S/\{0\}$, one would have $s^+(h) > 0$ since (3) holds. This is impossible and $\bar{T}(f(a) - S, f(\bar{x})) \cap S = 0$.

It is clear that if $\bar{x}$ were not efficient and $x_1 \in A$ with $f(x_1) \geqq_s f(\bar{x})$, that the definition of s^+ would imply that $s^+ f(x_1) > s^+ f(\bar{x})$ which contradicts the optimality of $\bar{x}$ for $(P(x^+))$. □

THEOREM 2. *Suppose that f is concave with respect to S and that A is convex. Suppose X and S satisfy the hypotheses of Proposition 2 or 3. Then $\bar{x}$ is properly efficient for (VMP) if and only if $\bar{x}$ is optimal for $(P(s^+))$ for some s^+ satisfying* (3).

Proof. Sufficiency was proved in Theorem 1. Suppose now that $\bar{x}$ is properly efficient. Since f is concave and A is convex, $f(A) - S = \{z | f(x) \geqq z, x \in A\}$ is convex. An elementary proposition in [3] shows that in this case,

$$f(A) - S - f(\bar{x}) \subset \bar{T}(f(A) - S, f(\bar{x})) = N \tag{4}$$

and that N is convex. Because $\bar{x}$ is assumed proper, $N \cap S = 0$. Since either Proposition 2 or 3 holds, s^+ satisfying (3) exists with $-s^+ \in N^+$. In particular, since (4) holds,

$$s^+(f(x) - s - f(\bar{x})) \leqq 0 \quad \forall x \in A, \quad s \in S.$$

Setting $s = 0$ shows that $\bar{x}$ is optimal for $(P(s^+))$ with $s^+(s) > 0 \; \forall s \in S/\{0\}$. □

In finite dimensions with coordinate ordering, this equivalence is exactly the same as Geffrion's. Thus for coordinate concave programs, Definitions 2 and 3 coincide. It is worth noting that the use of the set $\bar{T}(f(A) - S, f(\bar{x}))$ rather than the smaller $\bar{T}(f(A), f(\bar{x}))$ is motivated by the need for (4) to hold. If one desires the equivalence of Theorem 2 only for problems with $f(A)$ convex (which includes A convex, f linear) one need only require that

$$\bar{T}(f(A), f(\bar{x})) \cap S = 0. \tag{5}$$

Example. $f_1(x) = (-x^2, x, x)$ is an example of a coordinate concave function satisfying (1) or (2) on R^n at 0; $f_2(x) = (-x^2, x, 0)$ does not. This can be seen either directly from Definition 2 or from the respective presence and absence of positive multipliers when one applies Theorem 2.

If the hypotheses of Theorem 2 hold and the convex feasible set A is, in fact, $\{x | g(x) \geqq_B 0, x \in C\}$ for some function g mapping X into Z, concave with respect to B, and some convex C, one has the following "Lagrange" multiplier theorem.

THEOREM 3. *Suppose B is a convex cone with interior and that $g(x_0) \in B^0$. Suppose $\bar{x}$ is a proper efficient point for* (VMP) *with $A = \{x | g(x) \geqq 0, x \in C\}$. Then there is some continuous linear mapping T of Z into Y such that $T(B) \subset S$ and $Tg(\bar{x}) = 0$ with $\bar{x}$ properly efficient for the unconstrained concave problem*

$$\max f(x) + Tg(x) \quad \textit{subject to } x \in C \quad \text{(UCP)}.$$

Proof. Apply Theorem 2 to produce s^+ satisfying (3) with

$$s^+ f(\bar{x}) = \max s^+ f(x) \quad \text{subject to } g(x) \geqq 0, \quad x \in C.$$

The standard Lagrange multiplier theorem ([9]) guarantees that $u^+ \in B^+$ exists with $u^+ g(\bar{x}) = 0$ and

$$s^+ f(\bar{x}) \geqq s^+ f(x) + u^+ g(x) \quad \forall x \in C. \tag{6}$$

Choose $s \in S$ with $s^+(s) = 1$. Let $T_0: Z \to Y$ be defined by $T_0(z) = u^+(z)s$. Then $T_0(B) \subset S$, T_0 is continuous, linear and $T_0 g(\bar{x}) = 0$. Equation (6) can be rewritten as

$$s^+(f(x) + T_0 g(x)) \leqq s^+(f(\bar{x}) + T_0 g(\bar{x})), \quad x \in C,$$

from which it follows, using Theorem 2 again, that $\bar{x}$ is a proper efficient point for (UCP) with $T = T_0$. □

6. Differential conditions. Consider now the Pareto maximization problem

$$\max f(x) \quad \text{subject to } g(x) \in B, \quad x \in C \quad \text{(P)},$$

where $f: X \to Y$, $g: X \to Y$ are Fréchet differentiable functions between normed spaces and $C \subset X$, $B \subset Z$ are arbitrary sets.

DEFINITION 5. The *generalized constraint condition* on g is said to hold at $\bar{x}$ if there is some closed convex cone G such that $G \cap K \subset T(A, \bar{x})$, where $K = \{h | g'(\bar{x})(h) \in P(B, g(\bar{x}))\}$.

(This is necessarily slightly stronger than Zlobec's condition [9] in which $P(A, \bar{x})$ replaces $T(A, \bar{x})$.) As before, A denotes $g^{-1}(B) \cap C$.

DEFINITION 6 [4]. Suppose K and G satisfy the constraint condition. $H(G)$ is said to hold when

(a) $K^+ + G^+$ is closed,

(b) $H = \{u^+ \cdot g'(\bar{x}) | u^+ \in P^+(B, g(\bar{x}))\}$ is closed, (in the weak* topology).

$H(G)$ is satisfied in particular when K, G, B are polyhedrally convex in finite dimensions. The author in his thesis has given fairly general conditions for $H(G)$ to hold.

THEOREM 4. *Suppose $\bar{x}$ is a (local) proper efficient point for* (P) *and that G satisfies the generalized constraint qualification with $H(G)$ holding. Suppose either $(S^+)^0 \neq \emptyset$ or that S is pointed and has a compact base. There is some $s^+ \in S^+$ with $s^+(s) > 0$ if $s \in S/\{0\}$, and some $u^+ \in P^+(B, g(\bar{x}))$ such that*

$$s^+ f'(\bar{x}) - u^+ g'(\bar{x}) \in -G^+.$$

Proof. By hypothesis, $S \cap \bar{T}(f(A) - S, f(\bar{x})) = 0$. It is an elementary property of tangent cones that,

$$f'(\bar{x})(\bar{T}(A, \bar{x})) \subset \bar{T}(f(A), f(\bar{x})). \tag{7}$$

Combining these two containments with $K \cap G \subset \bar{T}(A, \bar{x})$, one sees that $\text{cl}\,(f'(\bar{x})(K \cap G)) \cap S = 0$.

(This last containment is essentially Kuhn and Tucker's notion of properness if one takes $S = R^{n+}$). All the hypotheses of Proposition 2 or 3 are met with $N = \text{cl}\,(f'(\bar{x})(K \cap G))$. There is some s^+ satisfying (3) with $s^+ f'(\bar{x})(h) \leqq 0 \; \forall h \in K \cap G$. This means

$$s^+ f'(\bar{x}) \in -(K \cap G)^+ = -(K^+ + G^+) \tag{8}$$

(using $H(G)$ and property (iii) of Definition 4). A straightforward separation argument shows that $\bar{H} = K^+$. This combined with $H(G)$ and (8) yields

$$s^+ f'(\bar{x}) + u^+ g'(\bar{x}) \in -G^+, \tag{9}$$

where $s^+(s) > 0$ if $s \in S/\{0\}$ and $u^+ \in P^+(B, g(\bar{x}))$. □

In the standard finite-dimensional programming problem, C, B are coordinate cones and the Kuhn–Tucker constraint condition implies that $K \cap P(C, \bar{x}) \subset T(A, \bar{x})$. This means that Theorem 4 includes the Pareto maximization of any such program with respect to any pointed cone in R^n. Thus one sees that Geffrion's first order necessary condition is subsumed by Theorem 4.

As in the case of real-valued objective functions, weak sufficiency conditions can be described for (P) using the theory developed by Guignard [2].

In another direction if one does not require $H(G)$ to hold, one still has

$$s^+ f'(\bar{x}) \in \operatorname{cl}(\bar{H} + G^+),$$

which is much like Zlobec's asymptotic results in [9].

7. Conclusion. The paper provides a tangent cone definition of proper efficiency which coincides with Geffrion's for concave programs and coordinate orderings and which enables one to develop the theory in a much more general framework. It seems possible that some requirement of properness could be fruitfully imposed on various other notions of maximization allowing one to characterize various classes of extreme points in tangent cone terms. Using compact derivatives [10] one can extend the results to arbitrary convex spaces.

Acknowledgments. This paper was partially prepared while the author was working on his D. Phil under Dr. M. A. H. Dempster of Balliol College, Oxford. Without his continued interest and direction it would not have been written.

REFERENCES

[1] A. M. GEFFRION, *Proper efficiency and the theory of vector maximization*, J. Math. Anal. Appl., 22 (1968), pp. 618–630.

[2] M. GUIGNARD, *Generalized Kuhn–Tucker conditions for mathematical programming problems in Banach space*, this Journal, 7 (1969), pp. 232–241.

[3] V. L. KLEE, *Separation properties of convex cones*, Proc. Amer. Math. Soc., 6 (1955), pp. 313–318.

[4] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proc. Second Berkeley Symp. on Mathematical Statistics and Probability, University of California Press, Berkeley, Calif., 1950, pp. 481–492.

[5] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[6] A. PERESSINI, *Ordered Topological Vector Spaces*, Harper and Row, New York, 1967.

[7] K. RITTER, *Optimization in linear spaces II*, Math. Ann., 183 (1969), pp. 169–180.

[8] P. P. VARAIYA, *Nonlinear programming in Banach spaces*, SIAM J. Appl. Math., 19 (1967), pp. 239–244.

[9] S. ZLOBEC, *Asymptotic Kuhn–Tucker conditions for mathematical programming in a Banach space*, this Journal, 9 (1970), pp. 505–512.

[10] S. ZLOBEC AND H. MASSAM, *Various definitions of the derivative in mathematical programming*, presented at the VIII Internat. Symp. on Mathematical Programming, Stanford Univ., Palo Alto, Calif., 1973.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

A CONVERGENT SCHEME FOR BOUNDARY CONTROL OF THE HEAT EQUATION*

William C. Chewning died early in 1975. This paper, a result of a felicitous though brief collaboration, is dedicated to his memory and to the mathematics which, but for this tragic accident, he should have lived to create. (T.I.S.)

WILLIAM C. CHEWNING† AND THOMAS I. SEIDMAN‡

Abstract. It is desired to steer the solution of the heat equation in $\mathscr{D} \subset \mathbb{R}^n$ from a given initial state $u_0 = u(0, \cdot)$ exactly to a given terminal state $u_T = u(T, \cdot)$ by controlling the Dirichlet data φ. An algorithm is presented which, when such control is possible, provides a sequence converging in L_2 to the optimal control.

1. Introduction. Let $\mathscr{D} \subset \mathbb{R}^n$ be a bounded domain with piecewise smooth boundary $\mathscr{B}$. We consider the equation

$$u_t = \Delta u, \quad u = u(t, x) \quad \text{for } 0 < t < T, \quad x \in \mathscr{D}, \tag{1}$$

$$u(0, x) = u_0(x), \qquad x \in \mathscr{D}, \tag{2}$$

$$u(t, x) = \varphi(t, x), \qquad 0 < t < T, \quad x \in \mathscr{B}, \tag{3}$$

with $u_0 \in L_2(\mathscr{D})$ and $\varphi \in L_2(\mathscr{S})$ where $\mathscr{S} = (0, T) \times \mathscr{B}$. We may wish to further restrict φ to vanish on $\mathscr{S}_1 = (0, T) \times \mathscr{B}_1 (\mathscr{B}_1 \subset \mathscr{B})$ so $\varphi \in L_2(\mathscr{S}_0) \subset L_2(\mathscr{S})$ with $\mathscr{B}_0 = \mathscr{B} \backslash \mathscr{B}_1$ relatively open in $\mathscr{B}$. The *terminal state* $u(T, \cdot) \in L_2(\mathscr{D})$ depends continuously—indeed, compactly—on the *initial state* u_0 and *boundary data* φ. If a terminal state $u_T \in L_2(\mathscr{D})$ is specified:

$$u(T, x) = u_T(x), \qquad x \in \mathscr{D}, \tag{4}$$

then, to the extent that the data φ is at our disposal, we may view (1)–(4) as a control problem with φ as the control, carrying the initial state u_0 to the specified terminal state u_T subject to (1).

In practical applications one would like to approximate numerically a suitable control $\varphi \in L_2(\mathscr{S}_0)$, given u_0 and u_T. There are two complications: (a) there does not always exist such a control for arbitrary $(u_0, u_T) \in L_2(\mathscr{D}) \times L_2(\mathscr{D})$ and (b) when such a control exists it is not unique. These are handled as follows: (a) we *assume* that the specified pair (u_0, u_T) is such that a suitable φ exists, and (b) we seek the *optimal control* φ_* having *minimum norm* in $L_2(\mathscr{S}_0)$. It is known (see [2], [4]) that for $u_T = 0$ (or u_T in the range of the map: $u_0 \mapsto u(T, \cdot)$ for $\varphi = 0$) there exists a control $\varphi \in L_2(\mathscr{S})$ (in this context called a *null-control*) for each $u_0 \in L_2(\mathscr{D})$ with the optimal control φ_* depending continuously on u_0. In general, the set of controls for given (u_0, u_T), when nonempty, is a translate of a closed subspace (the nullspace of the map: $\varphi \mapsto u(T, \cdot)$ for $u_0 = 0$) so φ_* is

* Received by the editors March 12, 1975, and in revised form April 8, 1976.

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

‡ Department of Mathematics, University of Maryland, Baltimore County, Baltimore, Maryland 21228.

well determined. Under the assumption that u_T is reachable from u_0, we construct a sequence $\{\varphi_k\}$ of approximate controls converging to φ_* in $L_2(\mathscr{S}_0)$.

2. A useful identity. We follow the discussion in [3] in considering, in conjunction with (1)–(4), the "dual problem"

$$-v_t = \Delta v, \quad v = v(t,x) \quad \text{for } 0<t<T, \quad x \in \mathscr{D}, \tag{5}$$

$$v(T,x) = v_T(x), \qquad x \in \mathscr{D}, \tag{6}$$

$$v(t,x) = 0, \qquad 0<t<T, \quad x \in \mathscr{B}, \tag{7}$$

with $v_T \in L_2(D)$. We let

$$\begin{aligned} v_0(x) &= v(0,x), && x \in \mathscr{D}, \\ \psi(t,x) &= \frac{\partial v}{\partial \nu}, && 0<t<T, \quad x \in \mathscr{B}_0. \end{aligned} \tag{8}$$

Integrating $(uv)_t = uv_t + u_t v = v\Delta u - u\Delta v$ over $(0, T)\times\mathscr{D}$ gives

$$\int_{\mathscr{D}} [u_T v_T - u_0 v_0] = \int_0^T \int_{\mathscr{D}} [v\Delta u - u\Delta v] = -\int_0^T \int_{\mathscr{B}} \varphi\psi$$

on applying Green's identity and noting (7) and that $\varphi = 0$ for $x \in \mathscr{B}\backslash\mathscr{B}_0$. Thus,

$$\langle u_T, v_T\rangle_{\mathscr{D}} = \langle u_0, v_0\rangle_{\mathscr{D}} - \langle \varphi, \psi\rangle_{\mathscr{S}} \tag{9}$$

with the inner products in $L_2(\mathscr{D})$ and $L_2(\mathscr{S})$ and with φ, u_0, u_T related by (1)–(4) and v_T, v_0, ψ related by (5)–(8). For a set $\mathscr{V}$ in $L_2(\mathscr{D})$ let $\theta(\mathscr{V})$ be the set of triplets (v_T, v_0, ψ) related by (5)–(8) with $v_T \in \mathscr{V}$ and let $\mathbf{M}(\mathscr{V}) \subset L_2(\mathscr{S}_0)$ be the span of $\{\psi : (v_T, v_0, \psi) \in \theta(\mathscr{V})\}$. Let $\mathbf{M}$ be the closure of $\mathbf{M}(L_2(\mathscr{D}))$.

THEOREM 1. *Let $\mathscr{V}$ be any total set in $L_2(\mathscr{D})$ (i.e., $\langle u, v\rangle_{\mathscr{D}} = 0$ for all $v \in \mathscr{V}$ implies $u = 0$). Then*:

(a) *given u_0, u_T, φ is a control for* (1)–(4) *if and only if it satisfies* (9) *for every $(v_T, v_0, \psi) \in \theta(\mathscr{V})$,*

(b) *if any such control exists, there is a unique one in* $\mathbf{M}$ *and that is the optimal control φ_*.*

Proof. The proof, above, of (9) actually gives, from (1)–(3), (5)–(8), that

$$\langle u(T,\cdot), v_T\rangle_{\mathscr{D}} = \langle u_0, v_0\rangle_{\mathscr{D}} - \langle \varphi, \psi\rangle_{\mathscr{S}_0}$$

so if φ satisfies (9) as stated one has, subtracting, $\langle u(T,\cdot) - u_T, v_T\rangle_{\mathscr{D}} = 0$ for all $v_T \in \mathscr{V}$ which implies (4). It follows that controls are entirely characterized by their action as linear functionals on $\mathbf{M}$. The Riesz representation theorem asserts that any such functional can be obtained as the inner product with a unique element of $\mathbf{M}$ and this choice clearly minimizes the norm of the control; the nullspace of the map: $\varphi \mapsto u(T,\cdot)$ for $u_0 = 0$ is just the orthocomplement in $L_2(\mathscr{D})$ of $\mathscr{M}$. □

3. The algorithm. Assume, here, that $u_0, u_T \in L_2(\mathscr{D})$ are such that boundary controls exist in $L_2(\mathscr{S}_0)$. For any $\varphi \in L_2(\mathscr{S}_0)$, define the *residual* $r[\varphi]$ to be $[u_T - u(T,\cdot)]$ where u is given by (1)–(3). Let $\mathscr{V} = \{v^1, \cdots\}$ be any total sequence

in $L_2(\mathscr{D})$ and let $\{\psi^1, \cdots, \}$ be the associated (by (5)–(8)) elements of $\mathbf{M}(\mathscr{V})$; set $\mathscr{V}_k = \{v^1, \cdots, v^k\}$ and $\mathbf{M}_k = \mathbf{M}(\mathscr{V}_k) = \text{sp}\{\psi^1, \cdots, \psi^k\}$ in $\mathbf{M}$.

THEOREM 2. *For each k there exists a unique $\varphi_k \in L_2(\mathscr{S}_0)$ which is the element of minimum norm such that $r[\varphi]$ is orthogonal to* sp $\mathscr{V}_k$; *this element φ_k is in $\mathbf{M}_k$. Finally, φ_k is the unique solution in $\mathbf{M}_k$ of the finite linear system*

$$\langle r[\varphi], v^j\rangle_{\mathscr{D}} = 0, \qquad j = 1, \cdots, k. \tag{10}$$

Proof. For $r[\varphi]$ to be orthogonal to $\mathscr{V}_k$ is equivalent to (10) which, by (9), is equivalent to the system

$$\langle \varphi, \psi^j\rangle_{\mathscr{S}_0} = c_j, \qquad j = 1, \cdots, k, \tag{11}$$

where, defining v_0^j by (5)–(8) with $v_T = v^j$, one sets

$$c_j = \langle u_0, v_0^j\rangle_{\mathscr{D}} - \langle u_T, v^j\rangle_{\mathscr{D}} \tag{12}$$

for each j. Any control φ giving (1)–(4) has $r[\varphi] = 0$ so, as it has been assumed that such controls exist, (11) is consistent; the set of solutions of (10) or (11) is clearly closed and convex so a unique solution φ_k exists having minimum norm. From the form of (11), whether an element satisfies (10) or (11) depends only on its action as a linear functional on M_k. Thus, by the same Riesz theorem argument used for Theorem 1, the minimum norm solution φ_k is in $\mathbf{M}_k$ and is the unique solution in $\mathbf{M}_k$ of (10) or, equivalently, of (11). □

COROLLARY. *The element φ_k is uniquely determined by*

$$\varphi_k = x_{1,k}\psi^1 + \cdots + x_{k,k}\psi^k, \tag{13}$$

where the $x_j = x_{j,k}$ are obtained by solving

$$\sum_{j=1}^{k} g_{i,j}x_j = c_i, \qquad i = 1, \cdots, k, \tag{14}$$

with $\{c_i\}$ given by (12) *and $((g_{i,j}))$ the Gramian of $(\psi^1, \cdots, \psi^k)$, i.e.,*

$$g_{i,j} = \langle \psi^i, \psi^i\rangle_{\mathscr{S}_0}, \qquad i, j = 1, \cdots, k. \tag{15}$$

Proof. As $\varphi_k \in \mathbf{M}_k$ one has (13); substitute this into (11) to get (14), (15). Observe that (14) need not have a unique solution unless assumptions are made guaranteeing the linear independence of $(\psi^1, \cdots, \psi^k)$ but, as the number of independent equations is obviously equal to dim $\mathbf{M}_k$, the element φ_k given by (13), (14) is nevertheless uniquely determined. □

Computationally, one would obtain the $\{\psi^i\}$ and $\{v_0^j\}$ using (5)–(8) and then the $\{c_j\}$ and the matrix entries $\{g_{i,j}\}$ using (12) and (15), after which one would solve (14) and use (13) to define the approximate control φ_k. The computation of $\{\psi^i\}$, $\{v_0^j\}$ involves numerical solution of the heat equation in $\mathscr{D}$ (unless the $\{v^i\}$ can be taken to be eigenfunctions of Δ in $\mathscr{D}$ with (7)) which can be done to arbitrarily

good accuracy. The computation of the $\{c_j\}$, $\{g_{i,j}\}$ involves numerical integration over $\mathscr{D}$ and $\mathscr{S}_0$, respectively, which can also be done to arbitrary accuracy. Thus, (13), (14) can be used to obtain φ_k to arbitrarily good accuracy and *the algorithm is computationally feasible.* A preliminary (approximate) orthonormalization of the $\{\psi^j\}$ ("reflected back" to corresponding combinations of the $\{v^j\}$ for recomputation) will guarantee the well-conditioning of (13), (14) although this may make accurate computation of the $\{c_j\}$ more difficult if the new $\{v^j\}$ are large.

THEOREM 3. *The sequence of approximate controls* $\{\varphi_k\}$ *given in Theorem* 2 *converges in* $L_2(\mathscr{S}_0)$ *to the optimal control* φ_*.

Proof. Clearly $\|\varphi_1\| \leqq \|\varphi_2\| \leqq \cdots \leqq \|\varphi_*\|$ since these are defined by minimization with more and more constraints. The weak sequential precompactness of bounded sets in Hilbert space implies the existence of weakly convergent subsequences of $\{\varphi_k\}$, i.e., $\varphi_{k(i)} \rightharpoonup \varphi_{**}$. For each fixed j, φ_{**} satisfies the jth equation of (11) since each $\varphi_{k(i)}$ does once $k(i) \geqq j$. Thus, by Theorem 1, φ_{**} is a suitable control (*Remark*: *This shows that* $\|\varphi_k\| \to \infty$ *if no control giving* (1)–(4) *exists.*) Further, as $\|\varphi_{k(i)}\| \leqq \|\varphi_*\|$ one has $\|\varphi_{**}\| \leqq \|\varphi_*\|$ whence, by the definition of φ_*, one has $\varphi_{**} = \varphi_*$. As this is true for the limit of *every* weakly convergent subsequence, we must have $\varphi_k \rightharpoonup \varphi_*$. Finally, we note that $\|\varphi_*\| = \|w - \lim \varphi_k\| \geqq \lim \|\varphi_k\|$ implies (cf. [9]) strong convergence in $L_2(\mathscr{S}_0)$: $\varphi_k \to \varphi_*$. □

Instead of considering control of the heat equation (1) by Dirichlet conditions (3), it would be possible to pose corresponding problems for the equation

$$u_t = \mathbf{L}u, \quad u = u(t, x) \quad \text{for } 0 < t < T, \quad x \in \mathscr{D}, \tag{16}$$

where, e.g., $\mathbf{L}$ is a second order elliptic operator in divergence form:

$$\mathbf{L}u = \nabla \cdot p\nabla u - qu \tag{17}$$

with p, q given (smooth) functions on $[0, T] \times \bar{\mathscr{D}}$ ($p > 0$) and for boundary conditions

$$\alpha u + \beta \frac{\partial u}{\partial \nu} = \varphi, \qquad 0 < t < T, \quad x \in \mathscr{B}, \tag{18}$$

where α, β are given (smooth) functions on $\mathscr{S}$ ($\alpha^2 + \beta^2 = 1$).

We replace (5) and (7), respectively, by the equation

$$-v_t = \mathbf{L}v, \quad v = v(t, x) \quad \text{for } 0 < t < T, \quad x \in \mathscr{D}, \tag{19}$$

and the boundary conditions

$$\alpha v + \beta \frac{\partial v}{\partial \nu} = 0, \qquad 0 < t < T, \quad x \in \mathscr{B}. \tag{20}$$

We replace the second equation of (8) by

$$\psi(t, x) = \alpha \frac{\partial v}{\partial \nu} - \beta v, \qquad 0 < t < T, \quad x \in \mathscr{B}_0, \tag{21}$$

It is easily seen that the basic identity (9) continues to hold if the inner product over $\mathscr{S}$ (or over $\mathscr{S}_0$) is defined with p used as a weight function ($\langle \varphi, \psi \rangle_{\mathscr{S}} = \int p\varphi\bar{\psi}$).

Then the algorithm and arguments used for (1)–(4) in Theorems 1, 2, 3 can be used for the more general problem with no essential changes.

4. Numerical results. Some computational experiments have been performed. While certainly inadequate to provide a real "feel" for the usefulness of the algorithms for practical computation they nevertheless seem worth presenting.

Consider, first, a problem involving the one-dimensional heat equation:

$$
\begin{aligned}
&u_t = u_{xx}, && 0<t<0.4, \quad 0<x<1,\\
&u(t,0)=0, \quad u(t,1)=\varphi(t), && 0<t<0.4,\\
&u(0,x)=\sin \pi x, && 0\leqq x\leqq 1.
\end{aligned} \tag{22}
$$

We take $\mathcal{V}_k = (\sin j\pi x : j = 1, \cdots, k)$ and have Table 1 (in which φ_k denotes the computed control).

TABLE 1

k	4	6	8	10	12	14	16	18
$\|\varphi_k\|^2$	0.04106	0.05389	0.06227	0.06811	0.07240	0.07568	0.07823	0.07896
$\|\varphi_k-\varphi_{k-2}\|^2$	—	0.01283	0.00838	0.00584	0.00429	0.00328	0.00254	0.00075

As anticipated from the proof of Theorem 3, $\|\varphi_k\|$ increases monotonically with k. Of greater interest is the bottom row of the table which gives *some* concrete indication of the rate of convergence. Observe that the table strikingly confirms the orthogonality of the increments $(\varphi_k - \varphi_{k-2})$; see § 5, below.

Since the computed approximations to the optimal control will not, in general, be (suboptimal) exact controls, an experiment was performed to see how effective the "steering" of a computed control might be. Since diffusion equations have a strong smoothing effect even without purposive control, the effectiveness of the computed null-control was measured by comparing the resulting terminal state u_T^* with the corresponding terminal state u_T^0 of the uncontrolled solution ($\varphi = 0$). This experiment was performed for a problem involving a variable coefficient diffusion equation:

$$
\begin{aligned}
&u_t = (e^{-x/5}u_x)_x + \tfrac{1}{2}(1+x)u, && 0<t<0.3, \quad 0<x<1,\\
&u(t,0)=0, \quad u(t,1)=\varphi(t), && 0<t<0.3,\\
&u(0,x)=100x(1-x), && 0\leqq x\leqq 1.
\end{aligned} \tag{23}
$$

Only the approximation φ_4 was computed with $\mathcal{V}_4 = (\sin \pi x, \sin 2\pi x, \sin 3\pi x, \sin 4\pi x)$. See Table 2. We have $\|u_T^*\|^2/\|u_T^0\|^2 = 0.33$.

TABLE 2

x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$u_T^0(x)$	0.0	0.640	1.236	1.730	2.071	2.220	2.153	1.866	1.379	0.736	0.0
$u_T^*(x)$	0.0	0.020	0.010	−0.036	−0.090	−0.075	0.070	0.218	−0.226	−1.598	−1.605

It is not surprising that some roughness is introduced near the controlled boundary as the theory involves L_2 rather than pointwise convergence. Indeed, theory predicts that some (optimal) controls will be quite wild as $t \to T$, which would certainly lead us to expect this sort of roughness. If desired, this can be avoided—at the cost of increasing the L_2 norm of the control—either by seeking a control which is optimal with respect to a stronger norm (see § 6, below) or by extending the problem to $x \in (0, a)$ with $a > 1$ (extending the initial data) and after solving the extended problem with control at $x = a$, using the values of that solution at $x = 1$ as a suboptimal control for the original problem. This last "trick" also can be used to compute controls for more general (even time-dependent) boundary conditions.

5. Remarks. The discussion so far has followed the manuscript prepared and submitted before William Chewning's death. The resulting substantial delays mean that this final version is being prepared about a year later and, while it seems appropriate to have presented the original material in somewhat of its original form, we proceed now to introduce the perspective of more recent work by the second author [5], [6], [7].

If one looks carefully at the algorithm and its proof it becomes clear that their intrinsic logic makes negligible use of properties specific to the control problem to which it is applied but, rather, can be viewed abstractly as a general computational approach to a wide variety of problems. The proof above of Theorem 3 does not even use the linearity of the problem or the Hilbert space setting and has been abstracted in [6] to apply to nonlinear problems in the setting of a uniformly convex Banach space. The computational algorithm of Theorem 2 and its corollary *does* use these structures and, in this case, a far less sophisticated convergence proof (giving additional useful information) has been presented in [5] and some material on convergence rates has been obtained in [7]. We now review § 3 from the more abstract perspective of [5], [7].

Let $\mathbf{A}$: $L_2(\mathscr{S}) \to L_2(\mathscr{D})$ be the (continuous—indeed, compact) mapping: $\varphi \mapsto u(T, \cdot)$ defined by (1)–(3) with $u_0 = 0$. Then (9) shows that $\mathbf{A}^*$: $L_2(\mathscr{D}) \to L_2(\mathscr{S})$ is just the mapping: $v(T, \cdot) \mapsto -\psi$ defined by (5)–(8). If we also define $\mathbf{F}$: $L_2(\mathscr{D}) \to L_2(\mathscr{D})$ by $\mathbf{F}u_0 = u(T, \cdot)$ given by (1)–(3) with $\varphi = 0$, then the control problem consists of solving

$$\mathbf{A}\varphi = \mathbf{b} \tag{24}$$

for φ with $\mathbf{b} = u_T - \mathbf{F}u_0$. For $u_T = 0$, solvability of (24) for every u_0 implies continuity of the null-control map $\mathbf{C}$: $L_2(\mathscr{D}) \to L_2(\mathscr{S})$, taking u_0 to the corresponding optimal control φ_*.

Since $\mathscr{R}(\mathbf{A})$ is (dense but) not closed in $L_2(\mathscr{D})$, $\mathbf{A}$ does not have a pseudoinverse (cf. [1]) but each $\mathbf{H}_k\mathbf{A}$ does, where $\mathbf{H}_k$: $L_2(\mathscr{D}) \to \mathbb{R}^k$ is defined by setting

$$(\mathbf{H}_k y)_j = \langle y, v^j \rangle_{\mathscr{D}} = \int_{\mathscr{D}} y v^j$$

and (12)–(15) is easily seen to define φ_k by letting this pseudoinverse act on $\mathbf{H}_k\mathbf{b}$.

Theorem 1 of [5] asserts that $\mathbf{M}_k = \mathbf{A}^* \mathcal{V}_k = \mathcal{R}((\mathbf{H}_k \mathbf{A})^*)$ and $\mathbf{S}_k = \{x: \mathbf{H}_k \mathbf{A} x = \mathbf{H}_k \mathbf{b}\}$ meet orthogonally with

$$\mathbf{S}_k \cap \mathbf{M}_k = \{\varphi_k\}, \quad \varphi_k = \text{projection of } \varphi_* \text{ on } \mathbf{M}_k, \tag{25}$$

so that $\varphi_k \to \varphi_*$ since $\operatorname{sp}\{\mathbf{M}_k\}$ is dense in $\mathcal{N}(\mathbf{A})^\perp = \overline{\mathbf{R}(\mathbf{A}^*)} = \mathbf{M}$ by the assumed totality of $\mathcal{V} = \{v^1, \cdots\}$. The fact that the approximants are obtained by orthogonal projection shows that *each φ_k is the best possible approximation (nearest point) to φ_* in $\mathbf{M}_k$* and that $(\varphi_k - \varphi_j) \perp \varphi_j$ for $k > j$ so $\|\varphi_k\|^2 = \|\varphi_j\|^2 + \|\varphi_k - \varphi_j\|^2$ (compare the start of the proof in § 3 of Theorem 3 and the first table in § 4).

6. Rate of convergence. Since *every* $\varphi \in \mathbf{M}$ is the optimal control associated with *some* specification of u_T (with u_0 fixed), the convergence can be arbitrarily slow no matter how $\{v^1, \cdots\}$ is chosen. On the other hand, we see [7] that restriction by a regularity condition (particularly on restricting consideration to certain null-control problems) permits establishment of a convergence rate.

It is easy to see that typical regularity conditions can be formulated as requiring, for problems (24), that the (minimum norm) solution φ_* be in the range of a compact embedding map $\mathbf{E}: \mathbf{X} \to L_2(\mathcal{S})$ for some Hilbert space $\mathbf{X}$. With $x_* \in \mathbf{X}$ so $\varphi_* = \mathbf{E} x_*$ one has, for the computed approximant φ_k, the convergence rate

$$\|\varphi_k - \varphi_*\| \leqq \rho_k \|x_*\| \tag{26}$$

with

$$\rho_k = \sup\{\|\mathbf{E}x - \mathbf{M}_k\| : \|x\| = 1\} \to 0. \tag{27}$$

By Theorem 2 of [7],

$$\varepsilon_{k+1} \leqq \rho_k \leqq \inf_K \left[\varepsilon_{K+1}^2 + \sum_1^K \|\mathbf{E} e_j - \mathbf{M}_k\|^2 \right]^{1/2}, \tag{28}$$

where $\{(e_j, \varepsilon_j^2)\}$ are the eigenpairs of $\mathbf{E}^*\mathbf{E}$ taken so $\{e_j\}$ is an orthogonal basis and $\varepsilon_1 \geqq \varepsilon_2 \geqq \cdots > 0$.

In applying this notion to the control problem we observe that if we restrict our attention to autonomous null-control problems for which it is known that *a null-control exists for every u_0 in $L_2(\mathcal{D})$ and every $T > 0$*—see [2], [4], [8]—then we may obtain a rate using $\mathbf{X} = L_2(\mathcal{D})$ and the control map. For null-control, the problem has the form: $\mathbf{A}\varphi = \mathbf{F} u_0$ and controllability means $\mathcal{R}(\mathbf{F}) \subset \mathcal{R}(\mathbf{A})$ so the control map $\mathbf{C}$: $u_0 \mapsto \varphi =$ minimum norm control is linear and continuous. Note that we may construct a null-control $\tilde{\varphi}$ for a given u_0 by taking $\tilde{\varphi}$ to be 0 for $0 < t < T'$ and then controlling $u(T', \cdot) = \mathbf{F}' u_0$ to $\mathbf{0}$ so, on $[T', T]$, $\tilde{\varphi}$ is $\mathbf{C}'\mathbf{F}' u_0$ where $\mathbf{F}'$ is the solution operator for (16)–(18) on $[0, T']$ with $\mathbf{0}$ control and $\mathbf{C}'$ is the control map for $[T', T]$. Now, with a minor abuse of notation, the optimal null-control $\varphi_* = \mathbf{C} u_0$ is just the projection $\mathbf{P}\mathbf{C}'\mathbf{F}' u_0$ of $\tilde{\varphi}$ on $\mathbf{M}$ ($\mathbf{P}$ the orthoprojection onto $\mathbf{M}$ in $L_2(\mathcal{S})$). Now $\mathbf{P}$, $\mathbf{C}'$ are bounded and $\mathbf{F}'$ is compact so $\mathbf{C}$ is compact and we may let $\mathbf{C}$ play the role of the "regularity map" above.

Letting $\{\varepsilon_k^2\}$ be the eigenvalues of $\mathbf{C}^*\mathbf{C}$ in decreasing order with multiplicities, a standard variational characterization of the eigenvalues gives

$$\begin{aligned}\varepsilon_k^2 &= \inf\{\sup\{\langle u, \mathbf{C}^*\mathbf{C}u\rangle: \|u\|=1,\ u\perp\mathbf{U}\}: \dim\mathbf{U}=k\}\\ &= \inf\{\|\mathbf{C}|_{\mathbf{U}^\perp}\|^2: \dim\mathbf{U}=k\}\\ &\leqq \|\mathbf{PC}'\|^2 \inf\{\|\{\mathbf{F}'|_{\mathbf{U}^\perp}\|^2: \dim\mathbf{U}=k\}\\ &= \|\mathbf{PC}'\|^2 \exp[-2\lambda_k T'],\end{aligned} \tag{29}$$

where the $\{-\lambda_k\}$ are the eigenvalues of $\mathbf{L}$. (Note that with $\mathbf{L}$ as in (17) for bounded $\mathscr{D}\subset\mathbb{R}^n$ one has $\lambda_k \sim c_0 k^{2/n}$ asymptotically as $k\to\infty$.) Using (28), (29) in (26) gives the convergence rate:

$$\|\varphi_k - \varphi_*\| \leqq c\|u_0\| \exp[-T' k^{2/n}] \tag{30}$$

for any $T' < T$ and c depending on T', etc., but not on k or u_0. This *optimal* convergence rate will be attained if one could have $\mathbf{M}_k = \operatorname{sp}\{\mathbf{C}u_1, \cdots, \mathbf{C}u_k\}$ with u_j the jth eigenfunction of $\mathbf{C}^*\mathbf{C}$. Unfortunately, at present little is known about these eigenfunctions $\{u_j\}$ or about $\{\mathbf{C}u_j\}$ so that it is difficult either to choose the $\{v^j\}$ optimally (it is not known whether $\mathbf{C}u_j \in \mathscr{R}(\mathbf{A}^*)$ so optimal choice of v^j may not be possible) or to use the right-hand inequality of (28) to estimate the convergence rate for a given sequence $\{v^j\}$. Unfortunately, also, there is no known nontrivial ($u_0 \neq 0$) null-control problem for which an exact analytic solution is available as a test case.

We make one final remark. The discussion above has been in the context of $L_2(\mathscr{S})$ controls and $L_2(\mathscr{S})$ convergence. One obvious modification is to seek the optimal control in $H_0^m([0, T]\to L_2(\mathscr{B}))$ when such controls exist, proceeding to compute approximately either φ_* or $\partial^m\varphi_*/\partial t^m$ by a modified version of the algorithm above. This would provide a control which would be smoother as $t\to T-$, eliminating the "roughness" noted in the second example of § 4, although at the expense of increasing the $L_2(\mathscr{S})$ norm. We also note that for the autonomous null-control problems just discussed it can be shown [8] that the algorithm gives not only pointwise but $C^m([0, T']\to L_2(\mathscr{B}))$ convergence (with—at least—the same convergence rate) for any $T' < T$; further, the optimal control can be shown to depend continuously on $\mathbf{L}$ in the setting.

REFERENCES

[1] D. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[2] D. L. RUSSELL, *A unified boundary value controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., LII (1973), pp. 189–211.

[3] T. I. SEIDMAN, *Problems of boundary control and observation for diffusion processes*, Rep. MRR 73-10, Univ. of Maryland, Baltimore County, Baltimore, 1973.

[4] ———, *Observation and prediction for the heat equation. III*, J. Differential Equations, 20 (1976), pp. 18–27.

[5] ———, *Solution of singular equations. I: Linear problems in Hilbert space*, Pacific J. Math., 61 (1975), pp. 513–520.

[6] ———, *Solution of singular equations. II: Non-linear problems*, Rep. MRR 75-10, Univ. of Maryland, Baltimore County, Baltimore, 1975.
[7] ———, *Computational approaches to ill-posed problems: general considerations*, Proc. Conference on Information Science and Systems, Johns Hopkins Univ., Baltimore, 1976.
[8] ———, *Exact boundary controllability for autonomous wave and diffusion processes*, to appear.
[9] K. YOSIDA, *Functional Analysis*, Springer, Berlin, 1966.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

A NOTE ON AN ALGORITHM BY J. RISSANEN*

B. R. DYE†

Abstract. An adaption is made to a well-known algorithm for calculating the realization of a sequence of matrices in order to allow more than one calculation of the realization to be made. A mean realization can then be obtained. A particular class of sequences is defined and then shown to be the most suitable for the adapted algorithm to use.

1. Introduction. In [1] and [2] Rissanen and Kailath present an algorithm for calculating the realization of a sequence of $p \times q$ matrices, $A_0, A_1, A_2, \cdots$, where the realization is in the form of a triple of matrices $F(N), H(N), K(N)$ such that

$$A_i = H(N) \cdot F^i(N) \cdot K(N) \quad \text{for } i = 0, 1, \cdots \tag{1}$$

and the size of the matrices is minimal.

Throughout this paper we shall use the notation of [2] and we shall refer only to the case where $p = 1$, i.e., the sequence A_i is a sequence of row matrices.

2. The main theorem.

THEOREM 1. *Given that Rissanen's algorithm produces the triple, $F(N)$, $H(N)$, $K(N)$ of dimension m as the realization of the sequence of $1 \times q$ matrices, $A_0, A_1, A_2. \cdots$, then*

$$P^*(m) \cdot P^{-1}(m) = C_F, \quad \textit{the companion matrix of } F(N).$$

Before proving this theorem we shall first establish some lemmas.

DEFINITION 1. If F is a square matrix and the characteristic equation $c_F(\lambda)$ of F is given by

$$c_F(\lambda) = a_1 + a_2\lambda + \cdots + a_n\lambda^{n-1} + \lambda^n$$

where n is the dimension of F, then the companion matrix C_F of F is given by

$$C_F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_n \end{bmatrix}.$$

LEMMA 1. *Given the same conditions as those in the statement of Theorem 1,*

* Received by the editors December 11, 1975, and in revised form March 17, 1976.

† School of Information Sciences, Hatfield Polytechnic, Hatfield, Hertfordshire AL10 9AB, England.

we have

$$C_F \cdot \begin{bmatrix} A_i \\ A_{i+1} \\ \cdot \\ \cdot \\ \cdot \\ A_{i+m-1} \end{bmatrix} = \begin{bmatrix} A_{i+1} \\ A_{i+2} \\ \cdot \\ \cdot \\ \cdot \\ A_{i+m} \end{bmatrix} \quad \text{for } i = 0, 1, \cdots.$$

Proof. First we note that, by the Cayley-Hamilton theorem, $F(N)$ satisfies its own characteristic equation, which we may write in the form

$$F^m = -a_1 I - a_2 F - \cdots - a_m F^{m-1}. \tag{2}$$

To prove the lemma it is sufficient to show that

$$-a_1 A_i - a_2 A_{i+1} - \cdots - a_m A_{i+m-1} = A_{i+m} \quad \text{for } i = 0, 1, \cdots. \tag{3}$$

Using (1), we see that the left-hand side of (3) may be written as

$$\begin{aligned} &-a_1 H \cdot F^i \cdot K - a_2 H \cdot F^{i+1} \cdot K - \cdots - a_m H \cdot F^{i+m-1} \cdot K \\ &= H \cdot F^i \cdot [-a_1 I - a_2 F - \cdots - a_m F^{m-1}] \cdot K \\ &= H \cdot F^i \cdot F^m \cdot K \quad \text{(by (2))} \\ &= A_{i+m} \quad \text{(by (1))} \end{aligned}$$

which proves (3) and hence the lemma.

COROLLARY 1. *If by $A_i(j)$ we mean the j-th element in the $1 \times q$ matrix A_i, then an immediate corollary of Lemma 1 is*

$$C_F \cdot \begin{bmatrix} A_i(j) \\ A_{i+1}(j) \\ \cdot \\ \cdot \\ \cdot \\ A_{i+m-1}(j) \end{bmatrix} = \begin{bmatrix} A_{i+1}(j) \\ A_{i+2}(j) \\ \cdot \\ \cdot \\ \cdot \\ A_{i+m}(j) \end{bmatrix} \quad \text{for } i = 0, 1, \cdots \text{ and } j = 1, 2, \cdots, q.$$

Thus C_F acts on the individual rows and columns of $A(m, N)$ and disregards the block-matrix structure.

LEMMA 2. *In* [2] *the elements of $Q(m, N)$ are denoted by q_{ij} and $s(i)$ is the least integer such that $q_{i,s(i)} \neq 0$. The algorithm sets all $q_{k,s(i)} = 0$ for $k > i$. Given the conditions in Theorem 1 the last row of $Q(m, N)$ is zero. If we form the matrix $Q^*(m-1)$ from columns $s(1), s(2), \cdots, s(m-1)$ of $Q(m-1, N)$, then $Q^*(m-1)$ will be nonsingular.*

Proof. m will not be greater than N so the columns certainly exist. $Q^*(m-1)$

is trivially nonsingular since it has the form

$$Q^*(m-1)=\begin{bmatrix} q_{1,s(1)} & q_{1,s(2)} & q_{1,s(3)} & \cdots & q_{1,s(m-1)} \\ 0 & q_{2,s(2)} & q_{2,s(3)} & \cdots & q_{2,s(m-1)} \\ 0 & 0 & q_{3,s(3)} & \cdots & q_{3,s(m-1)} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & q_{m-1,s(m-1)} \end{bmatrix}.$$

Proof of Theorem 1. We denote the elements of $A(m, N)$ by a_{ij}. From the factorization $A(m, N)=P(m)\cdot Q(m, N)$, we obtain directly

$$P(m)\cdot Q^*(m-1)=\begin{bmatrix} a_{1,s(1)} & a_{1,s(2)} & \cdots & a_{1,s(m-1)} \\ a_{2,s(1)} & a_{2,s(2)} & \cdots & a_{2,s(m-1)} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m-1,s(1)} & a_{m-1,s(2)} & \cdots & a_{m-1,s(m-1)} \end{bmatrix} \tag{4}$$

and also

$$P^*(m)\cdot Q^*(m-1)=\begin{bmatrix} a_{2,s(1)} & a_{2,s(2)} & \cdots & a_{2,s(m-1)} \\ a_{3,s(1)} & a_{3,s(2)} & \cdots & a_{3,s(m-1)} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m,s(1)} & a_{m,s(2)} & \cdots & a_{m,s(m-1)} \end{bmatrix}. \tag{5}$$

From Corollary 1, we may deduce the equality

$$C_F\cdot\begin{bmatrix} a_{1,s(1)} & a_{1,s(2)} & \cdots & a_{1,s(m-1)} \\ a_{2,s(1)} & a_{2,s(2)} & \cdots & a_{2,s(m-1)} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m-1,s(1)} & a_{m-1,s(2)} & \cdots & a_{m-1,s(m-1)} \end{bmatrix} \tag{6}$$

$$=\begin{bmatrix} a_{2,s(1)} & a_{2,s(2)} & \cdots & a_{2,s(m-1)} \\ a_{3,s(1)} & a_{3,s(2)} & \cdots & a_{3,s(m-1)} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m,s(1)} & a_{m,s(2)} & \cdots & a_{m,s(m-1)} \end{bmatrix}.$$

From (4), (5) and (6) together we get

$$C_F\cdot P(m)\cdot Q^*(m-1)=P^*(m)\cdot Q^*(m-1).$$

Since $Q^*(m-1)$ is nonsingular we may "cancel" it, thus proving the theorem.

3. The application of the theorem.

DEFINITION 2. We adapt the algorithm given in [1] and [2] to produce the realization F, G, H, where

$$F = P^*(m) \cdot P^{-1}(m),$$

$$G = \begin{bmatrix} A_0 \\ A_1 \\ \cdot \\ \cdot \\ \cdot \\ A_{m-1} \end{bmatrix}, \qquad H = [1, 0 \quad \cdots \quad 0].$$

DEFINITION 3. Given the sequence of $1 \times q$ matrices $A_0, A_1, A_2, \cdots$ the realization of the sequence $A_i, A_{i+1}, A_{i+2} \cdots$ is denoted by F_i, G_i, H_i for each $i = 0, 1, 2 \cdots$ so that, in particular, F_0, G_0, H_0 is just our original F, G, H.

DEFINITION 4. A sequence of $1 \times q$ matrices A_0, A_1, $A_2 \cdots$ is called *homogeneous* if the realization of each sequence $A_i, A_{i+1}, A_{1+2} \cdots$ for $i = 0, 1, 2 \cdots$ is of the same dimension.

THEOREM 2. *Given the homogeneous sequence of* $1 \times q$ *matrices* $A_0, A_1, A_2 \cdots$ *we calculate the realizations* F_i, G_i, H_i *in the form of Definition* 2 *for each* $i = 0, 1, 2, \cdots$. *Then* $F_i = F_j$ *for all* $i, j = 0, 1, \cdots$.

Proof. Certainly each F_i is a companion matrix of the same size. Suppose $F_i \neq F_j$ for some $i < j$. Let the characteristic equation of F_i be

$$c_{F_i}(\lambda) = a_1 + a_2\lambda + \cdots + a_m\lambda^{m-1} + \lambda^m,$$

And let the characteristic equation of F_j be

$$c_{F_j}(\lambda) = b_1 + b_2\lambda + \cdots + b_m\lambda^{m-1} + \lambda^m.$$

Following the proof of Lemma 1 we shall get, for $k \geqq j$,

$$-a_1A_k - a_2A_{k+1} - \cdots - a_mA_{k+m-1} = A_{k+m} \tag{7}$$

and

$$-b_1A_k - b_2A_{k+1} - \cdots - b_mA_{k+m-1} = A_{k+m}. \tag{8}$$

Subtracting (7) from (8) we get:

$$(a_1 - b_1)A_k + (a_2 - b_2)A_{k+1} + \cdots + (a_m - b_m)A_{k+m-1} = 0$$
$$\text{for } k = j, j+1, j+2, \cdots.$$

This contradicts our assumption that the sequence is homogeneous unless every coefficient is zero. So $F_i = F_j$ for all i, j.

THEOREM 3. *Given the realization* F, G, H *of the sequence of* $1 \times q$ *matrices, a necessary and sufficient condition for the sequence to be homogeneous is that* F *has no zero eigenvalues.*

Proof. F is in companion matrix form (see Definitions 1 and 2). F has zero eigenvalues if and only if a_1 is zero. The proof of the theorem is in two parts.

Necessity. Suppose F has zero eigenvalues; then a_1 is zero, and suppose F is of dimension m. Then, by considering (3) in the proof of Lemma 1, we have

$$-a_2A_{i+1}-\cdots-a_mA_{i+m-1}=A_{i+m} \qquad \text{for } i=0,1,\cdots,$$

and this tells us that the dimension of the realization of $A_1, A_2, A_3\cdots$ is $m-1$ and so the sequence is not homogeneous.

Sufficiency. Suppose $A_0, A_1, A_2\cdots$ is nonhomogeneous. Then there exists a k such that

(9) the realization $F_{k-1}, G_{k-1}, H_{k-1}$ of the sequence $A_{k-1}, A_k\cdots$ is of dimension m, say

and

(10) the realization F_k, G_k, H_k of the sequence $A_k, A_{k+1}\cdots$ is of dimension $m-1$.

Suppose

$$F_{k-1}=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_m \end{bmatrix}$$

and

$$F_k=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -b_1 & -b_2 & -b_3 & \cdots & -b_{m-1} \end{bmatrix}.$$

Then (9) and (10) imply

$$-a_1A_{k-1+i}-a_2A_{k+i}-\cdots-a_mA_{k+m-2+1}=A_{k+m-1+i} \tag{11}$$

and

$$-b_1A_{k+i}-b_2A_{k+1+i}-\cdots-b_{m-1}A_{k+m-2+i}=A_{k+m-1+i} \qquad \text{for } i=0,1,\cdots. \tag{12}$$

Now, subtracting (12) from (11) gives us

$$-a_1A_{k-1+i}-(a_2-b_1)A_{k+i}-\cdots-(a_m-b_{m-1})A_{k+m-2+i}=0 \qquad \text{for } i=0,1,\cdots. \tag{13}$$

Unless every coefficient in (13) is zero, it is saying that the sequence $A_{k-1}, A_k, A_{k+1}.\cdots$ has a realization with dimension at most $m-1$ and this

contradicts (9). So every coefficient is zero, in particular a_1, and therefore F_{k-1} has zero eigenvalues and so has F_0 by Theorem 2. Thus sufficiency is shown.

COROLLARY 2. *It follows directly from Theorem* 3, *that, given any not necessarily homogeneous sequence of* $1 \times q$ *matrices, then the nonzero elements in the last row of each* F_i *will be identical. This comes directly from the fact that every coefficient in* (13) *is zero.*

Theorems 1, 2 and 3 show that given any homogeneous sequence of $1 \times q$ matrices, then we may make as many calculations as we like of the F matrix and each calculation uses a different part of the given sequence. For example, if F $(=F_0)$ is of dimension $m-1$ say, then we may calculate F_i for $i=0, 2m, 4m, 6m \cdots$ and each calculation will use a completely disjoint portion of the given sequence (since $2m$ terms are used each time). The condition of homogeneousness serves to exclude sequences which have terms in them completely independent of all the rest and which therefore have dimensions artificially large. We have thus shown that a small adaption to a well-known algorithm allows a more accurate realization to be found. Reference [3] contains a description of other algorithms superior to the standard algorithms, which is relevant to this note.

REFERENCES

[1] J. RISSANEN, *Recursive identification of linear systems*, this Journal, 9 (1971), pp. 420–430.

[2] J. RISSANEN AND T. KAILATH, *Partial realization of random systems*, Automatica, 8 (1972), pp. 389–396.

[3] L. S. DE JONG, *Numerical aspects of realization algorithms in linear systems theory*, thesis, Dept. of mathematics, Technological University, Eindhoven, the Netherlands, 1975.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

OPTIMAL IMPULSE CONTROL OF A DIFFUSION PROCESS WITH BOTH FIXED AND PROPORTIONAL COSTS OF CONTROL*

SCOTT F. RICHARD†

Abstract. This paper concerns the optimal control of a system where the state is modeled by a homogeneous diffusion process in R^1. Each time the system is controlled a fixed cost is incurred as well as a cost which is proportional to the magnitude of the control applied. In addition to the cost of control, there are holding or carrying costs incurred which are a function of the state of the system. Sufficient conditions are found to determine the optimal control in both an infinite horizon case with discounting and a finite horizon case. In both cases the optimal policy is one of "impulse" control originally introduced by Bensoussan and Lions [2] where the system is controlled only a finite number of times in any bounded time interval and the control requires an instantaneous finite change in the state variable. The issue of the existence of such controls is not addressed.

1. Introduction. This paper concerns the optimal control of a system where the state is modeled by a homogeneous diffusion process in R^1 and where there are both fixed and proportional costs incurred by controlling the process. In addition to the costs of controlling the process, there is assumed to be a holding cost which is a function of the state of the system. Sufficient conditions are found for a control policy to be optimal. The question of the existence of such a policy is not considered.

This paper is a modification[1] of Bensoussan and Lions [1], [2] who were the first to consider the finite horizon problem with fixed costs only. In their case Bensoussan and Lions find that the optimal control policy is one of "impulse control," where the control is used at a series of stopping times to instantaneously move the state of the system by a finite amount. This jump type of control is, of course, necessitated by the incursion of a fixed cost every time the control is used. Bensoussan and Lions restrict themselves to the case where all costs are bounded and, in particular, rule out the case where holding costs rise linearly with the state of the system and the costs of control rise in proportion to the magnitude of the control. In this paper these restrictions are removed and a general holding cost function and proportional control costs are allowed. Furthermore, we consider here both the infinite horizon discounted cost case and the finite horizon case. Lastly, Bensoussan and Lions restrict the control to be nonnegative, but that restriction is removed herein.

2. The infinite horizon with discounted costs model. Let w_t be a Wiener process in R^1 and $\mathcal{F}_t$ be the increasing family of σ-algebras generated by w_t. Let $0 \leqq \tau_1 \leqq \tau_2 \leqq \cdots \leqq \tau_i \leqq \cdots$ be an increasing sequence of stopping times adapted to $\mathcal{F}_t$, such that only a finite number will occur in a bounded interval a.s. Denote by

* Received by the editors May 30, 1975, and in revised form August 14, 1975.

† Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

[1] In one sense this paper is not a generalization: Bensoussan and Lions consider the case of a diffusion in R^n.

$\mathcal{F}_{\tau_i}$ the minimum σ-algebra of events prior to τ_i. To each τ_i assign a random variable ξ_i which is $\mathcal{F}_{\tau_i}$ measureable. Without loss of generality assume that $\tau_i \to \infty$ a.s. as $i \to \infty$.

Let $Y(t)$ be defined by the stochastic differential equations

$$(1) \qquad \begin{cases} dy(t) = \mu\, dt + \sigma\, dw_t, & \tau_i \leqq t < \tau_{i+1} \quad \forall i \geqq 0; \\ y(\tau_i) = y(\tau_i^-) + \xi_i, \\ y(0) = x, \end{cases}$$

where

$$(2) \qquad \tau_0 \equiv 0 \quad \text{and} \quad \tau_{i+1}^- = \tau_i \quad \text{if } \tau_{i+1} = \tau_i.$$

Let an impulse control be denoted by v:

$$(3) \qquad v = (\tau_1, \xi_1; \cdots; \tau_i, \xi_i; \cdots).$$

Let the holding cost function $H(x)$ be continuous and nonnegative. The total cost function associated with the policy v is

$$(4) \qquad J_x(v) = E\left\{ \sum_{i=1}^{\infty} e^{-\beta\tau_i} B(\xi_i) + \int_0^{\infty} e^{-\beta s} H(y(s))\, ds \right\},$$

where $B(\xi)$ is the cost of control ξ given by

$$B(\xi) = K_\xi + k_\xi |\xi|,$$

$$(5) \qquad K_\xi = \begin{cases} K^+ > 0 & \text{for } \xi \geqq 0, \\ K^- > 0 & \text{for } \xi < 0, \end{cases}$$

and

$$k_\xi = \begin{cases} k^+ > 0 & \text{for } \xi \geqq 0, \\ k^- > 0 & \text{for } \xi < 0. \end{cases}$$

We, of course, seek a control $\hat{v}$ such that $J_x(\hat{v}) = \inf_v J_x(v)$.

Suppose there exists a function $u(x)$ such that[2]

$$(6) \qquad u \geqq 0, \quad u'(x) \text{ is absolutely continuous and bounded} \quad \text{and} \quad u''(x) \in L^2(R),$$

$$(7) \qquad u(x) \leqq \inf [B(\xi) + u(x+\xi)] \equiv Qu(x), \forall x;$$

$$(8) \qquad \beta u(x) - \mu u'(x) - \tfrac{1}{2}\sigma^2 u''(x) \leqq H(x) \quad \text{a.e. } x;$$

and

$$(9) \qquad (\beta u(x) - \mu u'(x) - \tfrac{1}{2}\sigma^2 u''(x) - H(x))(u(x) - Qu(x)) = 0.$$

Then we may define the optimal policy as follows.

Let the continuation region be defined by

$$(10) \qquad C = \{x : u(x) < Qu(x)\},$$

[2] For a heuristic (and quite intuitive) development of these conditions see Bensoussan and Lions [2].

which is an open subset of R since it is easily shown[3] that $Qu(x)$ is continuous. Define the impulse control $\hat{v}$ associated with $u(x)$ as follows. Let

$$\begin{cases} d\hat{y} = \mu\, dt + \sigma\, dw_t, \\ \hat{y}(0) = x, \end{cases} \tag{11}$$

and let

$$\hat{\tau}_1 = \inf_{t \geqq 0} \{\hat{y}(t) \notin C\} \tag{12}$$

and

$$\hat{\xi}_1 = \eta(\hat{y}(\hat{\tau}_1^-)), \tag{13}$$

where $\eta(x)$ is a real-valued measurable function[4] chosen so that

$$B(\eta(x)) + u(x + \eta(x)) = Qu(x) \quad \forall x. \tag{14}$$

In general then to define $\hat{\tau}_i$ and $\hat{\xi}_i$ consider

$$\begin{cases} d\hat{y} = \mu\, dt + \sigma\, dw_t, & t \geqq \hat{\tau}_i, \\ \hat{y}(\hat{\tau}_i) = \hat{y}(\hat{\tau}_i^-) + \hat{\xi}_i, \end{cases} \tag{15}$$

and let

$$\hat{\tau}_{i+1} = \inf_{t \geqq \hat{\tau}_i} \{\hat{y}(t) \notin C\} \tag{16}$$

and

$$\hat{\xi}_{i+1} = \eta(\hat{y}(\hat{\tau}_{i+1}^-)). \tag{17}$$

THEOREM 1. *If there exists a solution to* (6)–(9), *then*

$$u(x) = J_x(\hat{v}) \leqq J_x(v) \quad \forall v \tag{18}$$

and $\hat{v}$ defines the optimal impulse control.

Remark 1. If we have the additional constraint that $\xi \in K$, K compact, or $\xi \geqq 0$, then that constraint is imposed on ξ in (7) and on η in (14) and Theorem 1 remains valid.

---

[3] For $\varepsilon > 0$ by the uniform Lipschitz condition

$$-M\varepsilon \leqq u(x + \xi + \varepsilon) - u(x + \xi) \leqq M\varepsilon \quad \forall \xi$$

so that

$$\begin{aligned} K_\xi + k_\xi|\xi| + u(x + \xi) - M\varepsilon &\leqq K_\xi + k_\xi|\xi| + u(x + \xi + \varepsilon) \\ &\leqq K_\xi + k_\xi|\xi| + u(x + \xi) + M\varepsilon. \end{aligned}$$

Taking the infinum yields

$$Qu(x) - M\varepsilon \leqq Qu(x + \varepsilon) \leqq Qu(x) + M\varepsilon$$

so that

$$|Qu(x + \varepsilon) - Qu(x)| \leqq M\varepsilon.$$

[4] It is easily shown that for each x, $\eta(x)$ is chosen from a compact set. See Bensoussan and Lions [2].

To prove Theorem 1, we must first prove five preliminary lemmas.

LEMMA 1. *Let $F(x, t)$ be a continuous function for which $F_t(x, t)$, $F_x(x, t)$ and $F_{xx}(x, t)$ are continuous; then*

$$\begin{aligned} F(y(T^-), T)-F(x, 0) &= \sum_{\tau_i<T}[F(y(\tau_i), \tau_i)-F(y(\tau_i^-), \tau_i)] \\ &\quad +\int_0^T [F_t(y(t), t)+\mu F_x(y(t), t)+\tfrac{1}{2}\sigma^2 F_{xx}(y(t), t)]\,dt \\ &\quad +\int_0^T F_x(y(t), t)\sigma\,dw_t. \end{aligned} \tag{19}$$

Proof of Lemma 1. If τ_k is such that $\tau_k < T \leqq \tau_{k+1}$, then we may apply Ito's formula on the intervals $[\tau_{i-1}, \tau_i)$, $i = 1, \cdots, k$, and on the interval $[\tau_k, T)$ to find

$$\begin{aligned} F(y(\tau_i^-), \tau_i) &= F(y(\tau_{i-1}), \tau_{i-1})+\int_{\tau_{i-1}}^{\tau_i} F_x(y(t), t)\sigma\,dw_t \\ &\quad +\int_{\tau_{i-1}}^{\tau_i} [F_t(y(t), t)+\mu F_x(y(t), t)+\tfrac{1}{2}\sigma^2 F_{xx}(y(t), t)]\,dt \end{aligned} \tag{20}$$

and

$$\begin{aligned} F(y(T^-), T) &= F(y(\tau_k), \tau_k)+\int_{\tau_k}^{T} F_x(y(t), t)\sigma\,dw_t \\ &\quad +\int_{\tau_k}^{T} [F_t(y(t), t)+\mu F_x(y(t), t)+\tfrac{1}{2}\sigma^2 F_{xx}(y(t), t)]\,dt \end{aligned} \tag{21}$$

Summing (20) for $i = 1, \ldots, k$ and adding (21) yields (19).

LEMMA 2. *Let $F(x)$ be an arbitrary bounded twice continuously differentiable function with $|F'(x)| \leqq M$ and*

$$(GF)(x) \equiv -\beta F(x)+\mu F'(x)\tfrac{1}{2}\sigma^2 F''(x) \tag{22}$$

bounded. Then for $\beta > 0$,

$$-F(x) = E\sum_{i=1}^{\infty}[F(y(\tau_i))-F(y(\tau_i^-))]\,e^{-\beta\tau_i}+E\int_0^{\infty} e^{-\beta t}[(GF)(y(t))]\,dt. \tag{23}$$

Proof of Lemma 2. Apply Lemma 1 to $F(x, t) = e^{-\beta t}F(x)$ to obtain

$$\begin{aligned} e^{-\beta T}F(y(T^-))-F(x) &= \int_0^T e^{-\beta t}[-\beta F+\mu F'+\tfrac{1}{2}\sigma^2 F'']\,(y(t))\,dt \\ &\quad +\int_0^T \sigma F'(y(t))\,e^{-\beta t}\,dw_t+\sum_{\tau_i<T}[F(y(\tau_i))-F(y(\tau_i^-))]\,e^{-\beta\tau_i}. \end{aligned} \tag{24}$$

Letting $T \to \infty$ we find

$$\begin{aligned} -F(x) &= \int_0^{\infty} e^{-\beta t}(GF)(y(t))\,dt+\sum_1^{\infty}[F(y(\tau_i))-F(y(\tau_i^-))]e^{-\beta\tau_i} \\ &\quad +\int_0^{\infty} \sigma F'(y(t))\,e^{-\beta t}\,dw_t. \end{aligned} \tag{25}$$

Now

$$\int_0^\infty E\sigma^2[F'(y(t))]^2\, e^{-2\beta t}\, dt \leqq M^2\sigma^2 \int_0^\infty e^{-2\beta t}\, dt < \infty, \tag{26}$$

so that

$$E\int_0^\infty \sigma F'(y(t))\, e^{-\beta t}\, dw_t = 0. \tag{27}$$

Taking the expectation of (25) yields (23).

LEMMA 3. *Let $\phi(x)$ be continuous and in $L^2(R)$ and let $E\sum_{i=1}^\infty e^{-\beta\tau_i} < \infty$; then*

$$|Z(\phi)| = \left|E\int_0^\infty e^{-\beta t}\phi(y(t))\, dt\right| \leqq C\|\phi\|_{L^2(R)}. \tag{28}$$

Proof of Lemma 3. Consider the expectation

$$X_i(\phi) = E\int_0^\infty \chi_{[\tau_i,\tau_{i+1})}(t)|\phi(y(t))|\, e^{-\beta t}\, dt, \tag{29}$$

where from (1) we find that

$$y(t) = y(\tau_i) + \mu(t-\tau_i) + \sigma(w_t - w_{\tau_i}) \quad \text{for } \tau_i \leqq t < \tau_{i-1}. \tag{30}$$

Let

$$D(\tau_i, t) = \mu(t-\tau_i) + \sigma(w_t - w_{\tau_i}) \tag{31}$$

and let $s = t - \tau_i$ for $t \geqq \tau_i$ so that

$$D(\tau_i, t) = D(\tau_i, \tau_i + s) = \mu s + \sigma(w_{\tau_i+s} - w_{\tau_i}). \tag{32}$$

Denote

$$w_s^{\tau_i} = w_{\tau_i+s} - w_{\tau_i} \tag{33}$$

so that by the strong Markov property (see Gihman and Skorohod [3, p. 30]) $w_s^{\tau_i}$ is independent of $\mathscr{F}_{\tau_i}$ and is a Wiener process. Now

$$\begin{aligned} X_i(\phi) &= E\int_0^\infty \chi_{[0,\tau_{i+1}-\tau_i)}(s)|\phi(y(\tau_i)+D(\tau_i,\tau_i+s))|\, e^{-\beta(s+\tau_i)}\, ds \\ &\leqq \int_0^\infty E[e^{-\beta\tau_i}|\phi(y(\tau_i)+D(\tau_i,\tau_i+s))|]\, e^{-\beta s}\, ds. \end{aligned} \tag{34}$$

Let $p(x, s)$ be the probability density function for $D(\tau_i, \tau_i+s)$ so that

$$p(x,s) = \frac{\exp[-(x-\mu s)^2/(2s\sigma^2)]}{\sqrt{2\pi s}\,\sigma} \tag{35}$$

Hence

$$\begin{aligned} E[e^{-\beta\tau_i}|\phi(y(\tau_i)+D(\tau_i,\tau_i+s)|] &= \int_R E[e^{-\beta\tau_i}|\phi(y(\tau_i)+r)|p(r,s)]\, dr \\ &= \int_R E[e^{-\beta\tau_i}|\phi(x)|p(x-y(\tau_i),s)]\, dx \end{aligned}$$

since $D(\tau_i, \tau_i + s)$ is independent of $y(\tau_i)$ and of τ_i.
Therefore

$$|Z(\phi)| \leqq \sum_{i=0}^{\infty} X_i(\phi) \leqq \sum_{i=0}^{\infty} E \int_0^{\infty} e^{-\beta\tau_i} e^{-\beta s} \left[\int_R |\phi(x)| p(x - y(\tau_i), s)\, dx \right] ds$$

$$= \sum_{i=0}^{\infty} E \int_0^{\infty} \int_R e^{-(\beta/2)(\tau_i+s)} |\phi(x)|\, e^{-(\beta/2)(\tau_i+s)} p(x - y(\tau_i), s)\, dx\, ds \tag{36}$$

$$\leqq \left[\sum_{i=0}^{\infty} E \int_0^{\infty} \int_R e^{-\beta\tau_i} e^{-\beta s} \phi^2(x)\, dx\, ds \right]^{1/2}$$

$$\cdot \left[\sum_{i=0}^{\infty} E \int_0^{\infty} \int_R e^{-\beta s - \beta\tau_i} p^2(x - y(\tau_i), s)\, dx\, ds \right]^{1/2}$$

by the Cauchy–Schwarz inequality.

Now for any value of $y(\tau_i)$ we have

$$\int_R p^2(x - y(\tau_i), s)\, dx = \int_R \frac{\exp[-(x - y(\tau_i) - \mu s)^2/(s\sigma^2)]}{2\pi s\sigma^2}\, dx$$

$$= \frac{1}{2\sigma\sqrt{\pi s}} \int_R \frac{\exp[-(x - (\mu s + y(\tau_i))^2/(2(s\sigma^2/2))]}{\sqrt{2\pi(s\sigma^2/2)}}\, dx \tag{37}$$

$$= \frac{1}{2\sigma\sqrt{\pi s}}.$$

Thus

$$|Z(\phi)| \leqq \left[E \sum_{i=0}^{\infty} e^{-\beta\tau_i} \right] \|\phi\|_{L^2(R)} \frac{1}{\sqrt{\beta}} \left[\int_0^{\infty} \frac{e^{-\beta s}}{2\sigma\sqrt{\pi s}}\, ds \right]^{1/2}$$

$$\leqq \left[E \sum_{i=0}^{\infty} e^{-\beta\tau_i} \right] \|\phi\|_{L^2(R)} C_1(\beta) = C\|\phi\|_{L^2(R)}.$$

Remark 2. $Z(\phi)$ is a bounded linear functional on $\phi \in L^2(R)$ and continuous, hence for arbitrary $\phi \in L^2(R)$, $Z(\phi)$ is defined by extension such that $|Z(\phi)| \leqq C\|\phi\|_{L^2(R)}$.

LEMMA 4. *Let $F(x)$ be bounded and continuous on $\bar{R}$ with $(GF)(x) \in L^2(R)$, $F'(x)$ absolutely continuous and let $\sum_{i=1}^{\infty} E\, e^{-\beta\tau_i} < \infty$. Then*

$$F(x) = -\sum_{i=1}^{\infty} E[F(y(\tau_i)) - F(y(\tau_i^-))]\, e^{-\beta\tau_i} + E \int_0^{\infty} e^{-\beta t} [(-GF)(y(t))]\, dt. \tag{38}$$

Proof. Let $F_n(x)$ be a sequence of smooth functions bounded with two continuous bounded derivatives such that

$$F_n \to F \quad \text{uniformly on } \bar{R} \tag{39}$$

and

$$GF_n \to GF \quad \text{in } L^2(R). \tag{40}$$

Then from Lemma 2 we have

$$(41)\quad -F_n(x)=\sum_{i=1}^{\infty} E[F_n(y(\tau_i))-F_n(y(\tau_i^-))]\,e^{-\beta\tau_i}+E\int_0^{\infty} e^{-\beta t}[(GF_n)(y(t))]\,dt$$

Now by Lemma 3,

$$(42)\quad \left|E\int_0^{\infty} e^{-\beta t}[(GF-GF_n)(y(t))]\,dt\right| \leqq C\|GF-GF_n\|_{L^2(R)}\to 0.$$

Lastly, denote $\|F\|=\sup_x |F(x)|$ so that

$$\left|\sum_{i=1}^{\infty} E[F(y(\tau_i))-F_n(y(\tau_i))-(F(y(\tau_i^-))-F_n(y(\tau_i^-)))]\,e^{-\beta\tau_i}\right|$$

$$\leqq 2\|F-F_n\|\sum_{i=1}^{\infty} E\,e^{-\beta\tau_i}\to 0. \qquad \text{Q.E.D.}$$

LEMMA 5. *Let $u(x)$ satisfy* (6). *Assume that $\sum_{i=1}^{\infty} E\,e^{-\beta\tau_i}<\infty$,*

$$(43)\quad E\int_0^{\infty} e^{-\beta t}[(-Gu)(y(t))]\,dt<\infty$$

and

$$(44)\quad \sum_{i=1}^{\infty} E|\xi_i|\,e^{-\beta\tau_i}<\infty,$$

then

$$(45)\quad u(x)=\sum_{i=1}^{\infty} E[u(y(\tau_i^-))-u(y(\tau_i))]\,e^{-\beta\tau_i}+E\int_0^{\infty} e^{-\beta t}[(-Gu)(y(t))]\,dt.$$

Remark 3. By (43) we mean

$$(46)\quad E\left\{\int_0^{\infty} e^{-\beta t}[\beta u(y(t))-\mu u'(y(t))]\,dt\right\}+\tfrac{1}{2}\sigma^2 Z(u'')<\infty,$$

where $Z(\cdot)$ is given by (28), $u\geqq 0$ and $|u'|\leqq M$.

Proof. Clearly u satisfies a uniform Lipschitz condition.

$$(47)\quad |u(x)-u(y)|\leqq M|x-y|.$$

Define

$$(48)\quad u_\alpha(x)=e^{-\alpha x^2}u(x)\quad \text{for } 1\geqq\alpha>0,$$

so that $u_\alpha(x)$ is continuous and bounded on $\bar{R}$ and $Gu_\alpha\in L^2(R)$.

Furthermore, $u_\alpha'(x)$ is bounded for $0<\alpha\leqq 1$; thus $u(x)$ satisfies a uniform Lipschitz condition

$$(49)\quad |u_\alpha(x)-u_\alpha(y)|\leqq L|x-y|.$$

From Lemma 4 we have

$$(50)\quad u_\alpha(x)=-\sum_{i=1}^{\infty} E[u_\alpha(y(\tau_i))-u_\alpha(y(\tau_i^-))]\,e^{-\beta\tau_i}+E\int_0^{\infty} e^{-\beta t}[(-Gu_\alpha)(y(t))]\,dt.$$

From Lemma 3 we find that

$$\left|E\int_0^\infty e^{-\beta t}[-\tfrac{1}{2}\sigma^2 u''(y(t))]\,dt\right|<\infty, \tag{51}$$

and as $\alpha\to 0$ by Lemma 3,

$$\left|E\int_0^\infty e^{-\beta t}[-\tfrac{1}{2}\sigma^2(u''(y(t))-u_\alpha''(y(t)))]\,dt\right|\to 0. \tag{52}$$

Since $|u'(x)|\leqq M$ we have that

$$\left|E\int_0^\infty e^{-\beta t}[-\mu u'(y(t))]\,dt\right|\leqq|\mu|\frac{M}{\beta}<\infty \tag{53}$$

and

$$\left|E\int_0^\infty e^{-\beta t}[-\mu(u'(y(t))-u_\alpha(y(t)))]\,dt\right|\leqq|\mu|\|u'-u_\alpha'\|\frac{1}{\beta}\to 0 \tag{54}$$

as $\alpha\to 0$, where $\|F\|=\sup_x|F(x)|$.

Thus (43), (51) and (53) imply that

$$0\leqq E\int_0^\infty e^{-\beta t}\beta u(y(t))\,dt<\infty, \tag{55}$$

and by the monotone convergence theorem

$$E\int_0^\infty e^{-\beta t}\beta u_\alpha(y(t))\,dt\to E\int_0^\infty e^{-\beta t}\beta u(y(t))\,dt. \tag{56}$$

Lastly,

$$\left|E\sum_{i=1}^\infty[u_\alpha(y(\tau_i))-u_\alpha(y(\tau_i^-))]\,e^{-\beta\tau_i}\right|\leqq L\sum_{i=1}^\infty|\xi_i|\,e^{-\beta\tau_i}, \tag{57}$$

and by hypothesis (44) the expectation of the right-hand side of (57) is finite. Hence by the dominated convergence theorem as $\alpha\to 0$,

$$\begin{aligned}\sum_{i=1}^\infty E[u_\alpha(y(\tau_i))-u_\alpha(y(\tau_i^-))]\,e^{-\beta\tau_i}\\ \to\sum_{i=1}^\infty E[u(y(\tau_i))-u(y(\tau_i^-))]\,e^{-\beta\tau_i}<\infty.\end{aligned} \tag{58}$$

From (50), (52), (54), (56) and (58) the result (45) follows. We now complete the proof of Theorem 1.

Proof of Theorem 1. Suppose v is such that $J_x(v)<\infty$,[5] then by (7),

$$u(y(\tau_i^-))\leqq B(\xi_i)+u(y(\tau_i)). \tag{59}$$

Hence

$$\sum_{i=1}^\infty E[u(y(\tau_i^-))-u(y(\tau_i))]\,e^{-\beta\tau_i}\leqq\sum_{i=1}^\infty EB(\xi_i)\,e^{-\beta\tau_i}<\infty \tag{60}$$

[5] This justifies the assumption that only a finite number of stopping times may occur in any bounded time interval a.s.

since $J_x(v)<\infty$, which implies that

$$\sum_{i=1}^{\infty} E\, e^{-\beta\tau_i}<\infty \tag{61}$$

and

$$\sum_{i=1}^{\infty} E|\xi_i|\, e^{-\beta\tau_i}<\infty. \tag{62}$$

Furthermore, from (8) we find that

$$\int_0^{\infty} e^{-\beta t}E[(-Gu)(y(t))]\,dt \leqq \int_0^{\infty} e^{-\beta t}EH(y(t)))\,dt<\infty, \tag{63}$$

since $J_x(v)<\infty$. Thus applying Lemma 5 we have that

$$u(x)\leqq J_x(v). \tag{64}$$

We must first establish that $E\, e^{-\beta\hat{\tau}_i}u(\hat{y}(\hat{\tau}_i^-))<\infty$ for $i\geqq 1$. We do so by induction, the induction hypothesis following from

$$E\, e^{-\beta\hat{\tau}_1}u(\hat{y}(\hat{\tau}_1^-))\leqq u(x)+E\, e^{-\beta\hat{\tau}_1}M|y(\hat{\tau}_1^-)-x|<\infty$$

since it may easily be shown that for any stopping time $\tau\geqq 0$,

$$E\, e^{-\beta\tau}|y(\tau)|<\infty, \tag{65a}$$

where

$$dy(t)=\mu\,dt+\sigma\,dw_t \quad \text{for } 0\leqq t\leqq\tau \quad \text{and} \quad y(0)=0. \tag{65b}$$

Now assume $E\, e^{-\beta\hat{\tau}_i}u(\hat{y}(\hat{\tau}_i^-))<\infty$ and note that (7) implies

$$E\, e^{-\beta\hat{\tau}_i}u(\hat{y}(\hat{\tau}_i^-))=E\, e^{-\beta\hat{\tau}_i}B(\hat{\xi}_i)+E\, e^{-\beta\hat{\tau}_i}u(\hat{y}(\hat{\tau}_i))<\infty \tag{66}$$

which in turn implies that $E\, e^{-\beta\hat{\tau}_i}u(\hat{y}(\hat{\tau}_i))<\infty$ since both terms on the right-hand side of (66) are nonnegative. Hence,

$$\begin{aligned} E\, e^{-\beta\hat{\tau}_{i+1}}u(\hat{y}(\hat{\tau}_{i+1}^-)) &\leqq E\, e^{-\beta\hat{\tau}_{i+1}}u(\hat{y}(\hat{\tau}_i))+ME\, e^{-\beta\hat{\tau}_{i+1}}|\hat{y}(\hat{\tau}_{i+1}^-)-\hat{y}(\hat{\tau}_i)| \\ &\leqq E\, e^{-\beta\hat{\tau}_i}u(\hat{y}(\hat{\tau}_i))+ME\, e^{-\beta(\hat{\tau}_{i+1}-\hat{\tau}_i)}|\hat{y}(\hat{\tau}_{i+1}^-)-\hat{y}(\hat{\tau}_i)| \\ &<\infty \end{aligned}$$

by (65). Furthermore we have established that $E\, e^{-\beta\hat{\tau}_i}u(\hat{y}(\hat{\tau}_i))<\infty$ for $i\geqq 1$.

We note that (9) and (10) imply that for $x\in C$, (8) holds as an equality and thus $u''(x)$ is continuous for $x\in C$. Hence we may use Ito's lemma to find that

$$\begin{aligned} 0\leqq E\int_{\hat{\tau}_{i-1}}^{\hat{\tau}_i^-} e^{-\beta s}H(\hat{y}(s))\,ds &= E\int_{\hat{\tau}_{i-1}}^{\hat{\tau}_i^-} e^{-\beta s}[\beta u-\mu u'-\tfrac{1}{2}\sigma^2u''](\hat{y}(s))\,ds \\ &= E\, e^{-\beta\hat{\tau}_{i+1}}u(\hat{y}(\hat{\tau}_{i-1}))-E\, e^{-\beta\hat{\tau}_i}u(\hat{y}(\hat{\tau}_i^-))<\infty. \end{aligned} \tag{67}$$

Summing (67) for $i \geqq 1$ we find that

$$0 \leqq E\int_0^\infty e^{-\beta s}H(\hat{y}(s))\,ds = u(x) - \sum_{i=1}^{\infty} E\,e^{-\beta\hat{\tau}_i}[u(\hat{y}(\hat{\tau}_i^-)) - u(\hat{y}(\hat{\tau}_i))]$$

(68)

$$= u(x) - \sum_{i=1}^{\infty} E\,e^{-\beta\hat{\tau}_i}B(\hat{\xi}_i) \leqq u(x) < \infty,$$

from which it follows that

$$J_x(\hat{v}) = u(x) < \infty. \qquad \text{Q.E.D.} \tag{69}$$

3. The finite horizon model. We proceed in a manner exactly paralleling § 2. The proofs of several of the lemmas are shortened since they are modifications of previous lemmas.

Let $0 \leqq \tau_i \leqq \tau_2 \leqq \cdots \leqq \tau_i \leqq \cdots \leqq T$ be an increasing sequence of stopping times adapted to $\mathcal{F}_t$, such that only a finite number will occur in $[0, T]$ a.s. (or else the expected cost of the policy would be infinite). Let $y(t)$ and v be defined by (1)–(3). We modify the holding cost function and the cost of control to allow it to depend upon time, i.e., let $H(x, t)$ be continuous and nonnegative on $S = R \times (0, T)$ and define $B(\xi, t)$ on $\bar{S}$ by

$$B(\xi, t) = K_\xi(t) + k_\xi(t)|\xi|, \tag{70}$$

where

$$K_\xi(t) = \begin{cases} K^+(t) > 0 & \text{for } \xi \geqq 0, \\ K^-(t) > 0 & \text{for } \xi < 0, \end{cases} \tag{71}$$

and

$$k_\xi(t) = \begin{cases} k^+(t) > 0 & \text{for } \xi \geqq 0, \\ k^-(t) > 0 & \text{for } \xi < 0, \end{cases} \tag{72}$$

and K^+, K^-, k^+, k^- are continuous.

The total cost associated with the policy v is

$$J_x(v) = E\left\{\sum_{\tau_i \leqq T} B(\xi_i, \tau_i) + \int_0^T H(y(s), s)\,ds + \Psi(y(T))\right\},$$

where $\Psi(\cdot)$ is a nonnegative continuous penalty function.

Suppose there exists a function $u(x, t)$ such that

$$\begin{cases} u \in C^1(\bar{S}), \quad u \geqq 0, \quad u_x \text{ is absolutely continuous and bounded,} \\ u_{xx} \in L^2(S), \quad e^{-\alpha x^2}u_t \in L^2(S) \quad \text{for } 0 < \alpha \leqq 1, \quad \text{and} \quad u_t \leqq 0; \end{cases} \tag{73}$$

$$u(x, t) \leqq \inf_\xi [B(\xi, t) + u(x + \xi, t)] \equiv Qu(x, t) \quad \forall x, t; \tag{74}$$

$$-u_t - \mu u_x - \tfrac{1}{2}\sigma^2 u_{xx} \equiv -Gu \leqq H \quad \text{a.e.;} \tag{75}$$

$$(H + Gu)(u - Qu) = 0; \tag{76}$$

and

$$u(x, T) = \Psi(x). \tag{77}$$

Note that (73) and (77) imply that $\Psi(x)$ must satisfy

$$|\Psi(x)-\Psi(y)| \leqq \|u_x\|\,|x-y|. \tag{78}$$

Let the continuation region be defined by

$$C=\{(x,t): u(x,t)<Qu(x,t)\}, \tag{79}$$

which is an open subset of S. We define the optimal impulse control in the same manner as in § 2 but with η a function of both x and t chosen such that

$$B(\eta(x,t),t)+u(x+\eta(x,t),t)=Qu(x,t) \quad \forall (x,t). \tag{80}$$

Thus, for example, $\hat{\xi}_1=\eta(\hat{y}(\hat{\tau}_1^-),\hat{\tau}_1)$, etc.

THEOREM 2. *If there exists a solution to* (73)–(77), *then*

$$u(x,0)=J_x(\hat{v}) \leqq J_x(v) \quad \forall v \tag{81}$$

and $\hat{v}$ defines the optimal impulse control.

We again proceed with a series of lemmas.

LEMMA 6. *If $\phi(x,t)\in L^2(s)$ and $E\sum_{\tau_i<T}1<\infty$, then*

$$\left|E\int_0^T \phi(y(t),t)\,dt\right| \leqq C\|\phi\|_{L^2(s)}. \tag{82}$$

Proof of Lemma 6. This lemma can easily be proved by modifying the proof of Lemma 3. However, we may proceed even more easily by using a result of Bensoussan and Lions [2]. Their Lemma 2.1 can be used to show that for any $\tau_i \leqq \tau_{i+1} \leqq T$,

$$E\left[\left|\int_0^T \chi_{[\tau_i,\tau_{i+1})}(t)\phi(y(t),t)\,dt\right| \;\middle|\; \tau_i<T\right] \leqq C_1\|\phi\|_{L^2(s)}. \tag{83}$$

Thus

$$\begin{aligned}\left|E\int_0^T \phi(y(t),t)\,dt\right| &\leqq E\left\{\sum_{\tau_i<T} E\left[\left|\int_0^T \chi_{[\tau_i,\tau_{i+1})}(t)\phi(y(t),t)\,dt\right| \;\middle|\; \tau_i<T\right]\right\} \\ &\leqq E\left[\sum_{\tau_i<T}1\right] C_1\|\phi\|_{L^2(s)} \qquad (84)\\ &= C\|\phi\|_{L^2(s)}. \qquad \text{Q.E.D.}\end{aligned}$$

LEMMA 7, *Let $F(x,t)$ be bounded and continuous on $\bar{S}$ with $(GF)(x,t)\in L^2(S)$ and assuming $E\sum_{\tau_i\leqq T}1<\infty$. Then*

$$\begin{aligned}F(x,0)=EF(y(T),T)&-E\sum_{\tau_i\leqq T}[F(y(\tau_i),\tau_i)-F(y(\tau_i^-),\tau_i)] \\ &-E\int_0^T [(GF)(y(t),t)]\,dt.\end{aligned} \tag{85}$$

Proof of Lemma 7. Let $F^n(x,t)$ be a sequence of smooth bounded functions, with F_t^n, F_x^n and F_{xx}^n continuous and bounded such that

$$F^n \to F \text{ uniformly on } \bar{S} \tag{86}$$

and

$$GF^n \to GF \quad \text{in } L^2(S). \tag{87}$$

Using F^n in (19), adding $F^n(y(T), T) - F^n(y(T^-), T)$ to both sides and taking expectations yields

$$\begin{aligned} F^n(x, 0) = EF^n(y(T), T) - E \sum_{\tau_i \leqq T} [F^n(y(\tau_i), \tau_i) - F^n(y(\tau_i^-), \tau_i)] \\ - E \int_0^T [(GF^n)(y(t), t)]\, dt \end{aligned} \tag{88}$$

since F_x^n is bounded. Now by the same type of argument used in Lemma 4, we can show that (88) converges to (85). Q.E.D.

LEMMA 8. *Let $u(x, t)$ be a function satisfying* (73). *Assume that* $E \sum_{\tau_i \leqq T} 1 < \infty$,

$$E \sum_{\tau_i \leqq T} |\xi^i| < \infty \tag{89}$$

and

$$E \int_0^T [(-Gu)(y(t), t)]\, dt < \infty; \tag{90}$$

then

$$\begin{aligned} u(x, 0) = Eu(y(T), T) - E \sum_{\tau_i \leqq T} [F(y(\tau_i), \tau_i) - F(y(\tau_i^-), \tau_i)] \\ + E \int_0^T [(-Gu)(y(t), t)]\, dt \end{aligned} \tag{91}$$

Proof of Lemma 8. We proceed as in the proof of Lemma 5. The boundedness of u_x guarantees that u satisfies a uniform Lipschitz condition

$$|u(x, t) - u(y, t)| \leqq M|x - y|. \tag{92}$$

Define

$$u^\alpha(x, t) = e^{-\alpha x^2} u(x, t) \tag{93}$$

for $1 \geqq \alpha > 0$ so that $u^\alpha(x, t)$ is continuous and bounded on $\bar{S}$ and $Gu^\alpha(x, t) \in L^2(S)$. From Lemma 7 we have

$$\begin{aligned} u^\alpha(x, 0) = Eu^\alpha(y(T), T) - E \sum_{\tau_i \leqq T} [u^\alpha(y(\tau_i), \tau_i) - u^\alpha(y(\tau_i^-), \tau_i)] \\ + \int_0^T E[(-Gu^\alpha)(y(t), t)]\, dt. \end{aligned} \tag{94}$$

The remainder of the proof closely parallels the proof of Lemma 5. We need only remark that $Eu^\alpha(y(T), T) \to Eu(y(T), T)$ by the monotone convergence theorem and that $Eu(y(T), T) < \infty$ because all the other terms in (94) converge to finite limits by hypothesis. Q.E.D.

With the use of Lemma 8 instead of Lemma 5 and the observation that $u(y, T) = \Psi(y)$, the proof of Theorem 2 proceeds *mutatis mutandis* from the proof of Theorem 1.

Acknowledgment. I would like to acknowledge the helpful comments and enlightening discussion provided by Victor J. Mizel.

REFERENCES

[1] A. BENSOUSSAN AND J. L. LIONS, *Nouvelle formulation de problèmes de contrôle impulsionnel et applications*, C.R. Acad. Sci. Paris Ser. A, 276 (1973), pp. 1189–1192.

[2] ———, *Nouvelles methodes in contrôle impulsionnel*, Appl. Math. and Optimization Quart., 1 (1975), pp. 289–312.

[3] I. I. GIHMAN AND A. V. SKOROHOD, *Stochastic Differential Equations*, Springer-Verlag, New York, 1968.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

OPTIMAL CONTROL OF JUMP PROCESSES*

R. BOEL† AND P. VARAIYA‡

Abstract. The paper proposes an abstract model for the problem of optimal control of systems subject to random perturbations, for which the principle of optimality takes on an appealing form. This model is specialized to the case where the state of the controlled system is realized as a jump process. The additional structure permits operationally useful optimality conditions. Some illustrative examples are solved.

1. Introduction. This paper addresses the problem of the optimal control of dynamical systems subject to random perturbations. It does so in the following way. First, in § 3, an abstract mathematical model is proposed in which the choice of controller is modeled as choosing a probability measure over the measurable space of state trajectories. This idea was first developed by Beneš [1], [2] and Duncan and Varaiya [11] in order to prove existence of an optimal control when the perturbations form a Brownian motion. Second, in § 4, we derive optimality conditions for the abstract model using dynamic programming and elements of martingale theory in the way developed by Davis and Varaiya [9] for the Brownian motion case. Their approach in turn was motivated by the work of Rishel [20]; it also has some resemblance to earlier work by Kushner [16], and Stratonovich [26]. Some of the extensions of their results as given in § 4 are special cases of recent results of Striebel [25]. While the abstract model does serve to unify previous results, further comprehension of the scope of the model can be gained and an evaluation of its practical import can be made only by working through with more specialized problems with additional structure. Hence, in §§ 5 and 6, the case where the random perturbations constitute a jump process is discussed in detail. Related results using different methods have been reported by Rishel [21] and Stone [24] and we shall compare them later. We note that there are control problems with jump disturbances which must be modeled quite differently from the model of §§ 5 and 6. As examples of these we mention the work of Rishel [22] and Sworder [27].

2. Conventions and notations. Let $(\Omega, \mathcal{F})$ be a measurable space. Let $I=[0, T]$ or $[0, \infty)$ be a fixed time interval with the corresponding final time denoted T. A stochastic process is always a triple $(z_t, \mathcal{F}_t, \mathcal{P})$, $t \in I$, where $\mathcal{P}$ is a probability measure on $(\Omega, \mathcal{F})$, $(\mathcal{F}_t)$ is an increasing family of sub-σ-fields of $\mathcal{F}$ and (z_t) is a family of $(\mathcal{F}_t)$-adapted random variables with values in some unspecified measurable space. When the context makes it clear we write the

* Received by the editors January 3, 1975, and in final revised form June 2, 1976. This research was supported in part by the Joint Services Electronics Program Contract F44620-71-C-0087 and the National Science Foundation Grant GK-43024X.

† Department of Electrical Engineering and Computer Sciences and Electronics Research Laboratory, University of California, Berkeley, California 94720. The work of this author was supported in part by an ESRO-NASA International Fellowship.

‡ Department of Electrical Engineering and Computer Sciences and Electronics Research Laboratory, University of California, Berkeley, California 94720. Part of this paper was completed when this author was visiting the Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts.

stochastic process $z = (z_t, \mathscr{F}_t, \mathscr{P})$, $t \in I$, as (z_t) or $(z_t, \mathscr{F}_t)$ or $(z_t, \mathscr{P})$ or z; z_t, without the parentheses, usually denotes the random variable at time t instead of the process.

All probability spaces are assumed complete, and every increasing family of σ-fields, $(\mathscr{F}_t)$, is assumed right-continuous i.e., $\mathscr{F}_t = \bigcap_{s>t} \mathscr{F}_s$. An $(\mathscr{F}_t, \mathscr{P})$-martingale is a uniformly integrable martingale $(m_t, \mathscr{F}_t, \mathscr{P})$, $t \in I$, with $m_0 = 0$ a.s. The collection of all such martingales is denoted $\mathscr{M}^1(\mathscr{F}_t, \mathscr{P})$. In a similar way, we define $\mathscr{M}^2(\mathscr{F}_t, \mathscr{P})$, $\mathscr{M}^1_{\text{loc}}(\mathscr{F}_t, \mathscr{P})$, $\mathscr{M}^2_{\text{loc}}(\mathscr{F}_t, \mathscr{P})$, the classes of $(\mathscr{F}_t, \mathscr{P})$-uniformly square integrable, locally integrable, locally square integrable martingale, and it will be assumed that a version of these processes is chosen such that it has right-continuous sample paths with left-hand limits.

$\mathscr{A}^+(\mathscr{F}_t, \mathscr{P})$ is the class of all processes $(a_t, \mathscr{F}_t, \mathscr{P})$, $t \in I$, which vanish at 0, $a_0 = 0$ a.s., with right-continuous, nondecreasing sample paths, and which are uniformly integrable, $\sup_t Ea_t < \infty$. $\mathscr{A}(\mathscr{F}_t, \mathscr{P}) = \mathscr{A}^+(\mathscr{F}_t, \mathscr{P}) - \mathscr{A}^+(\mathscr{F}_t, \mathscr{P})$ is then the class of processes with integrable variation. The classes $\mathscr{A}^+_{\text{loc}}$, $\mathscr{A}_{\text{loc}}$ are defined in the usual way.

A family (z_t) of $(\mathscr{F}_t)$-adapted functions taking values in a metric space is said to be $(\mathscr{F}_t)$-predictable if there is a sequence of such families (z_t^n), $n = 1, 2, \cdots$, with left-continuous sample paths such that $\lim_{n\to\infty} z_t^n(\omega) = z_t(\omega)$ for all $(t, \omega) \in I \times \Omega$.

3. Abstract model of the control problem. The model proposed below is similar to the one presented and investigated in [25]. It consists of three interconnected parts: a description of the dynamical system, i.e., the way in which it is affected by the control action, a description of the set of allowable control laws, and a description of the cost associated with each control law. The assumptions imposed are given next.

We suppose given measurable spaces $(Z, \mathscr{Z})$, the *state* space, and $(\Omega, \mathscr{X})$, the *trajectory* or *sample* space. Also given is a function $x_t(\omega)$: $I \times \Omega \to Z$ which is measurable with respect to $\mathscr{B}_I \times \mathscr{X}$. Let $\mathscr{X}_t = \sigma\{x_s \mid s \leqq t\}$ and without losing generality we assume that $\mathscr{X} = \sigma\{x_t \mid t \in I\}$. We now assume

S_1. The behavior of the system under the action of any (admissible) control law u is completely described by the specification of a probability measure $\mathscr{P}^u$ on $(\Omega, \mathscr{X})$.

Thus for each control law u, $x^u = (x_t, \mathscr{X}_t, \mathscr{P}^u)$, $t \in I$, is a well-defined stochastic process. We are evidently modeling the system as a *controlled probability space* rather than as a controlled set of trajectories which is more customary. Of course in the deterministic context the latter model is the more natural one. We now describe the set of control laws.

We suppose given a measurable space $(\mathscr{V}, \mathscr{B}_u)$, the *control* space, where $\mathscr{V}$ is a metric space. Also given is an increasing family of σ-fields, $(\mathscr{Y}_t)$ called the family of *observations*, such that $\mathscr{Y}_t \subset \mathscr{X}_t$, $t \in I$. A collection $\mathscr{U}$ of functions $u_t(\omega)$: $I \times \Omega \to \mathscr{V}$ is a *collection* of (*admissible*) *control laws* if the following holds:

S_2. (i) (u_t) is $(\mathscr{Y}_t)$-adapted and $(u_t, \mathscr{Y}_t, \mathscr{P}^u)$, $t \in I$ is a measurable process.

(ii) $\mathscr{U}$ is closed under concatenation, i.e., if $u, v \in \mathscr{U}$, then so does (u, v, t) where $(u, v, t)(s) = u(s)$ for $s \leqq t$, $= v(s)$ for $s > t$.

(iii) For each $u \in \mathscr{U}$ and $A \in \mathscr{X}_t$, $\mathscr{P}^u(A)$ depends only on u_s, $s \leqq t$ i.e., if $v \in \mathscr{U}$ is such that $u_s \equiv v_s$, $s \leqq t$, then $\mathscr{P}^u(A) = \mathscr{P}^v(A)$; for each $u \in \mathscr{U}$, $A \in$

$\mathscr{X}$, $E^u(1_A \mid \mathscr{X}_t)$ does not explicitly depend on u_s, $s \leqq t$ i.e., if $v \in \mathscr{U}$ such that $u_s \equiv v_s$, $s \geqq t$, then $E^u(1_A \mid \mathscr{X}_t) = E^v(1_A \mid \mathscr{X}_t)$.

In the above, (iii) is a version of a causality condition and also expresses the notion that the past trajectory x_s, $s \leqq t$ serves as a state at t whereas (ii) is essential for dynamic programming. In (i) the requirement that u_t is $\mathscr{Y}_t$-measurable indicates that $\mathscr{Y}_t$ is the σ-field of observations available up to t.

We can now describe the cost of control. Associated with each $u \in \mathscr{U}$ is a unique cost $J(u)$ given by

$$J(u) = E^u\left[\int_I r_0^t c(t, u(t))\, d\Lambda^u(t) + r_0^T J_T\right] \tag{3.1}$$

where E^u denotes expectation with respect to $\mathscr{P}^u$, T denotes the final time of I, and the other terms are described below.

C_1. The *instantaneous cost* $c: I \times U \times \Omega \to R$ is a nonnegative function which is jointly measurable with respect to $\mathscr{B}_I \times \mathscr{B}_U \times \mathscr{X}$ ($\mathscr{B}_I$, $\mathscr{B}_U$ are the Borel sets of I, U), continuous with respect to u for fixed t, ω and measurable with respect to $\mathscr{X}_t$ for fixed t, u.

C_2. The *time rate* $\Lambda^u: I \times \Omega \to R$, defined for each $u \in \mathscr{U}$, is $(\mathscr{Y}_t)$-predictable and, for each ω, the sample path $t \to \Lambda^u(t, \omega)$ is right-continuous and increasing. Furthermore, $d\Lambda^{(u,v,t)}(s) = d\Lambda^u(s)$ for $s \leqq t$, $= d\Lambda^v(s)$ for $s > t$. (See S_2 above for a definition of (u, v, t)).

Since Λ^u can have discontinuous sample paths, the indefinite (Stieltjes) integral $\int_0^t r_0^s c(s, u(s))\, d\Lambda^u(s)$ can be discontinuous. The most useful examples of time rates are

(a) $\Lambda^u(t) \equiv t$; whenever the sample paths are absolutely continuous with respect to Lebesgue measure on I this case obtains.

(b) $\Lambda^u(t, \omega) = \sum 1_{\{t \geqq \tau_i(\omega)\}}$ which counts the number of $(\mathscr{X}_t)$-stopping times τ_i, $i = 1, 2, \cdots$, which occur before t.

(c) Λ^u is the predictable increasing process associated with the counting process in (b), and which can replace the latter in (3.1) whenever $c(t, u(t))$ is a $(\mathscr{X}_t)$-predictable process, since the values of the integrals coincide (see [19]).

C_3. The *discounting rate* $r_s^t(\omega)$: is a nonnegative function defined for $\omega \in \Omega$, s, t in I with $s \leqq t$. For fixed s, $r_s^t(\omega)$ is $(\mathscr{Y}_t)$-adapted, jointly $\mathscr{B}_I \times \mathscr{Y}$ measurable, and uniformly integrable, and has continuous sample paths for fixed ω. Furthermore, for each u

$$r_{t_1}^{t_3} = r_{t_1}^{t_2} r_{t_2}^{t_3} \quad \text{a.s. } \mathscr{P}^u \text{ for } t_1 \leqq t_2 \leqq t_3,$$

$$r_t^t = 1 \qquad \text{a.s. } \mathscr{P}^u \text{ for all } t.$$

C_4. The *terminal cost* $J_T: \Omega \to R$ is a nonnegative $\mathscr{X}$-measurable function. J_T is the cost incurred at or after the final time T. When $T = \infty$ it will be assumed that $J_T \equiv 0$.

C_5. For all $u \in \mathscr{U}$, $J(u) < \infty$.

The problem of optimal control is to find $u^* \in \mathscr{U}$ such that

$$J(u^*) = \inf_{u \in \mathscr{U}} J(u).$$

Such u^* is called an *optimal control.*

Remark 3.1. (i) The fixed time interval I can be replaced by a random interval $[0, \tau] \subset I$ where τ is a $(\mathscr{X}_t)$-stopping time. This can be achieved by setting $c(t, u, \omega) = 0$ for $t \geqq \tau(\omega)$ or by making Λ^u constant after τ. If τ does not depend on u one can set $r_0^t(\omega) = 0$, $t \geqq \tau(\omega)$.

(ii) The discounting rate $r_s^t(\omega)$ is not allowed to depend explicitly on u. In an economic context this implies that the controller cannot directly influence the interest rate. Of course, since the distribution of r_s^t is dependent upon $\mathscr{P}^u$ there is a possibility of introducing indirect control.

(iii) Except for the special results with complete information (i.e., $\mathscr{Y}_t \equiv \mathscr{X}_t$) or Markovian assumptions, the final cost J_T can depend explicitly on the control law u. Again, except for these special cases, $c(t, u, \omega)$ can be made to depend upon the past u_s, $s \leqq t$ of the control. These generalizations are not made here since the notational burdens become intolerable.

(iv) There are important applications, e.g. optimal stopping time problems, where the optimal control cannot be chosen to be predictable. The results here do not apply to this class of applications.

4. Optimality results for the abstract model. Since the proofs of the results are simple modifications of proofs published in [9] we have been content with citing the correspondence. The assumptions made in § 3 are enforced throughout.

4.1. Principle of optimality. Let $u, v \in \mathscr{U}$ and $t \in I$. We define

$$\psi(u, v, t) = E^{(u,v,t)}\left\{\int_t^T r_t^s c(s, v_s)\, d\Lambda_s^v + r_t^T J_T \,|\, \mathscr{Y}_t\right\}.$$

Evidently, from the assumptions made above,

$$\psi(u, v, t) \in L^1(\Omega, \mathscr{Y}_t, \mathscr{P}^u).$$

(The first part of S_2 (iii) implies that

$$E^{u,v,t}|\psi(u, v, t)| = E^u|\psi(u, v, t)|$$

justifying the notation $L^1(\Omega, \mathscr{Y}_2, \mathscr{P}^u)$.) The random variable $\psi(u, v, t)$ is the conditional expectation given the observation $\mathscr{Y}_t$ of the future cost beyond time t, evaluated at t, when u is adopted on $[0, t]$ and v is adopted beyond t. To evaluate these costs at time 0 it is only necessary to multiply $\psi(u, v, t)$ by r_0^t. Since L^1 is a complete lattice under the natural partial ordering for real-valued functions the following $\mathscr{P}_u$-essential infimum exists:

$$W(u, t) = \operatorname*{\Lambda}_{v \in \mathscr{U}} \psi(u, v, t) \in L^1(\Omega, \mathscr{Y}_t, \mathscr{P}^u).$$

Note that $W(u, 0) = J^* = \Lambda_{u \in \mathscr{U}} J(u)$ is the infimum of the achievable costs. The process $(W(u, t), \mathscr{Y}_t, \mathscr{P}^u)$ is called the *value function* corresponding to u. The fact that for different control laws u and v, the corresponding $\mathscr{P}_u$ and $\mathscr{P}_v$ can be singular, does not pose any problem since, in the following optimality conditions, $W(u, t)$ and $W(v, t)$, or related processes, need never be compared (one must interpret carefully expressions such as $\min_{u \in \mathscr{U}}$ in (5.41)).

The next definition was introduced by Rishel [20]. It was used in [9].

DEFINITION 4.1. $\mathcal{U}$ is said to be *relatively complete* with respect to W if for each $u \in \mathcal{U}$, $t \in I$, $\varepsilon > 0$ there exists $v \in \mathcal{U}$ such that

$$\psi(u, v, t) \leqq W(u, t) + \varepsilon \quad \text{a.s. } \mathcal{P}^u.$$

LEMMA 4.1. *$\mathcal{U}$ is relatively complete with respect to W.*

Proof. The proof is identical with that of [9, Lemma 3.1]. □

THEOREM 4.1. *For $t_1 \leqq t_2$ in I and $u \in \mathcal{U}$ we have*

$$W(u, t_1) \leqq E^u\left[\int_{t_1}^{t_2} r_{t_1}^s c(s, u_s)\, d\Lambda_s^u \,|\, \mathcal{Y}_{t_1}\right] + E^u\left[r_{t_1}^{t_2} W(u, t_2) \,|\, \mathcal{Y}_{t_1}\right], \tag{4.1}$$

$$W(u, T) = E^u[J_T \,|\, \mathcal{Y}_T]. \tag{4.2}$$

Furthermore, u is optimal if and only if equality holds in (4.1).

Proof. The proof depends on Lemma 4.1 and follows the same lines as that of [9, Thm. 3.1]. □

COROLLARY 4.1. *For $u \in \mathcal{U}$, the process*

$$\hat{J}_t^u = r_0^t W(u, t) + E^u\left[\int_0^t r_0^t c(s, u_s)\, d\Lambda^u(s) \,|\, \mathcal{Y}_t\right]$$

is a $(\mathcal{Y}_t, \mathcal{P}^u)$ sub-martingale. u is optimal if and only if this process is a martingale.

Proof. The proof is immediate from Theorem 4.1. □

Since the process

$$\left(E^u\left[\int_I r_0^s c(s, u_s)\, d\Lambda_s^u + r_0^T J_T \,|\, \mathcal{Y}_t\right]\right) \in \mathcal{M}^1(\mathcal{Y}_t, \mathcal{P}^u),$$

therefore the process $(w(u, t))$ is a $(\mathcal{Y}_t, \mathcal{P}^u)$-supermartingale, where

$$\begin{aligned} w(u, t) &= E^u\left[\int_t^T r_0^s c(s, u_s)\, d\Lambda_s^u + r_0^T J_T \,|\, \mathcal{Y}_t\right] - r_0^t W(u, t) \\ &= r_0^t[\psi(u, u, t) - W(u, t)]. \end{aligned} \tag{4.2'}$$

COROLLARY 4.2. *For $u \in \mathcal{U}$, the process $(w(u, t), \mathcal{Y}_t, \mathcal{P}^u)$ is a potential. u is optimal if and only if $w(u, t) \equiv 0$.*

Remark 4.1. (i) The model proposed above is a special case of the one presented by Striebel [25] and the results obtained above can be obtained from hers. In particular Corollary 4.1 is a version of [25, Thm. 3]. The additional structure that we have imposed will be used to obtain the more detailed results given below. It is possible to replace the "relative completeness" property by the slightly weaker "ε-lattice" property introduced by Striebel.

(ii) Following Samuelson [23] we can give a heuristic interpretation of the submartingale $(\hat{J}_t^u)$. Its value is the expected cost evaluated at t, using the observation $\mathcal{Y}_t$, given that u is adopted up to t and an optimal control is adopted beyond t. This expected value will increase if the nonoptimal control is used for a longer time, accounting for the sub-martingale property. If u is optimal, however, then the expected cost remains constant.

(iii) Theorem 4.1 can be rederived from Corollary 4.1. Hence the optional sampling theorem implies that in (4.1) we may replace the deterministic times t_1

and t_2 by any $(\mathcal{Y}_t)$-stopping times $\tau_1 \leqq \tau_2$ with values in I. This observation is often useful.

(iv) Sometimes, as in [2], [9], [11], there exists a probability measure $\mathcal{P}$ on $(\Omega, \mathcal{X})$ such that $\mathcal{P}^u \ll \mathcal{P}$ for all $u \in \mathcal{U}$. One can then introduce $L(u) = d\mathcal{P}^u/d\mathcal{P}$ and

$$\phi(u, v, t) = E\left\{L(u, v, t)\left[\int_t^T r_t^s c(s, v_s)\, d\Lambda_s^v + r_t^T J_T\right] \middle| \mathcal{Y}_t\right\},$$

$$V(u, t) = \operatorname*{\Lambda}_{v \in \mathcal{U}} \phi(u, v, t).$$

The previous results can be restated in terms of the unnormalized value function. While in an optimal filtering context working with such unnormalized quantities has certain advantages (see e.g. [6]), we are unable to observe similar advantages in the optimal control context.

(v) The random variable $w(u, t)$ expresses the loss incurred by using u beyond t as compared with an optimal control. From definition (4.2′) and from C_5 we can verify that it is potential of class (D). By [18, VI, T3 and T4] $w(u, t+)$ exists and is also a potential. Using (4.2′) we then get a right-continuous process $r_0^{t+} W(u, t+)$ satisfying (4.1). Even though in general $w(u, t+)$ and $W(u, t+)$ need not even be versions of $w(u, t)$ and $W(u, t)$, it is implicitly assumed from now on that these right-continuous process are meant. In special cases, such as value-decreasing controls (first step in proof of Lemma 4.2 of [9]) and complete information, with cost bounded by k (apply [18, VI, T16] to the submartingales $\psi(u, v, t) + k \cdot \Lambda_{(t,T]}^v$), $W(u, t)$ actually has a right-continuous modification, justifying the above notation. Hence by Meyer's decomposition theorem [18, VII, T31], there is a unique predictable process $(A_0^t w(u)) \in \mathcal{A}^+(\mathcal{Y}_t, \mathcal{P}^u)$ and a martingale $(m^w(u, t)) \in \mathcal{M}^1(\mathcal{Y}_t, \mathcal{P}^u)$ such that

$$w(u, t) = \tilde{J}(u) - A_0^t w(u) - m^w(u, t)$$

where $\tilde{J}(u) = w(u, 0) = J(u) - J^*$. We know, furthermore, that the following weak limit (in the sense of the $\sigma(L^1, L^\infty)$-topology) exists (see [18]).

$$\begin{aligned} A_0^t(u) &= \operatorname*{weak\,lim}_{h \to 0} \int_0^t \frac{1}{h} E^u[w(u, s) - w(u, s+h) \mid \mathcal{Y}_s]\, ds \\ &= \operatorname*{weak\,lim}_{h \to 0} \left\{ \int_0^t \frac{1}{h} E^u\left[\int_s^{s+h} r_0^\sigma c(\sigma, u_\sigma)\, d\Lambda_\sigma^u \mid \mathcal{Y}_s\right] ds \right. \\ &\qquad \left. - \int_0^t \frac{1}{h} E^u[r_0^s W(u, s) - r_0^{s+h} W(u, s+h) \mid \mathcal{Y}_s]\, ds \right\}. \end{aligned} \tag{4.3}$$

Now it is easy to see that there exists a predictable process $(\gamma(u)) \in \mathcal{A}^+(\mathcal{Y}_t, \mathcal{P}^u)$ such that

$$\gamma(u) = \operatorname*{weak\,lim}_{h \to 0} \int_0^t \frac{1}{h} E^u\left[\int_s^{s+h} r_0^\sigma c(\sigma, u_\sigma)\, d\Lambda_\sigma^u \mid \mathcal{Y}_s\right] ds. \tag{4.4}$$

From (4.3), (4.4) we may conclude that there exists a predictable process $(A_0^t w(u)) \in \mathscr{A}(\mathscr{Y}_t, \mathscr{P})$, viz. $A_0^t W(u) = \gamma(u) - A_0^t w(u)$, such that

$$A_0^t W(u) = \operatorname*{weak\,lim}_{h \to 0} \int_0^t \frac{1}{h} E^u[r_0^s W(u, s) - r_0^{s+h} W(u, s+h) \,|\, \mathscr{Y}_s]\, ds.$$

This is sufficient to apply Meyer's decomposition theorem to the process $(r_0^t W(u, t))$ and we may conclude that

$$\begin{aligned} r_0^t W(u, t) &= r_0^0 W(u, 0) - A_0^t W(u) + m^W(u, t) \\ &= J^* - A_0^t W(u) + m^W(u, t), \end{aligned} \tag{4.5}$$

where $(m^W(u, t)) \in \mathscr{M}(\mathscr{Y}_t, \mathscr{P}^u)$. Furthermore, since $(W(u, t))$ is evidently of class (D), the decomposition in (4.5) is unique.

In terms of this decomposition we can rewrite (4.1), after multiplying both sides by r_0^t, as

$$E^u[A_0^{t_2} W(u) - A_0^{t_1} W(u) \,|\, \mathscr{Y}_{t_1}] \leqq E^u\left[\int_{t_1}^{t_2} r_0^s c(s, u_s)\, d\Lambda_s^u \,|\, \mathscr{Y}_{t_1}\right] \quad \text{a.s. } \mathscr{P}^u, \tag{4.6}$$

and we have equality if and only if u is optimal. With these results in hand we can proceed as in the proof of [9, Thm. 4.1] to establish the next proposition.

THEOREM 4.2. *There exists a constant J^* and for every $u \in \mathscr{U}$ there exists a predictable process $(A_0^t(u)) \in \mathscr{A}(\mathscr{Y}_t, \mathscr{P}^u)$ such that*

$$E^u A_0^T(u) = J^* - E^u(r_0^T J_T), \tag{4.7}$$

and such that for $(\mathscr{Y}_t)$-stopping times $\tau_1 \leqq \tau_2$ with values in I,

$$E^u\left[-A_{\tau_1}^{\tau_2}(u) + \int_{\tau_1}^{\tau_2} r_0^s c(s, u_s)\, d\Lambda_s^u \,|\, \mathscr{Y}_{\tau_1}\right] \geqq 0 \quad \text{a.s. } \mathscr{P}^u. \tag{4.8}$$

A control law $u = u^*$ is optimal if and only if equality holds in (4.8) *for deterministic times $t_1 \leqq t_2$, and then, furthermore,*

$$J(u^*) = J^*,$$

$$r_0^t W(u^*, t) = E^u[A_t^T(u^*) + r_0^T J_T \,|\, \mathscr{Y}_t] \quad \text{a.s. } \mathscr{P}^u$$

Remark 4.2. This result is a considerable improvement over [9, Thm. 4.1] since there the inequality (4.6) and hence (4.7) is established only for those u which are "value decreasing", i.e., for which $(W(u, t))$ is a supermartingale. The same shortcoming can be noticed in [20]. Of course, if u is value decreasing, then in (4.5) $(A_0^t W(u))$ is an increasing process.

4.2. Local optimality conditions. One can divide both sides in (4.8) by $\tau_2 - \tau_1$ and take limits as $\tau_2 - \tau_1 \to 0$. The basic idea is to express $A_0^t(u)$ as an integral with respect to Λ^u. It appears necessary however to restrict attention to value decreasing controls.

So let $u \in \mathscr{U}$ be such that $(W(u, t))$ is a supermartingale. Then $(A_0^t W(u))$ is a predictable increasing process and (4.6) can be rewritten as

$$0 \leqq E^u[A_{t_1}^{t_2} W(u) \,|\, \mathscr{Y}_{t_1}] \leqq E^u\left[\int_{t_1}^{t_2} r_0^s c(s, u_s)\, d\Lambda_s^u \,|\, \mathscr{Y}_{t_1}\right] \quad \text{a.s. } \mathscr{P}^u.$$

For any nonnegative, $(\mathcal{Y}_t)$-predictable process (Θ_t) we have, for all $0 \leqq t_1 < t_2 \leqq T$, $A \in \mathcal{Y}_{t_1}$

$$(4.8') \qquad 0 \leqq E^u 1_A \int_{(t_1,t_2]} \theta_s \cdot dA_0^s W(u) \leqq E^u 1_A \int_{(t_1,t_2]} \theta_s \cdot \widehat{r_0^s c(s, u_s)}\, d\Lambda_s^u$$

where $\widehat{r_0^s c(s, u_s)}$ denotes the predictable projection of $E^u(r_0^s c(s, u_s) \mid \mathcal{Y}_s)$ (see [31, VT14]). This follows from [31, VT25], and an application of Fubini theorem to show that

$$E^u\Big[\int_{(t_1,t_2]} \theta_s [r_0^s c(s, u_s) - E^u(r_0^s c(s, u_s) \mid \mathcal{Y}_s)]\, d\Lambda_s^u \mid \mathcal{Y}_{t_1}\Big] = 0.$$

Hence, whenever the second integral in (4.8') vanishes, so does the first, and by the Radon–Nikodym theorem (applied to measure $\mu(\theta)$ on $R_+ \times \Omega$ with the predictable σ-field associated with $(\mathcal{Y}_t)$ [31, IV, D2]) there exists a predictable process $(\alpha_t(u), \mathcal{Y}_t, \mathcal{P}_u)$ such that $0 \leqq \alpha_t(u) \leqq 1$ $\mathcal{P}^u$ a.s., and

$$A_0^t W(u) = \int_0^t \alpha_s(u) \widehat{r_0^s c(s, u_s)}\, d\Lambda_s^u.$$

Using this representation we can restate Theorem 4.2 in a "local" version.

THEOREM 4.3. *There exists a constant J^* and for every value-decreasing $u \in \mathcal{U}$, there exists a $(\mathcal{Y}_t)$-predictable process $(\alpha_t(u))$, $0 \leqq \alpha_t(u) \leqq 1$, $\mathcal{P}^u$ a.s., such that*

$$(4.9) \qquad E^u \int_0^T \alpha_t(u) \widehat{r_0^t c(t, u_t)}\, d\Lambda_t^u = J^* - E^u(r_0^T J_T)$$

and

$$(4.10) \qquad [1 - \alpha_t(u)(\omega)] \widehat{r_0^t c(t, u_t)}(\omega) \geqq 0$$

for $d\Lambda^u \times d\mathcal{P}^u$ almost all (t, ω). A control law $u = u^*$ is optimal if and only if equality holds in (4.10). *Then, furthermore*

$$r_0^t W(u^*, t) = E^{u^*}\Big[\int_t^T \alpha_s(u^*) \widehat{r_0^s c(s, u_s^*)}\, d\Lambda_s^{u^*} + r_0^t J_T \mid \mathcal{Y}_t\Big] \quad \mathcal{P}^{u^*} \text{ a.s.}$$

Remark. If $r_0^t c(t, u_t) > 0$ a.s. $d\Lambda^u \times d\mathcal{P}^u$ one can also find an $(\alpha_t(u))$ such that

$$A_0^t W(u) = \int_0^t \alpha_s(u)\, d\Lambda_s^u$$

giving a slightly easier version of Theorem 4.3.

4.3. Complete information. Suppose $\mathcal{Y}_t \equiv \mathcal{X}_t$ so that at each time t the controller has complete information about the past. Then

$$\begin{aligned} \psi(u, v, t) &= E^{(u,v,t)}\Big[\int_t^T r_t^s c(s, v_s)\, d\Lambda_s^v + r_t^T J_T \mid \mathcal{Y}_t\Big] \\ &= E^v\Big[\int_t^T r_t^s c(s, v_s)\, d\Lambda_s^v + r_t^T J_T \mid \mathcal{Y}_t\Big] \\ &= \psi(v, v, t) \end{aligned}$$

by assumption S_2(iii). Hence $W(u, t)$ does not depend on u, and the preceding results are simpler. Nevertheless the process $A_0^t W(u)$ still depends upon past values of the control law u. Its "derivative" $\alpha_t(u)$ however will often be independent of values of u before t as seen in [9] and in the following sections.

5. Optimality results for jump processes. In this section the abstract model of § 3 is specialized to the case of a dynamical system whose state process is a (fundamental) jump process as studied in [5], [6]. The additional structure gives more content to the formal results established earlier. For a review of the definitions and properties of jump processes see [6, § 2].

5.1. The model and its limitations. The state space $(Z, \mathscr{Z})$ is now also a Blackwell space. Ω consists of all functions $\omega: I \to Z$ which are piecewise constant, right-continuous and have only a finite number of jumps in a finite time interval. $x_t(\omega)$: $I \times \Omega \to Z$ is just the evaluation function $x_t(\omega) = \omega(t)$. $\mathscr{X}_t$, $\mathscr{X}$ are defined as before.

The observations $(\mathscr{Y}_t)$ are obtained as follows. We suppose given a Blackwell space $(Y, \mathscr{Y})$ and a measurable map $\gamma: Z \to \mathscr{Z}$. Let $y_t = \gamma(x_t)$ and $\mathscr{Y}_t = \sigma\{y_s \mid s \leqq t\}$.

With (x_t) and (y_t) we can now associate the following discrete random measures.

$$(5.1)\quad P^x(B, t)(\omega) = \sum_{s \leqq t} 1_{\{x_{s-}(\omega) \neq x_s(\omega) \in B\}}$$

$$= \text{number of jumps of } x(\omega) \text{ which occur before } t \text{ and end in } B \in \mathscr{Z};$$

$$(5.2)\quad P^y(C, t)(\omega) = \sum_{s \leqq t} 1_{\{y_{s-}(\omega) \neq y_s(\omega) \in C\}}$$

$$= \text{number of jumps of } y(\omega) \text{ which occur before } t \text{ and end in } C \in \mathscr{Y}.$$

Note that $P^x(B, t)$ is $\mathscr{X}_t$-measurable and $P^y(C, t)$ is $\mathscr{Y}_t$-measurable.

We can now define the collection of admissible control laws $\mathscr{U}$ and the probability measures $\mathscr{P}^u$, $u \in \mathscr{U}$. Let $(\mathscr{V}, \mathscr{B}_u)$ be the control space, where $\mathscr{V}$ is a metric space. $\mathscr{U}$ is the collection of all functions $u_t(\omega)$: $I \times \Omega \to \mathscr{U}$ which are $(\mathscr{Y}_t)$-predictable. It is supposed that for each $u \in \mathscr{U}$ there is given a probability measure $\mathscr{P}^u$ on $(\Omega, \mathscr{X})$ such that the stochastic process $\mathscr{X}^u = (x_t, \mathscr{X}_t, \mathscr{P}^u)$, $t \in I$, is a jump process in the sense of [5.6]. (It is evident that S_2(i), S_2(ii) are satisfied by these assumptions.) Now from [5, Thm. 2.1] we know that to say that x_u is a jump process it is equivalent to say that there exist *continuous* processes $(\tilde{P}_u^x(B, t)) \in \mathscr{A}^+_{\text{loc}}(\mathscr{X}_t, \mathscr{P}^u)$ for each $B \in \mathscr{Z}$ such that

$$(5.3)\qquad (Q_u^x(B, t)) = (P^x(B, t) - \tilde{P}_u^x(B, t)) \in \mathscr{M}^1_{\text{loc}}(\mathscr{X}_t, \mathscr{P}^u).$$

Thus the action of $u \in \mathscr{U}$ is completely described by specifying the correspondence

$$u \to \{(\tilde{P}_u^x(B, t), \mathscr{X}_t, \mathscr{P}^u) \mid B \in \mathscr{Z}\}.$$

Compare this with a result of Jacod [15] which states that, under conditions satisfied in this section, there exists a one-to-one relationship between "kernels" of predictable $(\tilde{P}_u^x(B, t) | B \in Z)$ and all probability measures $\mathscr{P}^u$ on $(\Omega, \mathscr{X})$. To

guarantee assumption S_2(iii) and to simplify some notation later on we suppose that for $u \in \mathcal{U}$ and $B \in \mathcal{Z}$, $\tilde{P}_u^x(B, t)(\omega)$ is given by

$$\tilde{P}_u^x(B, t)(\omega) = \int_B \int_0^t f(z, s, u_s(\omega), \omega)\mu(dz, ds, u_s(\omega), \omega) \tag{5.4}$$

where the integral is an ordinary Stieltjes integral and the prespecified functions f and μ satisfy these conditions:

(i) $f(z, s, u) = f(z, s, u, \omega)$: $Z \times I \times \mathcal{V} \times \Omega \to R_+$ is jointly measurable, continuous in u for fixed z, s, ω and for fixed z, u, $(f(z, t, u, \omega))$ is $(\mathcal{Z}_t)$-predictable.

(ii) $\mu(B, t, u, \omega) = \mu(B, t, u)$: $\mathcal{Z} \times I \times \mathcal{V} \times \Omega \to R_+$ is jointly measurable and for each fixed B, u, $(\mu(B, t, u))$ is $(\mathcal{Y}_t)$-predictable, continuous and increasing. (In practice μ is usually a deterministic process.) These assumptions are technical, but μ and f can with care be interpreted as jump rate and distribution of different types of jumps.

Finally the cost $J(u)$ incurred by $u \in \mathcal{U}$ is supposed given by

$$J(u) = E^u\left[\int_Z \int_I r_0^s c(z, s, u_s)\tilde{P}_u^x(dz, ds) + r_0^T J_T\right] \tag{5.5}$$

where c satisfies the same conditions as f does, and r_s^t, J_T satisfy the conditions imposed in § 3. It is assumed that $J(u) < \infty$ for all u. This completes the description of the mathematical model.

Before turning to the analysis of the model we discuss its limits in terms of which empirical control problems can and which cannot be adequately reflected in the model. First of all, as far as the behavior of the state trajectories is concerned the most serious limitation is the requirement that $(\tilde{P}_u^x(B, t))$ have continuous sample paths. It is known (see e.g. [5]) that this restriction is equivalent to saying that the stopping times at which the state jumps, i.e. the times of discontinuity of $(x_t(\omega))$, are totally inaccessible. In intuitive terms this means that if the controller observes the first n jumps, then the probability with which it can predict the $(n+1)$st jump *exactly* is zero, for each $n = 0, 1, 2 \cdots$ (see [5, Lemma 2.4]). Now most problems of queuing, inventory control, machine failures etc. indeed have this property. But there are some problems which do not. For example suppose that in an inventory control problem there is a fixed (deterministic) delay between the time an order is placed and the time that the corresponding delivery is made; evidently the total inventory jumps when the delivery is made and this time of jump can be predicted exactly, and so the model proposed here is inadequate for this example. Now the only reason why we have insisted on the total inaccessibility of the jump times is so that we can use the martingale representation theorems derived in [5]. More recently, such theorems have been obtained without the restriction on the jump times (see [7], [8], [13], [15]) and therefore the results announced below should be extendable to arbitrary jump processes.

The second limitation of the model appears to the requirement that controls have to be predictable processes. One reason for this is based on empirical considerations. Since the time when the state jumps cannot be anticipated with positive probability, and since in empirical situations there is an infinitesimal delay before the controller can observe and react to a change in state, therefore the

predictability requirement seems appropriate to us. In any event since μ is continuous in t, therefore $\tilde{P}^x_u$ defined by (5.4) is always continuous in t even if u is measurable and not predictable. Hence the results below remain unchanged whether we permit u to be any measurable process so long as we always take the predictable projection of f (as well as of c in (5.5)), or whether one insists at the outset that u be predictable. Because many of the following results can also be obtained for μ discontinuous, and to avoid problems with predictable projections, the predictability assumption has been made.

Finally, the cost functional (5.5) may appear too limiting since in many situations one may wish to have the cost increase only when a jump occurs. Thus one would prefer to have as cost the amount

$$E^u\left[\int_Z \int_I r_0^s c(z, s, u_s) P^x_u(dz, ds) + r_0^T J_T\right]$$

$$= E^u\left[\sum_{\substack{s \in I \\ x_{s-} \neq x_s}} r_0^s c(x_s, s, u_s) + r_0^T J_T\right].$$

But since $P^x - \tilde{P}^x$ is a martingale and since the integrand above is predictable, the quantity above is equal to $J(u)$ given by (5.3) and so there is no loss in generality. (This equality does not obtain if u is not predictable.)

5.2. Preliminary analysis. To simplify notation we write $1_C(z) = 1_{\{\gamma(z) \in C\}}$. Then, from (5.2),

$$P^y(C, t) = \int_Z \int_0^t 1_C(z) P^x(dz, ds). \tag{5.6}$$

We calculate the unique processes $(\tilde{P}^y_u(C, t), \mathcal{Y}_t, \mathcal{P}^u)$ so that

$$(Q^y_u(C, t)) = (P^y(C, t) - \tilde{P}^y_u(C, t)) \in \mathcal{M}^1_{\text{loc}}(\mathcal{Y}_t, \mathcal{P}^u). \tag{5.7}$$

For an arbitrary process (g_t) let $(\hat{g}_t)$ be the $(\mathcal{Y}_t)$ predictable projection of $E^u(g_t \mid \mathcal{Y}_t)$ (the appropriate u will always be clear from the context.) Then from (5.4) and (5.6)

$$P^y(C, t) - \int_Z \int_0^t \widehat{1_C f(z, s, u_s)} \mu(dz, ds, u_s)$$

$$= \int_Z \int_0^t 1_C(z)[P^x(dz, ds) - f(z, s, u_s)]\mu(dz, ds, u_s)$$

$$+ \int_Z \int_0^t [1_C(z) f(z, s, u_s) - \widehat{1_C f(z, s, u_s)}]\mu(dz, ds, u_s)$$

which is a member of $\mathcal{M}^1_{\text{loc}}(\mathcal{Y}_t, \mathcal{P}^u)$ applying [37, VT25] and Fubini's theorem to the second term. Hence

$$\tilde{P}^y_u(C, t) = \int_Z \int_0^t \widehat{1_C(z) f(z, s, u_s)} \mu(dz, ds, u_s). \tag{5.8}$$

A calculation similar to that of $\tilde{P}_u^y$ above gives us the expected value, given $\mathcal{Y}_t$, of the increment of the instantaneous cost on the right hand side of (4.6). In terms of the cost functional (5.5),

$$E^u\Big[\int_Z\int_{t_1}^{t_2} r_0^s c(z,s,u_s)\tilde{P}_u^x(dz,ds)\,\Big|\,\mathcal{Y}_{t_1}\Big]$$

$$= E^u\Big[\int_Z\int_{t_1}^{t_2} r_0^s c(z,s,u_s)f(z,s,u_s)\mu(dz,ds,u_s)\,\Big|\,\mathcal{Y}_{t_1}\Big]$$

$$= E^u\Big[\int_Z\int_{t_1}^{t_2} \widehat{r_0^s c(z,s,u_s)f(z,s,u_s)}\mu(dz,ds,u_s)\,\Big|\,\mathcal{Y}_{t_1}\Big]$$

so that, by (4.6),

$$0 \leqq E^u\Big[-A_{t_1}^{t_2}W(u)+\int_Z\int_{t_1}^{t_2} r_0^s c(z,s,u_s)f(z,s,u_s)\mu(dz,ds,u_s)\,\Big|\,\mathcal{Y}_{t_1}\Big]$$

which means that the process

$$a_t = -A_0^t W(u)+\int_Z\int_0^t \widehat{r_0^s c(z,s,u_s)f(z,s,u_s)}\mu(dz,ds,u_s) \tag{5.9}$$

is a $(\mathcal{Y}_t, \mathcal{P}^u)$-sub-martingale. It is evidently of class (D) and is right-continuous since in (4.5) a right-continuous modification of $A_0^t W(u)$ can be chosen and so by Meyer's decomposition theorem there is a unique predictable process $(b_t)\in \mathcal{A}^+(\mathcal{Y}_t, \mathcal{P}^u)$ and $(m_t)\in\mathcal{M}^1(\mathcal{Y}_t, \mathcal{P}^u)$ so that

$$a_t = b_t + m_t.$$

But from (5.9) we know that (a_t) is also $(\mathcal{Y}_t)$-predictable. Hence (m_t) is a predictable process with integrable variation. It must therefore vanish so that $a_t = b_t$. Hence a_t itself is increasing so that (5.9) can be expressed as

$$0 \leqq -A_{t_1}^{t_2}W(u)+\int_Z\int_{t_1}^{t_2} \widehat{r_0^s c(z,s,u_s)f(z,s,u_s)}\mu(dz,ds,u_s) \quad \text{a.s. } \mathcal{P}^u \tag{5.10}$$

for every admissible control u. Furthermore we have equality if and only if u is optimal.

5.3. Optimality condition for partial information. Recall the following definition from [5]. A measurable function $\beta\colon Y\times I\times\Omega\to R$ is said to be in $L^1(\tilde{P}_u^y)$ if for fixed y, $\beta(y,\cdot)$ is a $(\mathcal{Y}_t)$-predictable process, and $E^u[\int_I\int_Z|\beta(y,s)|\tilde{P}_u^y(dy,ds)]<\infty$. β is said to be in $L^1_{\text{loc}}(\tilde{P}_u^y)$ if there is a sequence of $(\mathcal{Y}_t)$-stopping times $T_k\uparrow T$ a.s. $\mathcal{P}^u$ such that $(\beta 1_{\{t\leqq T_k\}})\in L^1(\tilde{P}_u^y)$ for each k.

We have the following version of Theorem 4.2.

THEOREM 5.1. *There exists a constant J^* and for every $u \in \mathcal{U}$ there exist a predictable process $(\bar{A}_0^t(u)) \in \mathcal{A}(\mathcal{Y}_t, \mathcal{P}^u)$ and a process $\beta^u \in L^1_{\text{loc}}(\tilde{P}^y_u)$ so that*

$$\bar{A}_0^T(u) - \int_Y \int_0^T \beta^u(y,s) P^y(dy, ds) = J^* - E^u[r_0^T J_T \mid \mathcal{Y}_T], \tag{5.11}$$

$$-\bar{A}_{\tau_1}^{\tau_2}(u) + \int_Z \int_{\tau_1}^{\tau_2} \overbrace{[\beta^u(\gamma(z), s) + r_0^s c(z, s, u_s)] f(z, s, u_s)} \mu(dz, ds, u_s) \geqq 0 \tag{5.12}$$

a.s. $\mathcal{P}^u$ for $(\mathcal{Y}_t)$-stopping times $\tau_1 \leqq \tau_2$ with values in I. A control law $u = u^*$ is optimal if and only if equality holds in (5.11) *for deterministic times $t_1 \leqq t_2$, and then, furthermore,*

$$J(u^*) = J^*,$$

$$r_0^t W(u^*, t) = J^* - \bar{A}_0^t(u^*) + \int_Y \int_0^t \beta^{u^*}(y,s) P^y(dy, ds). \tag{5.13}$$

Proof. Necessity. Let $u \in \mathcal{U}$. We have the representation (4.5),

$$r_0^t W(u,t) = J^* - A_0^t W(u) + m^W(u,t) \tag{5.14}$$

where $A_{t_1}^{t_2} W(u)$ satisfies (5.10),

$$0 \leqq -A_{t_1}^{t_2} W(u) + \int_Z \int_{t_1}^{t_2} r_0^s c(z, s, u_s) f(z, s, u_s) \mu(dz, ds, u_s) \tag{5.15}$$

with equality holding for u^*. By the martingale representation theorem [5, Thm. 3.4] there exists $\beta^u \in L^1_{\text{loc}}(\tilde{P}^y_u)$ such that

$$\begin{aligned} m^W(u,t) &= \int_Y \int_0^t \beta^u(y,s) Q_u^y(dy, ds) \\ &= \int_Y \int_0^t \beta^u(y,s) P^y(dy, ds) - \int_Y \int_0^t \beta^u(y,s) \tilde{P}^y_u(dy, ds) \\ &= \int_Y \int_0^t \beta^u(y,s) P^y(dy, ds) \\ &\quad - \int_Z \int_0^t \overbrace{\beta^u(\gamma(z), s) f(z, s, u_s)} \mu(dz, ds, u_s) \end{aligned} \tag{5.16}$$

by (5.8). Define

$$\bar{A}_0^t(u) = A_0^t W(u) + \int_Z \int_0^t \overbrace{\beta^u(\gamma(z), s) f(z, s, u_s)} \mu(dz, ds, u_s). \tag{5.17}$$

Substitution for $m^W(u,t)$ and $A_0^t W(u)$ from (5.16), (5.17) into (5.14) and (5.15) yields (5.11), (5.12) and (5.13).

Sufficiency. Now suppose (5.11), (5.12) holds. If we define $m^W(u,t)$ and $\bar{A}_0^t W(u)$ via (5.16) and (5.17), then (5.14) and (5.15) hold and the optimality of u^* follows from Theorem 4.2. □

Remark 5.1. (i) If we define $\Lambda^u(t) = \int_0^t \mu(Z, ds, u_s)$ then there exists a kernel $n(dz, t, u_t)$ such that $\mu(dz, dt, u_t) = n(dz, t, u_t)\Lambda^u(dt)$. Then Λ^u can act as a time rate and so exactly as in § 4.2 we can derive a local version of condition (5.12).

(ii) In many applications it is reasonable to suppose the existence of a probability measure $\mathcal{P}$ on $(\Omega, \mathcal{X})$ such that $\mathcal{P}^u \ll \mathcal{P}$ for all u. Then $\mathcal{P}^u$ can be described by specifying $\mathcal{P}$ and $L(u) = E[d\mathcal{P}^u/d\mathcal{P} \mid \mathcal{X}]$. Suppose further that $(x_t, \mathcal{P})$ is a jump process with compensating processes $(\tilde{P}^x(B, t), \mathcal{P})$, $B \in \mathcal{Z}$ given by

$$\tilde{P}^x(B, t) = \int_B \int_0^t f(z, s)\mu(dz, ds)$$

where $f(z, \cdot)$ is $(\mathcal{X}_t)$-predictable for each z and $(\mu(B, t))$ is a $(\mathcal{Y}_t)$-predictable increasing process. It can be shown then (see [6]) that for each u there is a process $\phi^u: Z \times I \times \Omega \to R$ such that

$$L_t(u) = E\left[\frac{d\mathcal{P}^u}{d\mathcal{P}} \middle| \mathcal{X}_t\right]$$

is given by

$$L_t(u) = \prod_{\substack{x_{s-} \neq x_s \\ s \leqq t}} [1 + \phi^u(x_s, s)] \exp\left[-\int_Z \int_0^t \phi^u(z, s) f(z, s)\mu(dz, ds)\right].$$

As a model (which satisfies the various assumptions of § 2) we can propose that the effect of a control u is determined by the process $(L_t(u))$ above in which

$$\phi^u(z, t, \omega) = \Phi(z, t, u_t, \omega)$$

where $\Phi: Z \times I \times U \times \Omega \to R$ is a fixed function. The processes $(\tilde{P}^x_u(B, t), \mathcal{X}_t, \mathcal{P}^u)$ are then given by (see [6])

$$\tilde{P}^x_u(B, t) = \int_B \int_0^t [1 + \Phi(z, s, u_s)]\tilde{P}^x(dz, ds).$$

The function Φ can be interpreted as the change in the rate at which jumps occur for $\mathcal{P}^u$ as compared with $\mathcal{P}$. In terms of this special model condition (5.12) reads as

$$-\bar{A}^{\tau_2}_{\tau_1}(\omega) + \int_Z \int_{\tau_1}^{\tau_2} \overbrace{[\beta^u(\gamma(z), s) + r^s_0 c(z, s, u_s)][1 + \Phi(z, s, u_s)]} f(z, s)\mu(dz, ds) \geqq 0.$$

5.4. Complete information. We assume that $y_t \equiv x_t$. Then, as observed in § 4.3, $W(u, t) = W(t)$ does not depend on u. However, it may appear that in the representation for $W(u, t)$ obtained in (5.13), (5.16) and (5.17), the processes $\bar{A}^t_0(u)$ and β^u still depend on u. To see that this is not the case, consider any two controls u, v. Then

$$\begin{aligned} r^t_0 W(t) &= J^* - \bar{A}^t_0(u) + \int_Z \int_0^t \beta^u(z, s)P^x(dz, ds) \\ &= J^* - \bar{A}^t_0(v) + \int_Z \int_0^t \beta^v(z, s)P^x(dz, ds). \end{aligned} \tag{5.18}$$

Now $(J^*-\bar{A}_0^t(u))$ and $(J^*-\bar{A}_0^t(v))$ are $(\mathscr{X}_t)$-predictable processes whereas the integrals in the equations above are piecewise constant with discontinuities at the jump times of the (x_t) process. It follows that

$$\int_Z \int_0^t \beta^u(z,s)P^x(dz,ds) = \int_Z \int_0^t \beta^v(z,s)P^x(dz,ds), \tag{5.19}$$

$$\bar{A}_0^t(u) = \bar{A}_0^t(v), \tag{5.20}$$

and so we have a considerably simpler version of Theorem 5.1.

THEOREM 5.2. *Suppose $y_t \equiv x_t$. Then u^* is optimal if and only if there exist a constant J^*, a predictable process $(\bar{A}_0^t) \in \mathscr{A}(\mathscr{X}_t, \mathscr{P}^u)$ and a process $\beta \in L^1_{\mathrm{loc}}(\tilde{P}^x_{u^*})$ so that*

$$\bar{A}_0^T - \int_Z \int_0^T \beta(z,s)P^x(dz,ds) = J^* - r_0^T J_T, \tag{5.21}$$

$$-\bar{A}_{t_1}^{t_2} + \int_Z \int_{t_1}^{t_2} [\beta(z,s) + r_0^s c(z,s,u_s)] f(z,s,u_s)\mu(dz,ds,u_s) \geqq 0 \tag{5.22}$$

a.s. $\mathscr{P}^u$ for all $u \in \mathscr{U}$ with equality holding for $u = u^*$. Then, furthermore,

$$J^* = J(u^*),$$

$$r_0^t W(t) = J^* - \bar{A}_0^t + \int_Z \int_0^t \beta(z,s)P^x(dz,ds). \tag{5.23}$$

Suppose henceforth (see Remark 5.1) that $\tilde{P}^x_u(dz,ds)$ has the form $\tilde{P}^x_u(dz,t) = \int_0^t n(dz,s,u_s)\lambda_s\,ds$ where $n(B,t,u_t)$ is a kernel satisfying the same assumptions as f in (5.4) (not necessarily $n(Z,t,u_t) = 1$) and (λ_t) is a nonnegative $(\mathscr{X}_t)$-predictable process independent of u. Then as shown in § 4.2, we can represent

$$\bar{A}_0^t = \int_0^t \tilde{\alpha}_s r_0^s c(s,u_s)\lambda_s\,ds = \int_0^t \alpha_s \cdot \lambda_s \cdot ds \tag{5.24}$$

for some (X_t)-predictable (α_t) (take predictable projection of $r_0^s c(s,u_s)$ if necessary). The local version of (5.21) now becomes

$$\left[-\alpha_t + \int_Z \beta(z,t) + r_0^t c(z,t,u_t)]n(dz,t,u_t)\right] \cdot \lambda_t \geqq 0 \tag{5.25}$$

for all (t,ω) with respect to $dt \times d\mathscr{P}^u$ measure, with equality when $u = u^*$. This gives us a version of the dynamic programming equation,

$$\lambda_t\left[-\alpha_t + \min_{u\in U} \int_Z [\beta(z,t) + r_0^t c(z,t,u)\right] n(dz,t,u) = 0, \tag{5.26}$$

and the minimum is achieved at $u^*(t,\omega)$ for almost all (t,ω) with respect to $dt \times d\mathscr{P}^{u^*}$ measure.

We shall now use (5.23) and (5.24) to directly relate $\bar{A}$ (or equivalently α and β) to the process $(r_0^t W(t))$. The basic idea is to note that $\bar{A}_0^t$ on the right hand side in (5.23) is continuous whereas the integral term is piecewise constant with discontinuities occurring only at the jump times of the (x_t) process. Thus the

discontinuous changes in rW account for β and the continuous changes account for α. To identify these changes we need a more detailed representation of rW. Set $T_0 \equiv 0$ and let $T_1 < T_2 < \cdots$ be the jump times of x defined by

$$T_{k+1}(\omega) = \inf\{t > T_k(\omega) \,|\, x_t(\omega) \neq x_{T_k}(\omega)\}, \qquad k = 0, 1, \cdots.$$

It is shown in [5] that $\mathscr{X}_t = \sigma(x_{T_k \wedge t}, T_k; 0 \leqq k < \infty)$, $\mathscr{X}_{T_n +} = \mathscr{X}_{T_n} = \sigma(x_{T_k}, T_k; 0 \leqq k \leqq n)$. Since $(r_0^t W(t))$ is adapted to $(\mathscr{X}_t)$, therefore there exist functions $w_k(t, t_0, z_0, \cdots, t_k, z_k)$, measurable in their arguments, so that

$$\begin{aligned} r_0^t W(t) &= \sum_{k \geqq 0} 1_{\{T_k \leqq t < T_{k+1}\}} w_k(t, T_0, x_{T_0}, \cdots, T_k, x_{T_k}) \\ &= J^* - \bar{A}_0^t + \sum_{k \geqq 0} 1_{\{T_k \leqq t < T_{k+1}\}} \sum_{l=1}^{k} \beta_k(x_{T_k}, T_k, x_{T_l \wedge t}, T_l \wedge t) \end{aligned} \tag{5.27}$$

with the Stieltjes-integral in (5.23) replaced by a sum with appropriate functions β_k.

The discontinuities of rW, which occur only at the T_k's, can now be identified as

$$\begin{aligned} &r_0^{T_k} W(T_k) - r_0^{T_k -} W(T_k -) \\ &\quad = w_k(T_k, T_0, x_{T_0}, \cdots, T_k, x_{T_k}) - w_{k-1}(T_k, T_0, x_{T_0}, \cdots, T_{k-1}, x_{T_{k-1}}). \end{aligned} \tag{5.28}$$

Hence the function β can be related to rW by

$$\beta(z, t) = \sum_{k \geqq 0} 1_{\{T_k < t \leqq T_{k+1}\}} b_k(z, t) \tag{5.29}$$

where, from (5.28),

$$\begin{aligned} b_{k-1}(z, t) = {} & w_k(t, T_0, x_{T_0}, \cdots, T_{k-1}, x_{T_{k-1}}, t, z) \\ & - w_{k-1}(t, T_0, x_{T_0}, \cdots, T_{k-1}, x_{T_{k-1}}). \end{aligned} \tag{5.30}$$

Using the uniqueness of the Doob–Meyer decomposition [18, VII T21] and the absolute continuity of $\tilde{P}_u^x(B, t)$, $\alpha_t \lambda_t$ can be identified by

$$\mathscr{L}_t^u[rW] = -\alpha(t)\lambda(t) + \int_Z \beta(z, t) n(dz, t, u_t) \tag{5.31}$$

where [18, VII T29] gives

$$\mathscr{L}_t^u[rW] = \operatorname*{weak\,lim}_{h \to 0} \frac{1}{h}\{E^u[r_0^{t+h} W(t+h) \,|\, \mathscr{X}_t] - r_0^t W(t)\}. \tag{5.32}$$

(Here weak lim means limit in the $\sigma(L^1, L^\infty)$ topology). From (5.27) it is obvious that the continuous part of $r_0^t W(t)$, described by $w_k(t, T_0, x_{T_0}, \cdots, T_k, x_{T_k})$ between jumps, behaves exactly like $J^* - \bar{A}_0^t$ which is absolutely continuous. This can also be stated as

$$\begin{aligned} &1_{\{T_k \leqq t \leqq t+h, T_{k+1}\}}[w_k(t+h, T_0, x_{T_0}, \cdots, T_k, x_{T_k}) - w_k(t, T_0, x_{T_0}, \cdots, T_k, x_{T_k})] \\ &\quad = \bar{A}_0^t - \bar{A}_0^{t+h} \end{aligned} \tag{5.33}$$

and the derivative of w_k is minus the derivative of $\bar{A}_0^t$. We can now obtain a formula for $\mathscr{L}_t^u[rW]$ as follows. We observe that in the stochastic interval $T_k \leqq t \leqq T_{k+1}$,

$$
\begin{aligned}
&E^u[r_0^{t+h}W(t+h)\,|\,\mathscr{X}_t]-r_0^tW(t)\\
&\qquad=[w_k(t+h,\,T_0,\cdots,x_{T_k})-w_k(t,\,T_0,\cdots,x_{T_k})]\\
&\qquad\quad\cdot\mathscr{P}^u[x \text{ does not jump in } [t,\,t+h]\,|\,\mathscr{X}_{T_k},\,T_{k+1}>t]\\
&\qquad+\int_Z\int_0^h[w_{k+1}(t+s,\,T_0,\cdots,x_{T_k},\,t+s,\,z)\\
&\qquad\qquad\qquad-w_k(t,\,T_0,\cdots,T_k,\,x_{T_k})]\\
&\qquad\quad\cdot\mathscr{P}^u[T_{k+1}-T_k\in ds,\,x_{T_{k+1}}\in dz\,|\,\mathscr{X}_{T_k},\,T_{k+1}>t]+o(h)
\end{aligned}
\tag{5.34}
$$

where absolute continuity of $\tilde{P}_u^x(B,t)$ (or $F_k^u(B,t)$, see (5.36)) implies that the probability of 2 or more jumps is $o(h)$. Now,

$$
\lim_{h\to 0}\mathscr{P}^u[x \text{ does not jump in } [t,\,t+h]\,|\,\mathscr{X}_{T_k},\,T_{k+1}>t]=1, \tag{5.35}
$$

$$
\mathscr{P}^u[T_{k+1}-T_k\in ds,\,x_{T_{k+1}}\in dz\,|\,\mathscr{X}_{T_k},\,T_{k+1}>t]=\frac{F_k^u(dz,\,ds)}{1-F_k^u(Z,\,t-T_k)}, \tag{5.36}
$$

where, by definition,

$$
F_k^u(B,\,t)=\mathscr{P}^u[T_{k+1}-T_k\leqq t,\,x_{T_{k+1}}\in B\,|\,\mathscr{X}_{T_k}]. \tag{5.37}
$$

Finally, as shown in [7], [8] and [15],

$$
\begin{aligned}
\tilde{P}^u(B,\,t)&=\int_B\int_0^t n(dz,\,s,\,u_s)\lambda_s\,ds\\
&=\sum_{i=0}^{k-1}\left[\int_0^{T_{i+1}-T_i}\frac{F_i^u(B,\,ds)}{1-F_i(Z,\,s)}\right]+\int_0^{t-T_k}\frac{F_k(B,\,ds)}{1-F_k(Z,\,s)}\qquad\text{for } T_k\leqq t\\
&<T_{k+1}.
\end{aligned}
\tag{5.38}
$$

From (5.33)–(5.38) we obtain

$$
\begin{aligned}
\mathscr{L}_t^u[rW]=\frac{\partial w_k}{\partial t}(t,\,T_0,\cdots,x_{T_k})+\int_Z[w_{k+1}(t,\,T_0,\cdots,x_{T_k},\,t,\,z)\\
-w_k(t,\,T_0,\cdots,x_{T_k})]n(dz,\,t,\,u_t)\lambda_t\\
\text{for } T_k\leqq t<T_{k+1}.
\end{aligned}
\tag{5.39}
$$

Substituting from (5.31) and (5.39) into (5.26) gives the next result.

THEOREM 5.3. *Suppose $y_t\equiv x_t$ and suppose that for $u\in\mathscr{U}$*

$$
\tilde{P}_u^x(dz,\,ds)=n(dz,\,s,\,u_s)\lambda_s\,ds. \tag{5.40}
$$

Then u^* is optimal if and only if there exist measurable functions $w_k(t, t_0, \cdots, t_k, z_k)$, which are absolutely continuous in t, so that

$$\frac{\partial w_k}{\partial t}(t, T_0, \cdots, x_{T_k}) + \min_{u \in \mathcal{U}} \lambda_t \int_Z [w_{k+1}(t, T_0, \cdots, x_{T_k}, t, z) - w_k(t, T_0, \cdots, x_{T_k}) + r_0^t c(z, t, u)] n(dz, t, u) = 0, \quad \text{for } T_k \leqq t < T_{k+1}, \tag{5.41}$$

$$w_k(T, T_0, \cdots, x_{T_k}) = J_T, \quad \text{for } T_k \leqq T < T_{k+1} \tag{5.42}$$

and the minimum in (5.40) *is achieved at $u = u^*(t, \omega)$ a.s. $\mathcal{P}^u$. Then, furthermore*

$$r_0^t W(t) = \sum_{k \geqq 0} 1_{\{T_k \leqq t < T_{k+1}\}} w_k(t, T_0, \cdots, x_{T_k}). \tag{5.43}$$

Remark. Equation (5.41) will indeed give a predictable $u^*(t, \omega)$, since in the left-closed stochastic interval $T_k(\omega) \leqq t < T_{k+1}(\omega)$ (in $R_+ \times \Omega$), its solution is a function of x_{T_i}, T_i, $i = 0, 1, \cdots, k-1$ and of T_k.

We are now in a position to compare our results with those of Rishel [21]. First of all his model of the dynamics of the jump processes is a special case of the one used in Theorem 5.3. Secondly, the observation σ-fields, $\mathcal{Y}_t$, that he permits are much more general even than those of Theorem 5.1. For he only requires that $(\mathcal{Y}_t)$ be "locally increasing", i.e., for each t there is $\eta > 0$ so that $\mathcal{Y}_t \subset \mathcal{Y}_s$ for $s \in [t, t+\eta]$. Thirdly, the structure of the cost functional is the same as the one used here. For an admissible control u let $J_t(u) = E\{\text{cost incurred in } [t, T] \text{ using } u \,|\, \mathcal{X}_t\}$. The process $(J_t(u), \mathcal{X}_t, \mathcal{P}^u)$ can be expressed as

$$J_t(u) = \sum_{k \geqq 0} 1_{\{T_k \leqq t < T_{k+1}\}} j_k^u(t, T_0, \cdots, T_k, x_{T_k})$$

as in (5.27). Rishel derives differential equations for the j_k^u similar to our equation (5.39). Finally he compares $J_t(u^*)$, for an optimal control u^* with $J_t(v)$ where v is a control obtained from u^* by a local perturbation. The necessary condition $E[J_t(u^*) - J_t(v) \,|\, \mathcal{Y}_t] \leqq 0$ is translated into a necessary condition on the j_k^u (see [21, Thm. 6]). Since u^* is compared with controls obtained by a local perturbation, therefore these necessary conditions are weaker as compared say with Theorem 5.3 above.

5.5. Markovian case. To simplify the discussion in this section we suppose $T < \infty$ and $r \equiv 1$. Now suppose $y_t \equiv x_t$ and suppose as in (5.40) that

$$\tilde{P}_u^x(dz, ds) = n(dz, s, u_s) \lambda_s \, ds,$$

where n and λ have the form

$$n(dz, s, u, \omega) = n(dz, s, u, x_{s-}(\omega)), \tag{5.44}$$

$$\lambda(s, \omega) = \lambda(s, x_{s-}(\omega)). \tag{5.45}$$

Similarly, suppose that in the cost functional (5.5) we have

$$c(z, s, u, \omega) = c(z, s, u, x_{s-}(\omega)), \tag{5.46}$$

$$J_T(\omega) = J_T(x_T(\omega)). \tag{5.47}$$

Next, call $u \in \mathcal{U}$ a *Markovian control* if u_t is of the form $u_t(\omega) = u_t(x_{t-}(\omega))$. Let $\mathcal{U}^M$ be the set of Markovian controls. Blumenthal and Getoor [33, p. 63ff] have shown that $(x_t, X_t, \mathcal{P}^u)$ is a Hunt process under these conditions. Essentially, this is a quasi-left continuous, strong Markov process. The martingale representation results of Kunita–Watanabe then apply, and the integrand can be written as

$$\beta(z, s, \omega) = \beta(z, s, x_{s-}(\omega)).$$

With these assumptions it is reasonable to expect that a Markovian control is optimal in the class $\mathcal{U}$, i.e.,

$$\Lambda_{u \in \mathcal{U}^M} J(u) = \Lambda_{u \in \mathcal{U}} J(u), \tag{5.48}$$

and it will then follow that the (complete information) value function $W(t)$ has a representation $W(t, \omega) = w(t, x_t(\omega))$.

To prove this assertion we begin by defining the Markovian value function. For u, v in $\mathcal{U}^M$, as before let

$$\begin{aligned} \psi(u, v, t) &= E^{(u,v,t)}\left\{\int_Z \int_t^T c(z, s, v_s)\tilde{P}_v^x(dz, ds) + J_T \,|\, \mathcal{F}_t^x\right\} \\ &= E^v\left\{\int_Z \int_t^T c(z, s, v_s)\tilde{P}_v^x(dz, ds) + J_T \,|\, x_t\right\} \\ &= \eta(v, t, x_t) \quad \text{say,} \\ V(t, x_t) &= \Lambda_{v \in \mathcal{U}^M} \eta(v, t, x_t). \end{aligned} \tag{5.49}$$

To show that $V(t, x_t) = W(t)$ it is enough, as we will see, to prove a version of the optimality principle, Theorem 4.1, for the function V and $u \in \mathcal{U}^M$. It is here that we face a difficulty because the proof of Theorem 4.1 relies on Lemma 4.1 and in the proof of the latter critical use is made of the fact that u_t can depend arbitrarily on $\mathcal{Y}_t$ and that these are *increasing*; whereas here u_t can depend arbitrarily only on $\sigma(x_{t-})$ and these are certainly *not* increasing.

We shall circumvent this difficulty by assuming that it is possible to approximate the time-continuous optimal control problem by a time-discrete problem. Since for the latter an optimality principle is available, we will be able to obtain such a result for the original problem.

For each $t \in I$ and integer N let $t = t_1 < t_2 < \cdots < t_{2^N} = T$ be equally spaced instances of time and let $\mathcal{U}_t^N$ be the set of all (u_s), $s \geqq t$ of the form

$$u_s(\omega) = u_s(x_{t_k}(\omega)) \quad \text{for } t_k < s \leqq t_{k+1}.$$

We impose the following assumption of approximation.

A. For all $\varepsilon > 0$, $t \in I$, $u \in \mathcal{U}^M$ there exists K such that for all $N \geqq K$ there exists $v \in \mathcal{U}_t^N$ with $\eta(v, t, x_t) \leqq \eta(u, t, x_t) + \varepsilon$.

THEOREM 5.4. *Suppose* (5.44)–(5.47). *Then for* $t_1 \leqq t_2$ *in* I *and* $u \in \mathcal{U}^M$ *we have*

$$V(t_1, x_{t_1}) \leqq E^u\left[\int_Z \int_{t_1}^{t_2} c(z, s, u_s)\tilde{P}_u^x(dz, ds) \,|\, x_{t_1}\right] + E^u[V(t_2, x_{t_2}) \,|\, x_{t_1}], \tag{5.50}$$

$$V(T, x_T) = J_T(x_T). \tag{5.51}$$

If equality holds in (5.50) *for* $u = u^*$ *then* u^* *is optimal in* $\mathcal{U}^M$. *Finally if* A *holds, then this condition is necessary for optimality.*

Proof. For $u \in \mathcal{U}^M$ we have

$$V(t_1, x_{t_1}) \leqq E^u\Big[\int_Z \int_{t_1}^{t_2} c(z, s, u_s)\tilde{P}_u^x(dz, ds)\,|\,x_{t_1}\Big] + \Lambda_{v \in \mathcal{U}^M} E^u[\eta(v, t_2, x_{t_2})\,|\,x_{t_1}]$$

with equality if and only if u is optimum. Since obviously

$$\Lambda_{v \in \mathcal{U}^M} E^u[\eta(v, t_2, x_{t_2})\,|\,x_{t_1}] \geqq E^u[\Lambda_{v \in \mathcal{U}^M} \eta(v, t_2, x_{t_1})\,|\,x_{t_1}], \tag{5.52}$$

therefore the sufficiency part of the assertion follows. Now suppose assumption A holds. To prove the final assertion it is enough to show that the reverse inequality holds in (5.52). Fix $\varepsilon > 0$. We must show that there is $v \in \mathcal{U}^M$ so that

$$E^u[\eta(v, t_2, x_{t_2})\,|\,x_{t_1}] \leqq E^u[\Lambda_{v \in \mathcal{U}^M} \eta(v, t_2, x_{t_2})\,|\,x_{t_1}] + \varepsilon. \tag{5.53}$$

Using assumption A, we can find N such that

$$E^u[\Lambda_{v \in \mathcal{U}_{t_2}^N} \eta(v, t_2, x_{t_2})\,|\,x_{t_1}] \leqq E^u\Big[\Lambda_{v \in \mathcal{U}^M} \eta(v, t_2, x_{t_2})\,|\,x_{t_2}\Big] + \frac{\varepsilon}{2}. \tag{5.54}$$

Next, we apply discrete backwards dynamic programming to obtain $v' \in \mathcal{U}_{t_2}^N$ so that

$$\eta(v', t_2, x_{t_2}) \leqq \Lambda_{v \in \mathcal{U}_{t_2}^N} \eta(v, t_2, x_{t_2}) + \frac{\varepsilon}{2}. \tag{5.55}$$

Corresponding to this $v' \in \mathcal{U}_{t_2}^N$, there exists a $v \in \mathcal{U}^M$ such that $v_s = v'_s$, $s \geqq t$, so that

$$\eta(v, t_2, x_{t_2}) \leqq \Lambda_{u \in \mathcal{U}_{t_2}^N} \eta(u, t_2, x_{t_2}) + \frac{\varepsilon}{2} \tag{5.56}$$

rewriting (5.55). From (5.54)–(5.56) we see that v satisfies (5.53). The assertion is proved. □

Now let $V_t = V(t, x_t)$. Fix $u \in \mathcal{U}^M$ and consider the process $(V_t, \mathcal{X}_t, \mathcal{P}^u)$. Then, using the same argument which led to (4.5), we obtain the representation

$$V_t = J_M - A_0^t(u) + m^V(u, t), \tag{5.57}$$

where $J_M = \inf\{J(u)\,|\,u \in \mathcal{U}^M\}$, $m^V(u) \in \mathcal{M}(\mathcal{X}_t, \mathcal{P}^u)$, and for $t_1 \leqq t_2$

$$A_{t_1}^{t_2} = \text{weak} \lim_{h \to 0} \int_{t_1}^{t_2} \frac{1}{h} E^u[V_s - V_{s+h}\,|\,\mathcal{X}_s]\,ds.$$

By the Hunt property $E^u[V_s - V_{s+h}\,|\,\mathcal{X}_s] = E^u[V_s - V_{s+h}\,|\,x_s]$ so that $A_{t_1}^{t_2}$ is measurable with respect to $\mathcal{X}_{t_1}^{t_2} = \sigma(x_s; t_1 \leqq s \leqq t_2)$. (This implies also that $m(u, t_2) - m(u, t_1)$ is $\mathcal{X}_{t_1}^{t_2}$-measurable; i.e., $m(u)$ is an additive functional of the

Markov process $(x_t, \mathscr{X}_t, \mathscr{P}^u)$.) We therefore obtain the following version of Theorem 4.2.

THEOREM 5.5. *Suppose* (5.44)–(5.47). *There exists a constant* J_M *and for every* $u \in \mathscr{U}^M$ *there exists a predictable process* $(A_0^t(u)) \in \mathscr{A}(\mathscr{X}_t, \mathscr{P}^u)$ *such that*

$$E^u A_0^T(u) = J_M - E^u J_T, \tag{5.58}$$

and such that for $t_1 \leqq t_2$, $A_{t_1}^{t_2}(u)$ *is* $\mathscr{X}_{t_1}^{t_2}$*-measurable and*

$$E^u\left[-A_{t_1}^{t_2}(u) + \int_Z \int_{t_1}^{t_2} c(z, s, u_s)\tilde{P}_u^x(dz, ds) \,|\, x_{t_1}\right] \geqq 0 \quad \text{a.s. } \mathscr{P}^u. \tag{5.59}$$

Suppose equality holds in (5.59) *for some* $u = u^*$ *in* $\mathscr{U}^M$. *Then* u^* *is optimal in* $\mathscr{U}^M$, $J(u^*) = J_M$ *and*

$$V_t = E^{u^*}[A_t^T(u^*) + J_T \,|\, x_t] \quad \text{a.s. } \mathscr{P}^{u^*}. \tag{5.60}$$

Finally if A *holds, then this condition is necessary for optimality.*

We return to the representation (5.57). Since $m^V(u)$ is an additive functional it can be represented as (applying results of [32, § 5])

$$m^V(u, t) = \int_Z \int_0^t \beta^u(z, s)[P^x(dz, ds) - \tilde{P}^x(dz, ds)]$$

where $\beta^u \in L^1_{\text{loc}}(\tilde{P}_u^x)$ is of the form

$$\beta^u(z, s, \omega) = \beta^u(z, s, x_{s-}(\omega)).$$

As before (cf. (5.17)) let

$$\begin{aligned} \bar{A}_0^t(u) &= A_0^t(u) + \int_Z \int_0^t \beta^u(z, s)\tilde{P}_u^x(dz, ds) \\ &= A_0^t(u) + \int_Z \int_0^t \beta^u(z, s) n(dz, s, u_s)\lambda_s \, ds \end{aligned}$$

and we may conclude again (see (5.19), (5.20)) that for u, v in $\mathscr{U}^M$

$$\bar{A}_0^t = \bar{A}_0^t(v) = \bar{A}_0^t \quad \text{say,}$$

$$\int_Z \int_0^t \beta^u(z, s)P^x(dz, ds) = \int_Z \int_0^t \beta^v(z, s)P^x(dz, ds).$$

Furthermore there exists a predictable process (α_t) such that

$$\bar{A}_0^t = \int_0^t \alpha_s \lambda_s \, ds.$$

But $\bar{A}_{t_1}^{t_2}$ is $\mathscr{X}_{t_1}^{t_2}$-measurable and $\lambda_t(\omega) = \lambda(t, x_{t-}(\omega))$ by (5.45). Hence α_t is of the form $\alpha_t(\omega) = \alpha(t, x_{t-}(\omega))$. The local version of (5.59) now becomes (cf. (5.26))

$$-\alpha(t, x_{t-}(\omega)) + \min_{u \in \mathscr{U}} \int_Z [\beta(z, t) + c(z, t, u)]n(dz, t, u) = 0 \tag{5.61}$$

and the minimum is achieved at $u^*(t, x_{t-}(\omega))$ for almost all (t, ω) with respect to $d\Lambda \times d\mathscr{P}^{u^*}$ measure. But from (5.61) it is evident that u^* is now an optimal control

in the class $\mathcal{U}$ of all controls and not just Markovian controls. It follows then that $V(t, x_t) = W(t)$. Theorem 5.3 simplifies as follows.

THEOREM 5.6. *Suppose* (5.44)–(5.47). *Suppose there exist* $u^* \in \mathcal{U}^M$ *and a measurable function* $V(t, x)$ *which is absolutely continuous in* t, *so that*

$$\frac{\partial V}{\partial t}(t, x_{T_k}) + \min_{u \in \mathcal{U}} \lambda(t, x_{T_k-}) \int_Z [V(t, z) - V(t, x_{T_k}) + c(t, z, u)] n(dz, t, u) = 0, \tag{5.62}$$

for $T_k \leqq t < T_{k+1}$,

$$V(T, x_{T_k}) = J_T(x_T(\omega)), \quad \text{for } T_k \leqq T < T_{k+1}, \tag{5.63}$$

and the minimum is achieved at $u^*(t, x_{t-}(\omega))$ *a.s. with respect to* $\mathcal{P}^{u^*}$ *measure. Then* u^* *is optimal in the class of all control laws, and furthermore* $V(t, x_t) = W(t)$. *Finally if* A *holds, then this condition is also necessary for optimality.*

We can compare the result above with the main result of Stone [24, Thm. 4.5]. Essentially our result is a special case of his result since the latter applies to *semi*-Markov processes and not just to Markov processes as we have insisted. Of course it is possible to obtain his result from ours by imbedding the semi-Markov process into a Markov process (see [24, Thm. 2.1]). One difference may be worth noting. Stone only considers controls which give rise to Markov processes with stationary transition probabilities; he is then able to use the infinitesimal generator of the process as the main tool of analysis. The martingale theoretic approach followed here permits us to dispense with the stationarity restriction.

6. Examples. We use the results derived above to solve some simple optimal control problems.

6.1. Queues.

(i) The simplest case imaginable is that of controlling the rate of a counting process over the interval $I = [0, T]$, $T < \infty$. Z is then the set of nonnegative integers. Let $U = [a, b]$ where $b > a \geqq 0$. Let $P^x(t)(\omega) =$ number of jumps of $x_s(\omega)$ in the interval $[0, t]$. Suppose $y_t \equiv x_t$, and for $u \in \mathcal{U}$ let

$$\tilde{P}^x_u = u_t$$

so that the controller can vary the rate of the process (x_t) to any desired value in $[a, b]$. Suppose $r_0^t \equiv 1$ and $c(t) = c(t, u_t, x_{t-})$, $J_T = J(x_{T-})$. Then the optimal control must be Markovian. The optimality criterion becomes

$$0 = \min_{a \leqq u \leqq b} \left\{ \frac{\partial V}{\partial t}(t, z) + [V(t, z+1) - V(t, z) + c(t, z, u)]u \right\}, \qquad z = 0, 1, \cdots, \tag{6.1}$$

with the boundary condition

$$V(T, z) = J(z). \tag{6.2}$$

One possible problem of this type, suggested by Professor D. Snyder, related to minimizing the damage to a sample in electron microscopy, is to seek u to maximize $\mathcal{P}^u(x_T = k)$ where k is a fixed integer. Since

$$\mathcal{P}^u(x_T = k) = E^u(1_{\{x_T = k\}}),$$

and since we are maximizing, the optimality criterion can be rewritten (setting $\hat{V} = -V$) as

$$0 = \max_{a \leqq u \leqq b} \left\{ \frac{\partial \hat{V}}{\partial t}(t, z) + [\hat{V}(t, z+1) - \hat{V}(t, z)]u \right\}, \tag{6.3}$$

$$\hat{V}(T, z) = 1_{\{z=k\}}. \tag{6.4}$$

Equation (6.3) gives the optimal Markovian control,

$$u^*(t, z) = \begin{cases} b & \text{if } \hat{V}(t, z+1) - \hat{V}(t, z) > 0, \\ a & \text{if } \hat{V}(t, z+1) - \hat{V}(t, z) < 0, \end{cases}$$

which upon substitution in (6.3) yields

$$0 = \frac{\partial \hat{V}}{\partial t}(t, z) + b \max[\hat{V}(t, z+1) - \hat{V}(t, z), 0] + a \min[\hat{V}(t, z+1) - \hat{V}(t, z), 0]$$

for $0 \leqq t \leqq T$, and $z = 0, 1, 2, \cdots$, and which can be solved recursively.

The closed loop optimal control $u^*(t, x_{t-})$ is given by

$$u^*(t, x_t) = a \cdot I_{\{\hat{V}(t, x_{t-}+1) - \hat{V}(t, x_{t-}) \leqq 0\}} + b \cdot I_{\{\hat{V}(t, x_{t-}+1) - \hat{V}(t, x_{t-}) > 0\}}$$

which is a predictable process ($\hat{V}(t, z)$ are deterministic functions.)

Remark 6.1. Suppose there were a second, independent Poisson process (N_t) and suppose the objective was to maximize

$$\mathcal{P}^u(x_T + N_T = k).$$

Suppose (N_t) cannot be observed or controlled, whereas (x_t) can, just as before. This is now a problem with partial information. Nevertheless, it is easy to see the optimality equation (6.3) is still valid here, with the boundary condition (6.4) replaced by

$$V(T, z) = \begin{cases} e^{-T} \dfrac{T^{k-i}}{(k-i)!} & \text{for } z = i, i = 0, \cdot, \cdots, k, \\ 0 & \text{for } z > k. \end{cases}$$

This follows from the fact that

$$\mathcal{P}^u(x_t + N_t = k) = \sum_{i=0}^{k} \mathcal{P}^u(x_t = i)\mathcal{P}(N_t = k - i).$$

Note that the problem becomes much more complicated as soon as x_t and N_t are dependent. The problem is then one of partial information, $V(u, t)$ depends on past controls, and Markov controls are not necessarily optimal. Solving the optimality equations of section then requires an unreasonable amount of calculations (as can be expected for a "dual optimal control" problem).

(ii) Consider the simplest problem of controlling a queue length by varying the service rate (or number of servers). The (x_t) process is now the queue length (Q_t) defined as follows. Let (A_t), (D_t) respectively represent the arrivals and departures. Then Q_t is defined by

$$Q_t = A_t - \int_0^t 1_{\{Q_{t-}>0\}}\, dD_s,$$

where the integrand manifests the fact that no departure can occur when the queue is empty. Now suppose that the arrival rate is a constant λ which cannot be controlled, but that the departure rate can be set to any $u \in U = \{0, 1, \cdots, N\}$. Then, in the notation of § 5.5,

$$\tilde{P}_u^x(dz, dt, Q_{t-}) = 1_{\{Q_{t-}+1\in dz\}}\lambda\, dt + 1_{\{Q_{t-}-1\in dz\}}1_{\{Q_{t-}>0\}}u_t\, dt$$

where the first term on the right corresponds to a jump of $+1$ in Q and the second term corresponds to a jump of -1.

Suppose the cost function is of the form

$$J(u) = E^u\left[\int_0^T c(s, u_s, Q_{s-})\, ds + f(Q_{T-})\right].$$

Then there is a Markovian optimal control and the value function $V(t, Q)$ satisfies

$$0 = \min_{u\in U}\left\{\frac{\partial V}{\partial t}(t, Q) + c(t, u, Q) + [V(t, Q+1) - V(t, Q)]\lambda \right. \tag{6.5}$$
$$\left. + [V(t, Q-1) - V(t, Q)]1_{\{Q>0\}}u\right\},$$

with boundary condition

$$V(T, Q) = f(Q). \tag{6.6}$$

Next, suppose that the cost is a linear function of the total waiting time and the total service time, i.e.,

$$c(t, u, Q) = au + Q, \quad f = 0,$$

where $a > 0$ is a constant. Hence from (6.5) the optimal control is "bang-bang". It can be exactly specified as

$$u^*(t, Q_{t-}) = \begin{cases} N1_{\{Q_{t-}>0\}} & \text{for } t\in[0, T-a], \\ 0 & \text{for } t\in[T-a, T]. \end{cases}$$

This follows because in the interval $[T-a, T]$,

$$V(t, Q) = (T-t)Q + \frac{\lambda}{2}(T-t)^2$$

so that

$$V(t, Q-1) - V(t, Q) = -(T-t) \tag{6.7}$$
$$< a, \quad \text{for } t\in(T-a, T)$$

and the fact that $V(t, Q-1) - V(t, Q)$ must increase with t.

Remark. Since $\tilde{P}_u(\{+1\}, t) = \int_0^t u_s \cdot ds = \int_0^t u_{s-} \cdot ds$, it is irrelevant in this example whether controls are chosen predictable or not. However $u^*(t, Q_{t-})$ of (6.7) is predictable in accordance with § 5.5.

(iii) A somewhat more complicated problem is that in which only one of two queues can be served at any given time (e.g. traffic light at an intersection). Each of the two queues, Q_t^1, Q_t^2 say, are described as above, and the possible values of the pair of service rates $u = (u^1, u^2) \in U = \{(0, 1), (1, 0)\}$.

$$
\begin{aligned}
0 = \min_{u \in U} \{ & c(t, u^1, u^2, Q^1, Q^2) + \frac{\partial V}{\partial t}(t, Q^1, Q^2) \\
& + [V(t, Q^1 + 1, Q^2) - V(t, Q^1, Q^2)]\lambda^1 \\
& + [V(t, Q^1, Q^2 + 1) - V(t, Q^1, Q^2)]\lambda^2 \\
& + [V(t, Q^1 - 1, Q^2) - V(t, Q^1, Q^2)]1_{\{Q^1 > 0\}} u^1 \\
& + [V(t, Q^1, Q^2 - 1] - V(t, Q^1, Q^2)]1_{\{Q^2 > 0\}} u^2 \}
\end{aligned} \tag{6.8}
$$

with the boundary condition

$$V(T, Q^1, Q^2) \equiv 0.$$

We are unable to derive an explicit form for the optimal control.

6.2. Investment. An example of a jump process with an infinite number of sizes of jump is the following. Assume that there are N stocks with $\pi_i(t)$ as the price of the ith stock. The ith price changes at random times with a rate λ_i and at these times the price changes from $\pi_i(t-)$ to $\pi_i(t) = \pi_i(t-) + \alpha_i(t)\pi_i(t-)$ where $\alpha_i(t) \geqq -1$ is a random variable with distribution function $n_i(d\alpha_i, t)$. Then an investor with wealth $K(t)$, who has invested a fraction $k_i(t)$ in the stock i, faces the accounting equation

$$dK_t = \sum_{i=1}^{N} k_i(t) K_{t-} \frac{d\pi_i(t)}{\pi_i(t-)}, \qquad \sum_{i=1}^{N} k_i(t) = 1. \tag{6.9}$$

(K_t) is therefore a jump process which has jumps of size $k_i K \alpha_i$ occurring at rates λ_i. Here, as before, the probability measure of the "state" process (K_t) depends on the "control" $(k_i(t))$, $i = 1, \cdots, N$. In a simpler setting it has been shown [30] that the problem of choosing $k = \{(k_i(t))\}$ to maximize $E^k(J(K_T))$, where J is the utility of wealth, can be reduced to a static optimization problem. We solve here a more general problem. Suppose the investor also has a wage income $y_t\, dt$ in $[t, t + dt]$, beyond his control, and can consume an amount $c_t\, dt$ of his wealth in the interval $[t, t + dt]$, where $c_t \geqq 0$ can be chosen freely and is therefore additional control. Then (6.9) is replaced by

$$dK_t = (y_t - c_t)\, dt + \sum_{i=1}^{N} k_i(t) K_{t-} \frac{d\pi_i(t)}{\pi_i(t-)}. \tag{6.10}$$

The investor's objective is to maximize

$$E^u\left[\int_0^T J(t, c_t)\, dt + J_T(K_T)\right] \tag{6.11}$$

where $u=\{(k_1(t)),\cdots,(k_N(t)),(c(t))\}$ is the control, J denotes utility from consumption and J_T denotes utility from terminal wealth. In this formulation (K_t) is no longer a jump process, because of the first term in the right-hand side of (6.10). A referee has pointed out that it is possible to regard (K_t) as a jump process by taking Z to be a space of continuous functions. The value between two successive jump times would then be the trajectory of (K_t) between these two instants of time. However, if we assume that the rate process $(\lambda_i(t))$ and the distributions (n_i) depend only upon K_{t-}, then $(K_t, \mathcal{P}_u)$ is a still a Markov process for a Markov control u, and it is easier to apply well-known results of Markov process theory. The infinitesimal generator $\mathcal{L}^u(\hat{V})$ of the value function

$$\hat{V}(t, K_t)=\sup_u E^u\left[\int_t^T J(t, c_t)\,dt+J_T(K_T)\right]$$

is, from (6.10),

$$\begin{aligned} & w\cdot\lim_{h\to 0}\frac{1}{h}\{E^u[\hat{V}(t+h, K_{t+h})\,|\,K_t]-\hat{V}(t, K_t)\} \\ & =\frac{\partial\hat{V}}{\partial t}(K_t, t)+\frac{\partial\hat{V}}{\partial K}(K_t, t)(y_t-c_t) \\ & \quad+\sum_{i=1}^N\lambda_i(t, K_t)\int_{-1}^{\infty}[\hat{V}(t, [1+\alpha_i k_i(t)]K_t)-\hat{V}(t, K_t)]n_i(d\alpha_i, t, K_t). \end{aligned} \tag{6.12}$$

(We could have permitted a Brownian motion component in (6.10) as studied in [17]).

The optimality criterion is

$$\begin{aligned} 0=\max_{u\in U}\Big\{ & J(t, c_t)+\frac{\partial\hat{V}}{\partial t}(t, K)+(y_t-c_t)\frac{\partial\hat{V}}{\partial K}(t, K) \\ & +\sum_{i=1}^N\lambda_i(t, K)\int_{-1}^{\infty}[\hat{V}(t, [1+\alpha_i k_i]K)-\hat{V}(t, K)]n_i(d\alpha_i, t, K)\Big\} \end{aligned} \tag{6.13}$$

with the boundary condition

$$\hat{V}(T, K)=J_T(K). \tag{6.14}$$

We can solve (6.13), (6.14) for the following special case. Assume $y_t\equiv 0$, $J(t, c)=c^\gamma/\gamma$, $J_T(K)=a(K^\gamma/\gamma)$, where $a>0$ and $0<\gamma\leqq 1$ are constants, and λ_i, n_i independent of K and t. Then (6.13), (6.14) have the following solution

$$\hat{V}(t, K)=f(t)\frac{K^\gamma}{\gamma},\qquad 0\leqq t\leqq T,\quad K\geqq 0,$$

where

$$f(t)=\left[\left(\frac{1-\gamma}{A}+a^{1/(1-\gamma)}\right)\cdot\exp\left(-A\cdot\frac{T-t}{1-\gamma}\right)+\frac{\gamma-1}{A}\right]^{1-\gamma}$$

The constant A and the optimal control are given by

$$c_t^* = \frac{K_t}{f(t)^{1/(1-\gamma)}};$$

$(k_i^*(t))$ are optimal solutions of the static problem

$$\max_{k_i \geq 0, \Sigma k_i = 1} \left\{ \sum_{i=1}^{N} \lambda_i \int_{-1}^{\infty} [(1+k_i\alpha_i)^\gamma - 1] n_i(d\alpha_i) \right\} = A.$$

Acknowledgment. This paper was extensively rewritten in the light of a thorough critique of the previous version by Dr. Pierre Brémaud. We are very grateful to him for his interest. The many constructive suggestions of an anonymous referee are also gratefully acknowledged.

REFERENCES

[1] V. BENEŠ, *Existence of optimal strategies based on specified 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] J. M. BISMUT, *Integrales convexes et probabilites*, INF. 7201/72001, Institut de Recherche d'Informatique et d'Antomatique, Rocquencourt, France, 1972.

[4] R. BOEL, *Optimal control of jump processes*, Memo ERL M-448, Electronics Research Laboratory, University of California, Berkeley, September 1974.

[5] R. BOEL, P. VARAIYA AND E. WONG, *Martingales on jump processes I: Representation results*, Memo ERL M-407, Electronics Research Laboratory, University of California, Berkeley, September 1973; this Journal, 13 (1975), pp. 999–1021.

[6] ———, *Martingales on jump processes II: Applications*, Memo ERL M-409, Electronics Research Laboratory, University of California, Berkeley, December 1973; this Journal, 13 (1975), pp. 1022–1061.

[7] C. S. CHOU AND P. A. MEYER, *Sur la représentation des martingales comme intégrales stochastiques dans les processus ponctuels*, Séminaire de Probabilités VIII, Lecture Notes in Mathematics, Springer-Verlag, Berlin, to appear.

[8] M. H. A. DAVIS, *The representation of martingales of jump processes*, this Journal, 14 (1976), pp. 623–638.

[9] M. H. A. DAVIS AND P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.

[10] ———, *Information states for linear stochastic systems*, J. Math. Anal. Appl., 37 (1972), pp. 384–402.

[11] T. E. DUNCAN AND P. VARAIYA, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.

[12] E. B. DYNKIN, *Markov Processes*, vol. I, Academic Press, New York, 1965.

[13] R. J. ELLIOTT, *Martingales of a jump process and absolutely continuous changes of measure*, presented at Symp. on Stochastic Systems, University of Kentucky, Lexington, 1975.

[14] M. FUJISAKI, G. KALLIANPUR AND H. KUNITA, *Stochastic differential equations for the non-linear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–40.

[15] J. JACOD, *Multivariate point processes: Predictable projection, Radon–Nikodym derivatives, representation of martingales*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 31 (1975), pp. 235–253.

[16] H. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.

[17] R. MERTON, *Optimum consumption and portfolio rules in a continuous time model*, J. Econom. Theory, 3 (1971), pp. 373–413.

[18] P. A. MEYER, *Probability and Potentials*, Blaisdell, Waltham, Mass., 1966.

[19] ———, *Intégrales stochastiques*, Séminaire de Probabilités I, Lecture Notes in Mathematics 39, Springer-Verlag, Berlin, 1967.

[20] R. RISHEL, *Necessary and sufficient dynamic programming conditions for continuous-time stochastic optimal control*, this Journal, 8 (1970), pp. 559–571.

[21] ———, *A minimum principle for controlled jump processes*, Control Theory, Numerical Methods and Computer Systems Modelling, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Economics and Mathematical Systems, 107, Springer-Verlag, Berlin, 1975, pp. 493–508.

[22] ———, *Dynamic programming and minimum principles for systems with jump Markov disturbances*, this Journal, 13 (1975), pp. 338–371.

[23] P. A. SAMUELSON, *Proof that properly discounted present values of assets vibrate randomly*, Bell J. of Econom. and Management Sci., 4 (1973), pp. 369–374.

[24] L. STONE, *Necessary and sufficient conditions for optimal control of semi-Markov jump processes*, this Journal, 11 (1973), pp. 187–201.

[25] C. STRIEBEL, *Martingale conditions for the optimal control of continuous time stochastic systems*, 1974, to appear.

[26] R. L. STRATONOVICH, *Conditional Markov Processes and their Application to the Theory of Optimal Control*, American Elsevier, New York, 1968.

[27] D. SWORDER, *Feedback control of a class of linear systems with jump parameters*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 9–14.

[28] J. VAN SCHUPPEN, *Estimation theory for continuous time processes, a martingale approach*, Memo ERL M-405, Electronic Research Laboratory, University of California, Berkeley, September 1973.

[29] J. VAN SCHUPPEN AND E. WONG, *Transformation of local martingales under a change of law*, Electronics Research Laboratory, Memo ERL M-385, University of California, Berkeley, May 1973; Ann. Probability, 2 (1974), pp. 879–888.

[30] P. VARAIYA, *The martingale theory of jump processes*, IEEE Trans. Automatic Control, AC-20 (1975), pp. 34–41.

[31] C. DELLACHERIE, *Capacités et Processus Stochastiques*, Springer-Verlag, Berlin, 1972.

[32] H. KUNITA AND S. WATANABE, *On square integrable martingales*, Nagoya Math. J., 30 (1967), pp. 209–245.

[33] R. M. BLUMENTHAL AND R. K. GETOOR, *Markov Processes and Potential Theory*, Academic Press, New York, 1968.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

CAUSAL REALIZATION FROM INPUT-OUTPUT PAIRS*

WILLIAM A. PORTER†

Abstract. Consider the finite subset $\{(x_i, y_i)\}$ of input-output pairs. A common design objective is to specify a system, S, such that $y_i = Sx_i$ holds on the set in question. Moreover, S should also be well behaved on a larger input space.

This rudimentary problem is typical of code block detectors, data transmission networks, computer controllers and with some refinement, can be viewed as a prototype problem for control compensator design. We note also that it is related to system identification. Indeed, having observed the input-output pairs any construction of S is a viable form of system identification.

In a recent study, the author solved a synthesis problem of the above type. In that study the input and output spaces are taken to be arbitrary Hilbert resolution spaces. A causal synthesis procedure was developed within this framework.

In the present study the linear solution is considered in more detail. We focus attention also on the Hilbert space $L_2(\nu)$. It is shown that the operator theoretic solution can be realized by a differential equation set of the form

$$\begin{aligned} \dot{z}(t) &= A(t)z(t) + b(t)x(t), \\ y(t) &= c(t)z(t), \qquad t \in \nu, \end{aligned} \tag{1}$$

where $\{A, b, c\}$ are explicitly specified from the input-output data.

1. Introduction.

Let X, Y be respective input and output spaces. Consider also the finite subsets $\{x_i\} \subset X$ and $\{y_i\} \subset Y$. A common design objective is to specify a system, S, such that $y_i = Sx_i$ holds on the sets in question. Moreover, S should also be well behaved on the larger input space X.

This rudimentary problem is typical of code block detectors, data transmission networks, computer controllers, and with some refinement can be viewed as a prototype problem for control compensator design. We note also that it is related to system identification. Indeed, having observed the input-output pairs $\{(x_i, y_i)\}$ any construction of S is a viable form of system identification.

The rudimentary problem takes on a more interesting form as constraints and detail are added. Typical constraints take the form of causality, linearity, multilinearity, stationarity and continuity requirements. Typical detail would include specifications of the input and output spaces and explicit choices for the input-output pairs. Of course, the choice of input-output pairs partially determines whether solutions exist satisfying the desired constraints.

In a recent study [1], the author solved, in rather general form, a synthesis problem of the above type. In that study the input and output spaces were taken to be arbitrary Hilbert resolution spaces. A causal synthesis procedure was developed within this framework. It was also determined when a linear causal solution exists and one such solution was provided. The entire development generalizes to Banach spaces without great difficulty.

* Received by the editors September 24, 1975.

† Computer, Information, and Control Program, University of Michigan, Ann Arbor, Michigan 48104. This research was sponsored in part by the United States Air Force Office of Scientific Research under Grant 73-2427.

In the present study the linear solution is considered in more detail. We focus attention also on the Hilbert space $L_2(\nu)$. It is shown that the operator theoretic solution of [1] can be realized by a differential equation set of the form

$$\begin{aligned} \dot{z}(t) &= A(t)z(t)+b(t)x(t), \\ y(t) &= c(t)z(t), \qquad t\in\nu, \end{aligned} \tag{1}$$

where $\{A, b, c\}$ are explicitly specified from the input-output data. During the development, an interesting connection with optimal control is exposed.

2. Some preliminaries. To facilitate the present development we shall restrict attention to $H=L_2(\nu)$, the usual Hilbert space of real square integrable functions. The inner product on $L_2(\nu)$ is denoted by $\langle\cdot,\cdot\rangle$ and $\{P^t\}$ is the family of projection operators defined by

$$(P^t x)(\beta)=\begin{cases} x(\beta), & \beta\leqq t, \\ 0, & \beta>t, \end{cases} \qquad t,\beta\in\nu.$$

Concerning the data set $\{(x_i, y_i)\}\subset H^2$, we shall assume first that the set $\{x_i\}$ is linearly independent. No loss of generality is incurred for if $x_k=\sum_{i\neq k}\alpha_i x_i$, then either $y_k\neq\sum_{i\neq k}\alpha_i y_i$, in which case a linear solution is not possible, or $y_k=\sum_{i\neq k}\alpha_i y_i$, in which case we delete (x_k, y_k) from the set and meet its constraint through linear extension.

It is convenient to make a stronger assumption, which can be removed later. We shall say that the set $\{x_i\}$ is *well-posed* provided $\{P^t x_i\}$ is linearly independent for all $t>0$, where $\nu=[0, d]$ (or $[0,\infty]$). We note, for instance, that the power functions $\{x_j(t)=t^{j-1}\}$ and the sinusoids $\{x_j(t)=\sin jt\}$ have this property.

In summarizing the synthesis procedure of [1], some definitions are helpful. For this we assume $\{x_i : 1,\cdots, n\}$ is well-posed and let

$$\eta_i[t,\beta]=\|P^t x_i\|^{-1}(P^t x_i)(\beta), \qquad i=1,\cdots,n. \tag{2}$$

We note that $\|\eta_i[t,\cdot\,]\|=1$ and that $\{\eta_i[t,\cdot\,]\}$ are linearly independent for all $t>0$. The $n\times n$ Grammian matrix, N, whose ijth element is computed by

$$N_{ij}(t)=\langle\eta_i[t,\cdot\,],\eta_j[t,\cdot\,]\rangle=\frac{\langle P^t x_i, x_j\rangle}{\|P^t x_i\|\cdot\|P^t x_i\|}, \tag{3}$$

is nonsingular for $t>0$ and we let $M(t)=N^{-1}(t)$. The row vector $\hat{y}(t)=(\hat{y}_1(t),\cdots,\hat{y}_n(t))$ is computed by

$$\hat{y}_j(t)=\|P^t x_j\|^{-1}y_j(t), \qquad j=1,\cdots,n. \tag{4}$$

Finally we define the column vector

$$\eta[t,\beta]=\operatorname{col}(\cdots,\eta_j[t,\beta],\cdots),$$

and construct the function

$$w(t,\beta)=\hat{y}(t)M(t)\eta[t,\beta], \qquad t,\beta\in\nu. \tag{5}$$

THEOREM 1 (see [1]). *If $\{x_i\}$ is well-posed, then*

$$(Su)(t) = \int_0^t w(t, \beta)u(\beta)\, d\beta.$$

In the following section we convert this result to differential equation form.

3. Computing *M*. The first matter of practical interest is the computation of $M(t) = N^{-1}(t)$. This fortunately has an elegant solution which is indirectly identified in the theorem.

THEOREM 2. *The matrix $N(t)$ of (3) is the self-adjoint solution of the differential equation*

$$\dot{N}(t) = \Pi(t) - \tfrac{1}{2}\{X(t)N(t) + N(t)X(t)\}, \qquad t \in \nu - \{0\}. \tag{6}$$

In this theorem $X(t)$ is the diagonal matrix

$$X(t) = \operatorname{diag}[\cdots, x_i^2(t)/\|P^t x_i\|^2, \cdots], \tag{7}$$

while $\Pi(t)$ is the symmetric matrix whose ijth element is given by

$$\Pi_{ij}(t) = x_i(t)x_j(t)/\|P^t x_i\| \cdot \|P^t x_j\|. \tag{8}$$

We note also that the initial condition on (6) can be taken from (3) at any $t \in \nu - \{0\}$. In fact N is known to be continuous [1] and hence $N(0) = \lim_{t \to 0^+} N(t)$ can also be used.

Proof. Equation (6) can be established by differentiation of (3). We shall not belabor these details but do list the following helpful intermediate identities $(d/dt)\langle P^t x, y\rangle = x(t)y(t)$:

$$\frac{d}{dt}\{\|P^t x\| \cdot \|P^t y\|\} = \frac{\|P^t y\|^2 x^2(t) + \|P^t x\|^2 y^2(t)}{2\|P^t x\| \cdot \|P^t y\|},$$

$$\frac{d}{dt}\left\{\frac{\langle P^t x, y\rangle}{\|P^t x\| \cdot \|P^t y\|}\right\} = -\frac{\langle P^t x, y\rangle}{\|P^t x\| \cdot \|P^t y\|}\left\{\frac{x^2(t)}{\|P^t x\|^2} + \frac{y^2(t)}{\|P^t y\|^2}\right\} + \frac{x(t)y(t)}{\|P^t x\| \cdot \|P^t y\|}.$$

The last identity is recognized as

$$\dot{N}_{ij}(t) = -\tfrac{1}{2}N_{ij}(t)\{X_{ii}(t) + X_{jj}(t)\} + \Pi_{ij}(t).$$

Using the diagonal form of $X(t)$ the theorem follows.

COROLLARY. *The matrix M is the self-adjoint solution of the equation*

$$\begin{aligned} &M(t) = \tfrac{1}{2}\{X(t)M(t) + M(t)X(t)\} - M(t)\Pi(t)M(t), \qquad t \in \nu, \\ &M(t') = N(t')^{-1}. \end{aligned} \tag{9}$$

This corollary is an immediate consequence of (6) and the identity $\dot{M} = -M(t)\dot{N}(t)M(t)$. When $\nu = [0, d]$, the choice $t' = d$ can be made with (9) solved in reverse time.

The computation of M through (9) provides an easy implementation of the operator S of Theorem 1. It suggests also that S might be realized in differential

form, and because (9) is of the Riccati type, an optimization problem may be related to our development.

4. The state variable realization. By way of notation we shall let $\Lambda(t)$ denote a diagonal matrix whose typical element is computed by

$$\Lambda_{ii}(t) = \|P^t x_i\|^{-1}, \qquad t \in \nu - \{0\},$$

and let $\mathbf{x}(t)$ denote the column vector formed by using the $\{x_i\}$ as entries. Our main result in this section is the following theorem.

THEOREM 3. *The equality* $z(t) = (Su)(t)$ *holds if and only if*

$$z(t) = \hat{y}(t)M(t)p(t),$$

$$\dot{p}(t) = -\tfrac{1}{2}X(t)p(t) + \Lambda(t)\mathbf{x}(t)u(t),$$

$$p(0) = 0.$$

Proof. Let us first demonstrate that the asserted equation is a realization of S. For this, note that the diagonal form of X means that

$$\Phi(t, \beta) = \exp\left\{-\frac{1}{2}\int_\beta^t X(s)\, ds\right\}.$$

However

$$\int_\beta^t \frac{x_j^2(s)}{\int_0^s x_j^2(\lambda)\, d\lambda}\, ds = \ln \|P^t x_j\|^2 - \ln \|P^\beta x_j\|^2,$$

and hence

$$\Phi(t, \beta) = \Lambda(t)\Lambda^{-1}(\beta), \qquad t, \beta \in \nu.$$

The integral form of the asserted differential system then is

$$z(t) = \int_0^t \hat{y}(t)M(t)\Phi(t, \beta)\Lambda(\beta)\mathbf{x}(\beta)u(\beta)\, d\beta$$

$$= \int_0^t \hat{y}(t)M(t)\Lambda(t)\mathbf{x}(\beta)u(\beta)\, d\beta, \qquad t \in \nu.$$

For $\beta < t$ it is obvious that

$$\eta[t, \beta] = \Lambda(t)\mathbf{x}(\beta),$$

which completes the argument.

To establish the converse we let Φ be unknown and equate

$$c(t)\Phi(t, \beta)b(\beta) = \hat{y}(t)M(t)\Lambda(t)\mathbf{x}(\beta), \qquad t \geqq \beta.$$

We choose

$$c(t) = \hat{y}(t)M(t),$$

with no loss of generality and then differentiate

$$\Lambda(t)\mathbf{x}(\beta) = \Phi(t, \beta)b(\beta),$$

resulting in

$$\dot{\Lambda}(t)\mathbf{x}(\beta) = A(t)\Phi(t, \beta)b(\beta) = A(t)\Lambda(t)\mathbf{x}(\beta).$$

It is easily shown that when $\{x_i\}$ is well-posed, a β can be found to arbitrarily orient $\mathbf{x}(\beta)$ and hence $\dot{\Lambda}(t) = A(t)\Lambda(t)$. However, differentiating Λ shows that, in fact,

$$\dot{\Lambda}(t) = -\tfrac{1}{2}X(t)\Lambda(t), \tag{10}$$

which completes the proof.

The realization of Theorem 3 is, of course, unique only to within a similarity change of variables. One such change of variable, namely $q(t) = M(t)p(t)$, is suggested by the form of the first realization. Noting that $\Pi(t) = \Lambda(t)\mathbf{x}(t)\mathbf{x}^*(t)\Lambda(t)$, it is easily verified that the following corollary holds.

COROLLARY. *The equality* $z(t) = (Su)(t)$ *holds if and only if*

$$\begin{aligned} &z(t) = \hat{y}(t)q(t), \qquad t \in \nu, \\ &\dot{q}(t) = \tfrac{1}{2}X(t)q(t) + M(t)\Lambda(t)\mathbf{x}(t)v(t), \qquad t \in \nu, \\ &v(t) = u(t) - \mathbf{x}^*(t)\Lambda(t)q(t), \qquad t \in \nu, \\ &q(0) = 0. \end{aligned}$$

This latter realization has an obvious feedback interpretation.

5. Relationship to optimal control. Equation (9) is recognized as the well-known Riccati equation which is an integral part of certain optimal control and filtering problems. In particular, if $\nu = [0, d]$, F, Q, R symmetric and $R(s) > 0$ on ν and if

$$\dot{q}(t) = A(t)q(t) + B(t)u(t), \qquad t \in \nu,$$

$$J(u) = [q(d), Fq(d)] + \int_0^d \{[q(s), Q(s)q(s)] + [u(s), R(s)u(s)]\}\, ds,$$

then the optimal control, u_0, minimizing J is given by

$$u_0(t) = -R^{-1}(t)B^*(t)K(t)q(t), \qquad t \in \nu,$$

where $K(t) = K^*(t)$ and

$$\begin{aligned} &\dot{K}(t) = -K(t)A(t) - A^*(t)K(t) + K(t)B(t)R^{-1}(t) \\ &\qquad\qquad B^*(t)K(t) - Q(t), \qquad t \in \nu, \\ &K(d) = F. \end{aligned} \tag{11}$$

Comparing (9) and (11) we see that

$$\begin{aligned} &A(t) = A^*(t) = -\tfrac{1}{2}X(t), \qquad t \in \nu, \\ &Q(t) = 0, \quad R(t) = I, \quad B(t)B^*(t) = \Pi(t), \quad t \in \nu. \end{aligned}$$

Noting that $\Pi(t) = \Lambda(t)\mathbf{x}(t)\mathbf{x}^*(t)\Lambda(t)$, we have

$$\begin{aligned} &\dot{r}(t) = -\tfrac{1}{2}X(t)r(t) + \Lambda(t)\mathbf{x}(t)u(t), \qquad t \in \nu, \\ (12) \qquad &u_0(t) = -\mathbf{x}^*(t)\Lambda(t)M(t)r(t), \qquad t \in \nu, \\ &J(u) = [r(d), M(d)r(d)] + \|u\|^2. \end{aligned}$$

The relationship between systems of (12) and Theorem 3 is quite easily summarized. As we have depicted in Fig. 1, an open loop plant characterized by $\{-\frac{1}{2}X(t); \Lambda(t)\mathbf{x}(t); M(t)\}$ is involved in both cases. With u arbitrary, zero initial state and output constraint $\hat{y}(t)$, the open loop plant is the system of Theorem 3. With the input driven by the indicated feedback law and arbitrary initial state the system is that of (12).

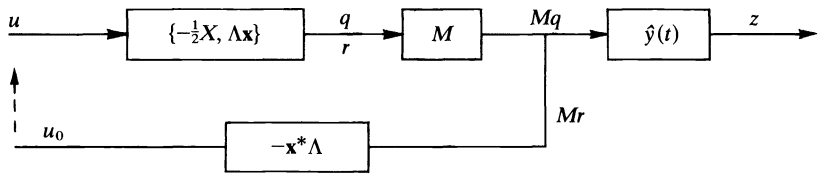


FIG. 1.

6. A numerical example. It is helpful to consider an example where a complete numerical solution is possible. For this we select $x_1(t) = 1$, $x_2(t) = t$, $x_3(t) = t^2$. It has been noted in [1] that N, and hence M are constant for this selection of inputs. Moreover [1] studies in some detail the characterizing properties of function classes which give rise to constant N.

Example 1. With $x_1(t) = 1$, $x_2(t) = t$, $x_3(t) = t^2$, the matrix N is constant with inverse (see [1]) given by

$$M = \begin{bmatrix} 9 & -12\sqrt{3} & 6\sqrt{5} \\ -12\sqrt{3} & 64 & -12\sqrt{15} \\ 6\sqrt{5} & -12\sqrt{15} & 36 \end{bmatrix}.$$

We note also that $\|P^t x_1\|^2 = t$, $\|P^t x_2\|^2 = t^3/3$, $\|P^t x_3\|^2 = t^5/5$. It is easily verified then that

$$X(t) = \frac{1}{t}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \qquad t > 0,$$

$$\Lambda(t)\mathbf{x}(t) = \frac{1}{\sqrt{t}}\begin{pmatrix} 1 \\ \sqrt{3} \\ \sqrt{5} \end{pmatrix}, \qquad t > 0,$$

and hence the system of Theorem 3 is completely specified once the output functions y_1, y_2, y_3 are chosen.

It is of interest to compute also the integral form of our system. For this it is only necessary to compute

$$w(t, \beta) = \hat{y}(t) M \Lambda(t) \mathbf{x}(\beta), \qquad t, \beta \in \nu.$$

Recalling that $\hat{y}_i(t) = y_i(t)/\|P^t x_i\|$, it is easily verified that

$$\begin{aligned} w(t, \beta) = {} & \frac{3y_1(t)}{t}[3 - 12\beta t^{-1} + 10\beta^2 t^{-2}] \\ & + \frac{12y_2(t)}{t^2}[-3 + 16\beta t^{-1} + -15\beta^2 t^{-2}] \\ & + \frac{30y_3(t)}{t^3}[1 - 6\beta t^{-1} + 6\beta^2 t^{-2}]. \end{aligned} \tag{13}$$

7. An alternative realization. The form of $w(t, \beta)$ above suggests a second differential realization that we now explore.

We note that in (13) the functions y_1, y_2, y_3 can be taken as arbitrary elements of $L_2(\nu)$. There are some practical limits, however, that we should note before continuing. First we note that with a simple integrator the input $x_1(t) = 1$ produces an output $y_1(t) = t$, and hence the ratio $y_1(t)/t$ in (13) resembles a comparative output/input measure. If our system is to be lowpass, then the ratios $y_i(t)/t^i$ should be bounded as $t \to 0$. If this does not hold, then the system will have direct transmission and an adjustment in the model is called for.

For convenience we introduce the notation

$$\bar{y}_i(t) = y_i(t)/t^i, \qquad i = 1, 2, 3,$$

and assume that $\bar{y}_i$ are bounded at $t \to 0$. Since $\beta \leqq t$, the kernel $w(t, \beta)$ is otherwise well-behaved and has the same continuity as the $\bar{y}_i$, whatever this may be.

Recall now the identity

$$\beta^n = [t - (t - \beta)]^n = \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} t^k (t - \beta)^{n-k}, \qquad n = 0, \cdots. \tag{14}$$

Using this identity on the terms in (13), we may express w as a function of $w(t, t - \beta)$, for instance,

$$3 - 12\beta t^{-1} + 10\beta^2 t^{-2} = 1 - 8t^{-1}(t - \beta) + 10t^{-2}(t - \beta)^2, \qquad \beta \leqq t.$$

Upon rearrangement, equation (13) takes the alternative form

$$\begin{aligned} w(t, \beta) = {} & [3\bar{y}_1(t) + 276\bar{y}_2(t) + 30\bar{y}_3(t)] \\ & + [-24\bar{y}_1(t) - 432\bar{y}_2(t) - 180\bar{y}_3(t)]t^{-1}(t - \beta) \\ & + [30\bar{y}_1(t) + 120\bar{y}_2(t) + 180\bar{y}_3(t)]t^{-2}(t - \beta)^2, \qquad 0 \leqq \beta \leqq t. \end{aligned} \tag{15}$$

For convenience we introduce the functions α_i, defining them in the obvious way through the equation

$$w(t, \beta) = \alpha_1(t) + \alpha_2(t)(t - \beta) + \alpha_3(t)t(t - \beta)^2/2.$$

Recall now that if

$$p(t) = \int_0^t \frac{(t-\beta)^n u(\beta)\, d\beta}{n!},$$

then p, u satisfy the $(n+1)$st order differential equation

$$p^{(n+1)}(t) = u(t), \qquad t \geqq 0,$$
$$p^{(n)}(0) = p'(0) = p(0) = 0.$$

Thus our operator can be synthesized as shown in Fig. 2.

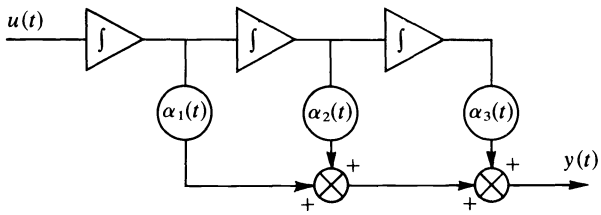


FIG. 2. *The differential realization*

8. Other extensions. In our development we have restricted attention to the single variate case which is typified by $H = L_2(\nu)$. It is convenient now to point out that extension to more general settings is easily accomplished. Consider first $L_2^m(\nu)$ and let $[\,\cdot\,,\cdot\,]$ denote the usual inner product on R^m. We then have

$$\langle P^t x, y\rangle = \int_0^t [x(\beta), y(\beta)]\, d\beta, \qquad x, y \in L_2^m(\nu).$$

The computation of $\|P^t x_i\|$ undergoes the usual adjustment.

Reviewing the development of §§ 2 and 3, we see that N, $\hat{y}$ and w of equations (3), (4) and (5) are well-defined in the new context. Concerning X, Π the new expressions become

$$X(t) = \operatorname{diag}[\,\cdots, [x_i(t), x_i(t)]/\|P^t x_i\|^2, \cdots], \tag{16}$$

and

$$\Pi_{ij}(t) = [x_i(t), x_j(t)]/\|P^t x_i\| \cdot \|P^t x_j\|, \tag{17}$$

with (6) and (9) still holding for N and M respectively. Theorem 3 is still valid with the proviso that the components of $\hat{y}$ and $\mathbf{x}$ are themselves m-tuples and that

$$\mathbf{x}(t)u(t) = ([x_1(t), u(t)], \cdots, [x_n(t), u(t)]), \qquad t \in \nu.$$

We note that the order of the differential system of Theorem 3 is dependent only on the cardinality of $\{(x_i, y_i)\}$ and not on whether x_i or y_i is scalar or vector-valued.

In the same spirit as the above extension, consider the case where the input-output pairs are derived from a distributive system. Let the spatial domain

be [0, 1] and ν denote the temporal domain. As a formalism we now define $[\cdot, \cdot]$ as follows:

$$[x(t), y(t)] = [x, y](t) = \int_0^1 x(t, s)y(t, s)\, ds.$$

Using this formalism, equations (16) and (17) are well-defined and Theorem 3 once more is valid. The point to this particular extension is that a finite element simulator results for a distributive system.

In another direction we note that extensive use is made of the "well-posed" condition. However, the dimension of the linear span of $\{P^t x_i\}$ is a monotone step function of "t", and, as such, nonwell-posed problems decompose into a finite collection of well-posed problems.

Other adjustments, which require more detailed explanation than space permits here, can be made which remove the well-posed assumption entirely.

9. Closure. It is interesting to note that the present study, together with [1], demonstrates a complete solution in operator form to the synthesis problem. Once the solution is in hand the concept of state is implicitly introduced as a realization mechanism. This then is a graphic example of the subsidiary nature of the state concept in system theory.

We note in closing that [1] also provides a polynomic solution to the nonlinear case. The polynomic operators in question also have a state variable realization. Our attention here has been centered on the linear case primarily because the associated state variable realization is less direct and hence a more richer and interesting topic.

REFERENCES

[1] W. A. PORTER, *Data interpolation, causality structure and system identification*, Information and Control, 29 (1975).

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

OPTIMAL CONTROL OF NONSYMMETRIC HYPERBOLIC SYSTEMS IN n VARIABLES ON THE HALF-SPACE*

RICHARD B. VINTER† AND TIMOTHY L. JOHNSON‡

Abstract. We study a quadratic control problem on the finite time interval with respect to the system of hyperbolic partial differential equations

$$\frac{\partial y}{\partial t} = \sum_i A_i \frac{\partial y}{\partial x_i} + f,$$

$$My_{\partial\Omega} = u,$$

$$y(0) = y_0,$$

on the spatial domain $\Omega = \{x \in \mathbb{R}^n | x_1 > 0\}$. For a special case it is shown that the control u may be synthesized in feedback form. The nonlinear operator equations involved in this synthesis are shown to have unique solutions within an appropriate class of functions.

1. Introduction. We consider quadratic cost boundary control of hyperbolic systems of partial differential equations. The development is built on recent results of Rauch [7] concerning well-posedness of mixed initial boundary value problems in L^2 for hyperbolic systems on the half-space which allow for varying coefficients, apply for an arbitrary number of space dimensions and do not require the systems to be symmetrizable.

This study is motivated by certain problems in distributed control where nonsymmetric hyperbolic systems of partial differential equations in more than one spatial variable arise [3]. Although quadratic cost control problems for hyperbolic equations are studied in [5] and [9], Lions supplies only a heuristic treatment for the boundary control problem and Russell limits attention to one spatial variable.

The optimal control for a fairly general class of quadratic control problems is first characterized through the solution of a two-point boundary value problem. Thus far, the development is a routine application of methods in [5]. Interest resides rather in the next step; that of realizing the control in feedback form in a special, but nontrivial, case when only the terminal cost is present. Corresponding to this special case we proceed to establish existence and uniqueness of the solution to an operator Riccati differential equation through which the feedback operators may be expressed. The general form of these results has previously been indicated in [3].

We wish to emphasize that the results reported here are only the first step towards a realistic study of engineering control problems. Problems of the type

* Received by the editors July 1, 1975. This research was carried out in part at the Decision and Control Sciences Group of the M.I.T. Electronic Systems Laboratory, and was supported by the National Science Foundation under Grant NSF-GK-41647.

† Department of Computing and Control, Imperial College of Science and Technology, London SW7 2BT, England. The research of this author was supported by the Commonwealth fund (Harkness Fellowships).

‡ Electronic Systems Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

considered herein may result from linearization of fundamental equations of continuum mechanics electrodynamics, and thermodynamics about equilibrium points or other fixed trajectories. In particular, we cannot generally expect the assumption of strict hyperbolicity imposed below to be met. Also, relevance of a study involving spatial domains which are half-spaces rests largely on the insights it may give into situations where the domains are more complicated yet the wave propagation velocity and the time interval of interest are small compared with the domain dimensions.

For one spatial variable (see [9]), existence and uniqueness in the large to solutions of the Riccati partial differential equation characterizing the optimal control may be *directly* studied through the classical construction of integrating along characteristic chains and Picard iteration using optimality considerations to establish an a priori inequality. This is possible because the Riccati partial differential equation turns out to be a semilinear system which is *diagonalizable* (even though it is in two spatial variables). The structure of the optimal control may therefore be examined by verifying that conditions on the solution of the Riccati partial differential equation under which the Riccati partial differential equation may be set up are indeed valid.

For more than one spacial variable the classical construction fails, since each step of the Picard iteration no longer in general reduces to solution of a system of ordinary differential equations on the characteristic surfaces. In the present study it becomes necessary, therefore, to undertake an *indirect* study of the operator Riccati equation through consideration of the two-point boundary value problem characterizing the optimal control. This is an approach associated with the name of Lions.

The problem is considerably more difficult when we pass to more than one spatial variable and the analysis is pushed through under conditions which are more stringent than those required in Russell's treatment. The reason is basically this: The boundary control is expressed through the boundary value of the "adjoint variable" $p(t)$. The method of characteristics is well-suited to handling problems for which the data on the adjoint equation are not compatible. For, even in this situation, the method establishes that p is piecewise continuous and this is sufficient to assure that the boundary value of $p(t)$ is well-defined outside a null set of the time interval. Abstract methods, such as are used here, are not so well suited since they cannot readily exploit the property that $p(t)$ is piecewise continuous and not merely a bounded measurable function. Therefore for the boundary value to be defined, we need to make several additional assumptions assuring compatibility of the data on the adjoint equation.

2. Preliminary notations and definitions. Let V be an open, connected subset of $\mathbb{R}^n$ with smooth boundary ∂V. $C_0^\infty(V; \mathbb{R}^k)$ is taken to be the space of infinitely differentiable maps $V \to \mathbb{R}^k$ with compact support. $C_{(0)}^\infty(V; \mathbb{R}^k)$ is the restriction of $C_0^\infty(\mathbb{R}^n; \mathbb{R}^k)$ to the closure of V. $L^2(V; \mathbb{R}^k)$ has its usual meaning of the space of Lebesgue square integrable functions (modulo null functions) with natural inner product. For economy of notation, $\langle \cdot, \cdot \rangle_{L^2(V;\mathbb{R}^k)}$ is often abbreviated to $\langle \cdot, \cdot \rangle_V$.

We shall have occasion to use certain Sobolev spaces. For integral $s \geqq 1$, $H^s(V; \mathbb{R}^k)$ is the completion of $C^\infty_{(0)}(V; \mathbb{R}^k)$ with respect to the norm

$$\|y\|_{H^s} = \left(\int_V \sum_{|\alpha| \leqq s} \|D^\alpha y(x)\|^2 \, dx \right)^{1/2} \tag{2.1}$$

(for $\alpha = \{\alpha_i\}$, $i = 1, \cdots, n$, nonnegative integers, D^α is the differential operator $\partial\alpha_1/\partial x_1^{\alpha^1} \cdots \partial^{\alpha_n}/\partial x_n^{\alpha_n}$ and $|\alpha| = \sum_{i=1}^h \alpha_i$). The closed subset H^s_0 of H^s is taken to be the completion of $C^\infty_0(V, \mathbb{R}^k)$ with respect to (2.1).

We shall for the most part be concerned with H^1, writing κ for the canonical injection $H^1 \to L^2$.

On those occasions when we wish to emphasize merely the domains of functions in L^2, H^s, etc., we write $L^2(V), \cdots$, for $L^2(V; \mathbb{R}^k) \cdots$.

We remind readers of the trace theorem (see [6] for a much more refined statement): for $n \geqq 2$, $s \geqq 1$ integers, take $\Omega \subset \mathbb{R}^n$, an open half-space. Then the restriction of $C^\infty_{(0)}(\Omega; \mathbb{R}^k)$ to $\partial\Omega$ defines a *bounded* linear map from a dense subset of $H^s(\Omega; \mathbb{R}^k)$ into $H^{s-1}(\partial\Omega; \mathbb{R}^k)$ which may in consequence be lifted to all of $H^s(\Omega; \mathbb{R}^k)$.

For $\mathcal{H}$ a real, separable Hilbert space, $[t_0, t_1]$ an interval in $\mathbb{R}$, $L^2([t_0, t_1]; \mathcal{H})$ and $C([t_0, t_1]; \mathcal{H})$ have their usual meanings of square integrable, strongly continuous, respectively, maps $[t_0, t_1] \to \mathcal{H}$. We also introduce $H^1([t_0, t_1]; \mathcal{H})$, the space of $\mathcal{H}$-valued distributions on $[t_0, t_1]$, such that h, Dh define L^2 functions.[1]

3. Mixed boundary initial value problems for hyperbolic systems. Here we present the results on mixed problems for hyperbolic systems which will be required below. These results have been built up in a series of papers [2], [4], [7]. Rauch has provided the final step in establishing that the present class of problems is well-posed in the L^2 sense for nonzero initial data and inhomogeneous boundary conditions.

We consider the mixed problem

$$\begin{aligned} &\frac{\partial y}{\partial t} = \mathcal{A}\left(\frac{\partial}{\partial x}\right) y + f, \qquad \mathcal{A} = \sum_{j=1}^m \frac{A_j \partial}{\partial x_j}(\cdot) + K, \\ &\text{boundary condition} \quad My_\Sigma = g, \\ &\text{initial condition} \quad y(0) = y_0, \qquad y, \text{ real } n\text{-vector}. \end{aligned} \tag{3.1}$$

The following sets are identified:

$$\begin{aligned} &T = [0, t_1], \\ &\Omega = \{x \in \mathbb{R}^m \mid (x_1 > 0\}, \qquad \partial\Omega = \{x \in \mathbb{R}^m \mid x_1 = 0\}, \qquad m > 1. \\ &Q =]0, t_1[\times \Omega, \qquad \Sigma =]0, t_1[\times \partial\Omega. \end{aligned}$$

[1] Equivalently, $H^1([t_0, t_1]; \mathcal{H})$ comprises $\mathcal{H}$-valued functions h on $[t_0, t_1]$ (modulo null functions) such that h is a.e. strongly differentiable and

$$h(t) = h(t_0) + \int_{t_0}^t Dh(\tau) \, d\tau \quad \text{all } t \in [t_0, t_1]$$

with h, Dh square integrable.

K, M, A_i, $i=1,\cdots,m$, are C^∞ matrix-valued functions with domain $\mathbb{R}\times\Omega$ or $\mathbb{R}\times\partial\Omega$ (as appropriate) each of which may be expressed as the sum of a constant function and a function of compact support.

We introduce:

Assumption 3.1 (Strict hyperbolicity). The determinant equation

$$\psi(s)=\det\left(-sI+\sum_{j=1}^{m}\lambda_j A_j(t,x)\right)=0$$

has n distinct real roots for all $\lambda\in\mathbb{R}^m$, $\lambda\neq 0$, all $(t,x)\in\mathbb{R}\times\Omega$.

Assumption 3.2 (Noncharacteristic boundary).

$$\det\left(A_1(t,x)\right)\neq 0$$

for all $(t,x)\in\mathbb{R}\times\partial\Omega$.

Assumption 3.3 (Determinate boundary values). For each $(t,x)\in\mathbb{R}\times\Omega$, A_1 has the normal form[2]

$$A_1=\begin{bmatrix}A^- & 0\\ 0 & A^+\end{bmatrix},\qquad A^-=\operatorname{diag}(a_1,\cdots,a_r),\qquad A^+=\operatorname{diag}(a_{r+1},\cdots,a_n),$$

where $a_i<0$, $i=1,\cdots,r$; $a_i>0$, $i=r+1,\cdots,n$, and for each $(t,x)\in\mathbb{R}\times\partial\Omega$, $M(t,x)$ is $r\times n$ and rank $M(t,x)=r$.

We need also to restrict the null space of the boundary operator M. For each $(\bar t,\bar x)\in\mathbb{R}\times\Omega$, define the $\mathbb{C}^{n\times n}$-valued function $\bar A(\cdot,\cdot)$ by

$$\bar A(s,k)=\bar A_1^{-1}\left(sI-i\sum_{j=2}^{m}k_j\bar A_j\right),$$

$$s\in\mathbb{C},\quad k=(k_2,\cdots,k_m)\in\mathbb{R}^{m-1},\quad \bar A_j=A_j(\bar t,\bar x)\quad\text{etc.}$$

Take $\bar{\mathcal{M}}(s,k)$ to be the generalized eigenmanifold

$$\bar{\mathcal{M}}(s,k)=\{x\in\mathbb{C}^m\mid(\bar A(s,k)-\sigma I)^n x=0 \text{ some integer } n, \text{ some } \sigma\in\mathbb{C}, \operatorname{Re}\{\sigma\}<0\}.$$

Write $\bar M$ for $M(\bar t,\bar x)$.

Assumption 3.4 (Condition on boundary space). For each $(\bar t,\bar x)\in\mathbb{R}\times\partial\Omega$ there exists some $\varepsilon>0$ such that, for all $r\times n$-matrices M' with[3] $\|\bar M-M'\|_{\mathrm{tr}}<\varepsilon$ and for all $k\in\mathbb{R}^{m-1}$, $s\in\mathbb{C}$,

$$\ker\{M'\}\cap\bar{\mathcal{M}}(s,k)=\{\theta\}\quad\text{(null element)}.$$

Partitioning the $r\times n$ matrix $M(\bar t,\bar x)$ as $[M^-(\bar t,\bar x)\ M^+(\bar t,\bar x)]$, where M^- is $r\times r$, a necessary condition that Assumption 3.4 hold is that $M^-(\bar t,\bar x)$ be nonsingular for each $(\bar t,\bar x)\in\mathbb{R}\times\partial\Omega$ [2]. We may assume therefore, without loss of generality, that M^- is the identity

$$M(\bar t,\bar x)=[I\,\vdots\,M^+(\bar t,\bar x)].$$

We shall be concerned with strong solutions to the mixed problem, defined as follows.

---

[2] Without loss of generality in view of the strict hyperbolicity assumption.

[3] $\|M\|_{\mathrm{tr}}^2=\operatorname{trace} MM^T$.

DEFINITION 3.1. For given $f, g, y_0 \in L^2$, we define $y \in L^2(Q; \mathbb{R}^n)$ to be a *strong* (L^2) *solution* to the mixed problem (3.1) iff there exists a sequence $\{y^n\}$, $y^n \in C^\infty_{(0)}(\bar{Q}; \mathbb{R}^n)$, and some $y_\Sigma \in L^2(\Sigma; \mathbb{R}^n)$ (termed the *strong* (L^2) *boundary value of* y on Σ) such that

$$\|y^n - y\|_Q, \quad \|y^n_\Sigma - y_\Sigma\|_\Sigma, \quad \left\|\left(\frac{\partial}{\partial t} - \mathcal{A}\right) y^n - f\right\|_Q,$$

$$\|My^n_\Sigma - g\|_\Sigma, \quad \|y^n(0) - y_0\| \to 0 \quad \text{as } n \to \infty.$$

For $t \in T$, $y(t) \in L^2(\Omega; \mathbb{R}^n)$ is called a *strong* (L^2) *solution at time* t if additionally,

$$\|y^n(t) - y(t)\|_\Omega \to 0. \tag{3.2}$$

We have the following fundamental result (see [7]).

THEOREM 3.1 (Well-posedness of the mixed problem). *Under Assumptions* 3.1–3.4, *for given* $f, g, y_0 \in L^2$, *the mixed problem* (3.1) *has a unique strong solution with unique strong boundary value* $y_\Sigma \in L^2$ *and the strong solution at time* t, $y(t)$, *for each* $t \in T$ *is uniquely defined by* (3.2). *Further we have the estimate*

$$\|y(t)\|_\Omega + \|y\|_Q + \|y_\Sigma\|_\Sigma \leqq \text{const.} \{\|f\|_Q + \|g\|_\Sigma + \|y_0\|_\Omega\} \tag{3.3}$$

uniformly in $f, g, y_0 \in L^2$, $t \in T$.

Assumptions 3.2–3.4 are in effect the weakest possible if the above mixed problem for strictly hyperbolic systems is to be well-posed in L^2 (see [4]).

We make two important observations.

Remark 1. The mixed problem (3.1) has a smooth solution for smooth data. More precisely, if

(a) $$y_0 \in C^\infty_0(\Omega), \quad f \in C^\infty_{(0)}(Q), \quad g \in C^\infty_0(\Sigma)$$

and

(b) $$f(t, x), \quad g(t, \sigma) \quad \text{vanish for } t \leqq 0,$$

then $y \in C^\infty_{(0)}(Q)$ and y_Σ, $y(t)$ are the restrictions of y to Σ, $\{t\} \times \Omega$ respectively (see [7]).

Remark 2. It is easily deduced that $t \mapsto y(t) : T \to L^2(\Omega)$ is strongly continuous for $y(t)$ as in Theorem 3.1: to see this we use estimate (3.3) (uniformly in t) and Remark 1 to construct an equicontinuous family of smooth functions $T \to L^2(\Omega)$ converging pointwise to $y(t)$ (see [11]). Notice also that the strong solution y defines an element in $L^2(T; L^2(\Omega; \mathbb{R}^n))$ which coincides with $t \mapsto y(t)$ within a null set.

Rauch has also supplied the following regularity result (see [7]).

THEOREM 3.2. *Suppose in addition to the hypotheses of Theorem* 3.1 *that, for some positive integer* s, $y_0 \in H^s_0(\Omega)$, $f \in H^s(Q)$, $g \in H^s(\Sigma)$ *with* $D^j_t f|_{t=0} = 0$, $D^j_t g|_{t=0}$, $0 \leqq j \leqq s-1$. *Then* y, $y(t)$, y_Σ *are* H^s *functions and* (3.3) *holds with respect to* H^s *norms uniformly in* f, g, y_0, t.

Under the added hypotheses of Theorem 3.2, we also have that $t \mapsto y(t) : T \to H^1(\Omega)$ is strongly continuous.

Study of the control problem will require introduction of the adjoint problem:

$$
\begin{aligned}
&\frac{-\partial p}{\partial t} = \mathscr{A}^*\left(\frac{\partial}{\partial x}\right)p + h,\\
&\mathscr{A}^* = -\sum_{j=1}^{m} A_j^T \frac{\partial}{\partial x_j} + \left(K^T - \sum_{j=1}^{m} \frac{\partial A_j^T}{\partial x_j}\right),\\
&M^* p_\Sigma = e,\\
&p(t_1) = p_1.
\end{aligned}
\tag{3.4}
$$

Here the matrix-valued function M^* with domain Σ is defined by

$$M^* = [-(A^+)^{-1}(M^+)^T(A^-) \vdots I]$$

and h, p_1, e are arbitrary elements in $L^2(Q; \mathbb{R}^n)$, $L^2(\Omega; \mathbb{R}^n)$ and $L^2(\Sigma; \mathbb{R}^{n-r})$ respectively. Notice that the adjoint system "runs backwards in time" with data given at time t_1. We recognize $\mathscr{A}^*$ as the formal adjoint of $\mathscr{A}$, i.e., for any $C^\infty_{(0)}$ functions a, $\tilde{a}$ with $Ma = 0$, $M^*\tilde{a} = 0$ on $\partial\Omega$ we have $\langle \mathscr{A}a, \tilde{a}\rangle = \langle a, \mathscr{A}^*\tilde{a}\rangle$.

Strong (L^2) solutions, strong solutions at time t, and strong boundary values for the adjoint problem (3.4) are defined analogously to strong solutions, etc., for the mixed problem.

Now it may be shown [7] that the adjoint problem satisfies Assumptions 3.1–3.4 *backwards in time*. It follows that analogues of Theorems (3.1), (3.2) regarding existence and regularity of strong solutions to the adjoint problem apply.

The appropriate version of the divergence theorem will be an essential tool in characterization of the solution to the control problem.

PROPOSITION 3.1 (divergence theorem). *Suppose that Assumptions* 3.1–3.4 *hold.* Let y, p *be strong solutions to the mixed and adjoint problems respectively for arbitrary* L^2 *data. Then*[4]

$$\langle f, p\rangle_Q = \langle y, h\rangle_Q + \langle y_\Sigma, A_1 p_\Sigma\rangle_\Sigma - \langle y_0, p(0)\rangle_\Omega + \langle y(t_1), p_1\rangle_\Omega. \tag{3.5}$$

For smooth solutions this follows immediately by parts integration. We obtain the result, in general, by consideration of sequences of smooth functions converging to y, p in L^2, exploiting the property that y, p are *strong* solutions to the mixed, adjoint, problems respectively (see [11] for details).

4. The control problem. We introduce the *state equation*

$$
\begin{aligned}
&\frac{\partial y}{\partial t} = \mathscr{A}y + f,\\
&My_\Sigma = u,\\
&y(0) = y_0.
\end{aligned}
\tag{4.1}
$$

[4] Given a strong solution y to the mixed problem, here and in the sequel, y_Σ, $y(t)$ always denote the strong boundary value and the strong solution at time t. Similar meaning attaches to p_Σ, $p(t)$ in relation to the adjoint problem.

Here $\mathscr{A}$ and M are as in § 3. f, u, y_0 are L^2 functions. We view y_0 and f as fixed. We are free to choose u which is termed the *control*. We know from the previous section that for each choice of u, (4.1) has a unique strong solution Q, a strong boundary value and a strong solution at time t, $t \in T$. We write these y^u, y^u_Σ, $y^u(t)$ to emphasize the dependence on u.

The *cost function* $u \mapsto J(u)$ is taken as

$$J(u) = \int_T \langle (y^u - z)(t), Q(t)(y^u - z)(t) \rangle_\Omega \, dt$$

$$+ \langle (y^u - z)(t_1), R(y^u - z)(t_1) \rangle_\Omega + \int_T \langle u(t), u(t) \rangle_{\partial\Omega} \, dt.$$

Here, $z \in L^2(Q; \mathbb{R}^n)$, $z(t_1) \in L^2(\Omega; \mathbb{R}^n)$ are given. $R \in \mathscr{L}(L^2(\Omega; \mathbb{R}^n))$, $t \mapsto Q(t) : T \to \mathscr{L}(L^2(\Omega; \mathbb{R}^n))$ is measurable (with respect to the strong operator topology) and essentially bounded. We assume that R, $Q(t)$ (for each $t \in T$) are self-adjoint and nonnegative.

Control problem. Minimize $J(u)$ over $u \in L^2(\Sigma; \mathbb{R}^r)$.

It is a routine matter to modify the development below to accommodate a "distributed control" term in the state equation and in the cost function, also to introduce a fixed inhomogeneous term in the boundary condition of the state equation and to penalize y^u_Σ (see [11]).

5. Characterization of the optimal control through a two-point boundary value problem. In view of the above assumptions,

$$J(u) = \pi(u, u) - 2L(u) + \text{const.},$$

where, as may be shown, $\pi(\cdot, \cdot) : L^2 \times L^2 \to \mathbb{R}$ is a continuous, coercive, symmetric, bilinear form and $u \mapsto L(u)$ is a bounded linear functional on L^2.

Standard results concerning minimization of quadratic forms (see, e.g., [5]) give us existence and uniqueness of the optimal control u_0 and its characterization:

u_0 is optimal $\Leftrightarrow$

$$\int_T \langle (y^{u_0} - z)(t), Q(t)(y^u - y^{u_0})(t) \rangle_\Omega \, dt + \langle (y^{u_0} - z)(t_1), R(y^u - y^{u_0})(t_1) \rangle_\Omega \, dt \tag{5.1}$$

$$+ \langle u_0, u - u_0 \rangle_\Sigma = 0 \quad \text{for all } u \in L^2(\Sigma; \mathbb{R}^r).$$

We now pattern arguments in [5] to refine this characterization through introduction of the adjoint equation:

PROPOSITION 5.1. *For given $u_0 \in L^2$, let p be the strong solution to the adjoint problem*

$$\frac{-\partial p}{\partial t} = \mathscr{A}^* p + \{t \mapsto Q(t)(y^{u_0} - z)(t)\},$$

$$M^* p_\Sigma = 0, \tag{5.2}$$

$$p(t_1) = R(y^{u_0} - z)(t_1).$$

Then[5]

$$u_0 \text{ is optimal} \Leftrightarrow u_0 = A^- p_\Sigma^-.$$

Details of the proof are given in [11]. The essential step is application of the divergence theorem (Proposition 3.1) to $y^u - y^{u_0}$ and p (for arbitrary $u \in L^2$).

6. Feedback synthesis of the control for a special case. We should like additionally to achieve a feedback synthesis of the optimal control, that is, show that u_0 may be determined pointwise in time through a function dependence on $y^{u_0}(t)$ independent of the initial condition y_0.

This we do for the following subclass of problems:

Terminal cost control problem. Minimize $\bar{J}(u)$ over $u \in L^2(\Sigma; \mathbb{R}^r)$. Here, $\bar{J}(u) = \langle (y^u - z)(t_1), R(y^u - z)(t_1)\rangle_\Omega + \langle u, u\rangle_\Sigma$ with $z(t_1) \in L^2(\Omega; \mathbb{R}^n)$, $R \in \mathscr{L}(L^2(\Omega; \mathbb{R}^n))$ satisfying both

(a) $R = R^*, R \geqq 0$,

(b) R carries $L^2(\Omega; \mathbb{R}^n)$ functions into $H_0^1(\Omega; \mathbb{R}^n)$ functions and $\kappa^{-1} R \in \mathscr{L}(L^2; H_0^1)$ (recall κ, the canonical injection $H_0^1 \to L^2$).

Example. Take $S \in H_0^1(\Omega \times \Omega; \mathbb{R}^{n \times n})$. Suppose that $S(x, x') = S^T(x', x)$ a.e. $(x, x') \in \Omega \times \Omega$ and $\iint_{\Omega \times \Omega} y^T(x) S(x, x') y(x')\, dx\, dx' \geqq 0$ all $y \in L^2$. Then the map $y(x') \overset{R}{\mapsto} \int_\Omega K(x, x') y(x')\, dx'$ with domain $L^2(\Omega)$ takes values in H_0^1 and satisfies the conditions (a), (b) above (see [11]).

Henceforth we limit attention to the terminal cost control problem. Here treatment is greatly simplified by the property that the adjoint equation admits a strong H^1 solution. This follows immediately from Theorem 3.2 and the assumed properties of R (see Proposition 5.1)

PROPOSITION 6.1. *For given $u \in L^2(\Sigma; \mathbb{R}^r)$, let p be the strong (H^1) solution to*

$$\begin{aligned} \frac{-\partial p}{\partial t} &= \mathscr{A}^* p, \\ M^* p_\Sigma &= 0, \\ p(t_1) &= R(y^u - z)(t_1). \end{aligned} \tag{6.1}$$

Then u is optimal $\Leftrightarrow u = \{t \mapsto A^- p_{\partial\Omega}^-(t)\}$.

In Proposition 6.1, we have only to justify replacing p_Σ by $\{t \mapsto p_{\partial\Omega}(t)\}$, where $p_{\partial\Omega}(t)$ is the trace on $\partial\Omega$ of the strong H^1 solution at time t of (6.1). But p_Σ, $\{t \mapsto p_{\partial\Omega}(t)\}$ define the same $L^2(T; L^2(\partial\Omega; \mathbb{R}^n))$ functions, in view of the definition of strong solutions and the trace theorem, being in effect the mean square and pointwise limit of the same sequence of smooth functions (see [11] for details).

We now introduce the natural spaces $\mathscr{F}$, $\mathscr{F}^*$ in which to seek solutions to the two-point boundary value problem associated with the terminal cost control problem.

[5] Here, and below, we partition the n vector p as $[p_1, \cdots, p_r \,|\, p_{r+1}, \cdots, p_n] = [p^- \; p^+]$.

Bearing in mind that the mixed problem is well-posed on $[\tau; t_1]$ regardless of the time $\tau \in [0, t_1[$ at which initial data is supplied, we may define $\mathscr{F}_\tau \subset L^2(]\tau, t_1[\times\Omega; \mathbb{R}^n)$: $y \in \mathscr{F}_\tau \Leftrightarrow y$ is the strong solution of

$$\frac{\partial y}{\partial t} = \mathscr{A}y + f,$$

$$My_\Sigma = g,$$

$$y(\tau) = y_0,$$

for some f, g, $y_0 \in L^2$. $\mathscr{F}_\tau^* \subset L^2(]\tau, t_1[\times\Omega; \mathbb{R}^n)$ is defined analogously in relation to the adjoint problem.

LEMMA 6.1. *For fixed $\tau \in [0, t_1[$ and fixed $a \in L^2(\Omega; \mathbb{R}^n)$, the optimality system*

$$\frac{\partial y}{\partial t} = \mathscr{A}y + f,$$

$$My_\Sigma = (A^-)p_\Sigma^-, \tag{6.2}$$

$$y(\tau) = a,$$

and

$$\frac{-\partial p}{\partial t} = \mathscr{A}^* p,$$

$$M^* p_\Sigma = 0, \tag{6.3}$$

$$p(t_1) = R(y - z)(t_1)$$

has a unique strong solution in $\mathscr{F}_\tau \times \mathscr{F}_\tau^*$.

Existence of a solution to the optimality system is immediate from Proposition 5.1; uniqueness may be deduced from the uniqueness of the characterization (5.1) (see [11] for details).

It follows from Lemma 6.1 that we may define a family of maps

$$\mathscr{P}(\tau): L^2(\Omega; \mathbb{R}^n) \to L^2(\Omega; \mathbb{R}^n), \qquad \tau \in [0, t_1[,$$

$$a \overset{\mathscr{P}(\tau)}{\mapsto} p(\tau),$$

where for each $\tau \in [0, t_1[$, (y, p) is the strong solution (in $\mathscr{F} \times \mathscr{F}^*$) to the optimality system (6.2), (6.3). We define $\mathscr{P}(t_1)$ by

$$a \overset{\mathscr{P}(t_1)}{\mapsto} R(a - z(t_1)), \qquad a \in L^2(\Omega).$$

Evidently $\mathscr{P}(\tau)$ is affine, and

$$\mathscr{P}(\tau)a = P(\tau)a + r(\tau).$$

$P(\tau)$ is computed as $a \overset{P(\tau)}{\mapsto} p(\tau)$, where now we delete the terms f, $z(t_1)$ from the optimality system; $r(\tau)$ is simply $\mathscr{P}(\tau)\theta$ (θ, null-element).

Again let u_0 be the optimal control for the terminal cost control problem. We have from Proposition 6.1,

$$u_0(t) = (A^-)(\mathcal{P}(\tau)y^{u_0}(\tau))^-_{\partial\Omega}, \qquad t \in T. \tag{6.4}$$

This achieves the feedback synthesis of the control. The remainder of the paper is given over to developing properties of the map $\mathcal{P}(\cdot)$.

7. Properties of the feedback operators. Here the main results are presented. We consider the terminal cost control problem throughout. The functions $P(t)$, $r(t)$ are as in § 6.

THEOREM 7.1 (Properties of $P(t)$). *Suppose that A_i, $i = 1, \cdots, n$, M and K are independent of time, further that by adjustment on a null set f defines an element in $C(T; L^2(\Omega; \mathbb{R}^n))$. Then $P(\cdot)$ is the unique map $T \to \mathcal{L}(L^2(\Omega; \mathbb{R}^n))$ satisfying*

(a) *range $\{P(t)\} \subset \kappa H^1(\Omega; \mathbb{R}^n)$ with $\kappa^{-1}P(t) \subset \mathcal{L}(L^2; H^1)$ each $t \in T$ and*

$$\sup_{t \in T} \|\kappa^{-1}P(t)\|_{\mathcal{L}(L^2, H^1)} < \infty,$$

(b) *$P(t_1) = R$, $P(t) = P^*(t)$, $M^*(P(t)a)^-_{\partial\Omega} = 0$ all $t \in T$, $a \in L^2(\Omega; \mathbb{R}^n)$,*

(c) *$t \mapsto \kappa^{-1}P(t) : T \to \mathcal{L}(L^2; H^1)$ is strongly continuous from the right, and*

(d) *for each a, $\tilde{a} \in L^2(\Omega; \mathbb{R}^n)$, $t \mapsto \langle a, P(t)\tilde{a}\rangle_\Omega : T \to \mathbb{R}$ is absolutely continuous with*

$$\begin{aligned} \frac{d}{dt}\langle P(t)a, \tilde{a}\rangle_\Omega = &-\langle \mathcal{A}^*P(t)a, \tilde{a}\rangle_\Omega - \langle \mathcal{A}^*P(t)\tilde{a}, a\rangle_\Omega \\ &+ \langle A^-(P(t)a)^-_{\partial\Omega}, A^-(P(t)\tilde{a})^-_{\partial\Omega}\rangle_{\partial\Omega} \quad \text{a.e. } t \in T. \end{aligned} \tag{7.1}$$

THEOREM 7.2 (Properties of $r(t)$). *Suppose again that A_i, $i = 1, \cdots, n$, M and K are independent of time and that f defines an element in $C(T; L^2(\Omega; \mathbb{R}^n))$. Then $r(\cdot)$ is the unique map $T \to L^2(\Omega; \mathbb{R}^n)$ such that*

(a) *$r(t)$ takes values in $\kappa H^1(\Omega; \mathbb{R}^n)$ and*

$$\sup_{t \in T} \|\kappa^{-1}r(t)\|_{H^1} < \infty,$$

(b) *$r(t_1) = -Rz(t_1)$, $M^*r_{\partial\Omega}(t) = 0$ each $t \in T$,*

(c) *$t \mapsto \kappa^{-1}r(t) : T \to H^1$ is strongly continuous from the right, and*

(d) *for each $a \in L^2(\Omega; \mathbb{R}^n)$, $t \mapsto \langle a, r(t)\rangle_\Omega : T \to \mathbb{R}$ is absolutely continuous and satisfies*

$$\begin{aligned} \frac{d}{dt}\langle r(t), a\rangle_\Omega = &-\langle \mathcal{A}^*r(t), a\rangle_\Omega + \langle A^-r^-_{\partial\Omega}(t), A^-(P(t)a)^-_{\partial\Omega}\rangle_{\partial\Omega} \\ &- \langle f(t), P(t)a\rangle_\Omega \quad \text{a.e. } t \in T, \end{aligned} \tag{7.2}$$

with $P(t)$ as in Theorem 7.1.

The proofs of Theorems 7.1, 7.2 are sketched in the next section. In outline, we interpret $\langle \mathcal{P}(0)y_0, y_0\rangle$ as $\min \bar{J}(u)$. This yields an identity which may be "differentiated" to give (7.1) and (7.2). Thus, in general approach, we use the methods of [5]. The novelty lies in the manner in which we use the regularity of the

solution to the adjoint equation (assured by Theorem 3.2) to justify the differentiation. We shall see also that proving uniqueness within the specified class presents special difficulties.

The assumption of time invariance in Theorems 7.1, 7.2 is a technical condition introduced to ensure that the constant in estimate (3.3) can be chosen independently of the time $t_0 \in [0, t_1[$ at which initial data is supplied and can almost certainly be dropped.

We remark that (7.1) may be interpreted as a partial differential equation in distributions on Ω (cf. [5, p. 157]): let $\mathscr{D}$ be the space of test functions on $\Omega (C_0^\infty(\Omega)$ functions equipped with the inductive limit topology) and let $\mathscr{D}'$ be the space of $\mathbb{R}^n$-valued distributions on Ω (space of continuous linear maps $\mathscr{D} \to \mathbb{R}^n$, equipped with its strong topology).[6] For each $t \in T$, $P(t) \in \mathscr{L}(L^2)$ in particular defines a continuous map $\mathscr{D}^n \to \mathscr{D}'$; by the kernel theorem [10] therefore, $P(t)$ has the representation

$$\langle \tilde{a}, P(t)a \rangle_\Omega = \int\int_{\Omega\times\Omega} \tilde{a}^T(x)P(x, x', t)a(x')\, dx\, dx',$$

$$a = \{a_i\}, \quad \tilde{a} = \{\tilde{a}_i\} \quad \text{all } a_i, \tilde{a}_i \in \mathscr{D},$$

where $P(\cdot, \cdot, t) \in \mathscr{D}'(\Omega\times\Omega; \mathbb{R}^{n\times n})$ is uniquely determined by $P(t)$.

We have then from Theorem 7.1,

$$\frac{d}{dt}P(x, x', t) = -\mathscr{A}_x^* P(x, x', t) - P(x, x', t)\mathscr{A}_{x'} + \int_{\partial\Omega} P(x, \sigma, t)^-(A_\sigma^-)(A_\sigma^-)P(\sigma, x', t)\, d\sigma \quad a.e.\ t \in T,$$

$$M^*(x)P(x, x', t) = 0 \quad \text{for } x \in \partial\Omega, \quad x' \in \Omega, \tag{7.3}$$

$$P(x, x', t_1) = R(x, x') \quad (R(x, x'), \text{ kernel of } R),$$

$$P(x, x', t) = P^T(x', x, t) \quad \text{each } t \in T.$$

$$\int_\Omega P(x, x', t)a(x')\, dx' \text{ defines an } H^1 \text{ element for each } a \in \mathscr{D}^n.$$

Of course, (7.3) is meaningful only in a distribution sense. In particular, $(d/dt)P(x, x', t) \in \mathscr{D}'(\Omega\times\Omega; \mathbb{R}^n)$ such that for all $a, \tilde{a} \in \mathscr{D}^n$,

$$\int\int_{\Omega\times\Omega} \tilde{a}(x)\frac{d}{dt}P(x, x', t)a(x')\, dx\, dx' = \frac{d}{dt}\int\int_{\Omega\times\Omega} \tilde{a}^T(x)P(x, x', t)a(x')\, dx\, dx'.$$

$\mathscr{A}_x^* P(x, x', t)$ is defined following the usual definition of differentiation on $\mathscr{D}'$; likewise $P(x, x', t)\mathscr{A}_{x'}$ is merely $-\Sigma_i(\partial/\partial x_i')P(x, x', t)A_i(x')$. $\int_{\partial\Omega} P(x, \sigma, t)^-(A_\sigma^-)^2 P(\sigma, x', t)^-\, d\sigma$ is the unique kernel corresponding to the continuous bilinear form $\mathscr{D}^n \times \mathscr{D}^n \to \mathbb{R}$,

$$a, \tilde{a} \mapsto \int_{\partial\Omega} \left\{ \int_\Omega \tilde{a}^T(x)P(x, \sigma, t)^-_{\partial\Omega}(A_\sigma^-)^2 \left(\int_\Omega P(\sigma, x', t)a(x')\, dx' \right)^- \right\} d\sigma,$$

[6] See, e.g., [1].

the traces being well-defined in view of the regularity assumptions on $P(t, x, x')$. The regularity assumptions also assure that the boundary condition $M^*(x)P(x, x', t) = 0$ is meaningful.

The foregoing establishes existence of solutions to the partial differential equation (7.3) within the class of functions whose values define kernels of continuous linear maps $\mathscr{D}^n \to \mathscr{D}'$; the solution is unique in the sense of Theorem 7.1.

We may lend a similar interpretation to (7.2).

8. Proof of Theorems 7.1, 7.2. We compress routine steps in the material of this section. For a much more expansive treatment, the reader is referred to [11].

We first take note of an identity which interprets $\langle \mathscr{P}(0)y_0, y_0\rangle_\Omega$ as $\min \bar{J}(u)$:

LEMMA 8.1. *Fix $t \in [0, t_1[$. For a, $\tilde{a} \in L^2(\Omega; \mathbb{R}^n)$ let $(y, p)(\tilde{y}, \tilde{p})$ be the unique solution in $\mathscr{F}_t \times \mathscr{F}_t^*$ to the optimality system* (6.1), (6.2) *with $y(t) = a$ $(\tilde{y}(t) = \tilde{a})$. Then*

$$\langle a, \mathscr{P}(t)\tilde{a}\rangle_\Omega = \langle y(t_1), R\tilde{y}(t_1)\rangle_\Omega + \int_t^{t_1} \langle A^- p^-_{\partial\Omega}(\tau), A^-\tilde{p}^-_{\partial\Omega}(\tau)\rangle_{\partial\Omega}\, d\tau. \tag{8.1}$$

Proof. The identity (8.1) follows by application of the divergence theorem (Proposition 3.1) to y, $\tilde{p}$ and the regularity of p, $\tilde{p}$ which permits us to replace p_Σ, $\tilde{p}_\Sigma$ by $\{t \mapsto p_{\partial\Omega}(t)\}$, $\{t \mapsto \tilde{p}_{\partial\Omega}(t)\}$ respectively (cf. remarks following Proposition 6.1).

LEMMA 8.2 (A basic estimate). *For $t \in [0, t_1[$, take (y, p) to be the unique solution in $\mathscr{F}_t \times \mathscr{F}_t^*$* to (6.2), (6.3). *Then*

$$\begin{aligned} &\|y(\tau)\|_{L^2(\Omega)} + \|p(\tau)\|_{H^1(\Omega)} + \|p_{\partial\Omega}(\tau)\|_{L^2(\partial\Omega)} \\ &\qquad \leqq \text{const.}\, \{\|y(t)\|_{L^2(\Omega)} + \|f\|_{L^2(\Omega\times[t,t_1])} + \|z(t_1)\|_{L^2(\Omega)}\} \end{aligned} \tag{8.2}$$

uniformly in t, $\tau \in [t, t_1]$, $y(t)$, f, z.

Proof. That $y(\tau)$ is estimated as stated follows from the coercivity of $\bar{J}(u)$ and Theorem 3.1; the estimate for $p(\tau)$ is then an immediate consequence of the regularity theorem (Theorem 3.2). Finally the trace theorem justifies inclusion of $\|p_{\partial\Omega}(\tau)\|_{L^2(\partial\Omega)}$ in the left-hand side of (8.2).

Theorem 3.2 tells us nothing about the regularity of $t \mapsto y(t)$. We do have though that $y(t)$ is weakly differentiable with respect to a certain class of bounded linear functionals:

LEMMA 8.3 (Weak differentiability of $y(t)$). *Take $t \in [0, t_1[$. Let y, p be the unique solutions in $\mathscr{F}_t \times \mathscr{F}_t^*$ to* (6.2), (6.3). *Suppose that $\tilde{p} \in H^1(\Omega; \mathbb{R}^n)$ and $M^*\tilde{p}_{\partial\Omega} = 0$. Then for $0 < \delta \leqq t_1 - t$,*

$$\begin{aligned} \langle y(t+\delta) - y(t), \tilde{p}\rangle_\Omega = &\int_t^{t+\delta} \langle y(\tau), \mathscr{A}^*\tilde{p}\rangle_\Omega - \int_t^{t+\delta} \langle A^- p^-_{\partial\Omega}(\tau), A^-\tilde{p}^-_{\partial\Omega}\rangle_{\partial\Omega}\, d\tau \\ &+ \int_t^{t+\delta} \langle f(\tau), \tilde{p}\rangle_\Omega\, d\tau \end{aligned} \tag{8.3}$$

and

$$|\langle y(t+\delta) - y(t), \tilde{p}\rangle_\Omega| \leqq \text{const.}\, \{\|y(t)\|_{L^2} + \|f\|_C + \|z(t_1)\|_{L^2}\} \cdot \|\tilde{p}\|_{H^1} \cdot \delta. \tag{8.4}$$

Proof. If y, p were smooth functions, (8.3) would be given by parts integration. We demonstrate (8.3) in general by considering sequences of smooth functions approximating y, p in L^2, H^1 respectively. The estimate (8.4) now follows from the previous lemma, the property that $\mathscr{A}^*$ is a first order operator (whence $\|\mathscr{A}^*\tilde{p}\|_{L^2} \leqq \text{const.}\ \|\tilde{p}\|_{H^1}$) and the trace theorem.

We may now deduce the following preliminary properties of $P(t)$, $r(t)$.

LEMMA 8.4.

(a) *For each $t \in T$, $r(t) \in H^1(\Omega; \mathbb{R}^n)$ and $P(t)$ takes values in $H^1(\Omega; \mathbb{R}^n)$ with*

$$\sup_{t\in T} \|P(t)\|_{\mathscr{L}(L^2;H^1)} < \infty, \qquad \sup_{t\in T} \|r(t)\|_{L^2} < \infty.$$

(b) $M^*(P(t)a)^-_{\partial\Omega} = 0$, $M^*(r(t))^-_{\partial\Omega} = 0$, $t \in T$, $a \in L^2(\Omega)$.

(c) *The maps $t \mapsto P(t) : T \to \mathscr{L}(L^2; H^1)$ and $t \mapsto r(t) : T \to H^1(\Omega; \mathbb{R}^n)$ are continuous from the right with respect to the strong operator topology and the strong topology respectively.*

(d) *$P(t) \in \mathscr{L}(L^2)$ is nonnegative and self-adjoint.*

Proof. (a), (b) are consequences of Lemma 8.2 and the definition of $P(t)$, $r(t)$. To prove (c), we make use of the properties that $t \mapsto y(t) : T \to L^2$; $t \mapsto p(t) : T \to H^1$ are strongly continuous (recall remarks following Theorems 3.1, 3.2) and the boundedness of $\|\mathscr{P}(t)\|_{\mathscr{L}(L^2;H^1)}$ (see [11] for details). (d) follows from Lemma 8.1.

PROPOSITION 8.1. *For each a, $\tilde{a} \in L^2(\Omega)$, $t \mapsto \langle P(t)a, \tilde{a}\rangle_\Omega : T \to \mathbb{R}$ and $t \mapsto \langle r(t), a\rangle_\Omega : T \to \mathbb{R}$ are absolutely continuous with*

$$\text{(a)} \qquad \frac{d}{dt}\langle P(t)a, \tilde{a}\rangle_\Omega = -\langle \mathscr{A}^*P(t)a, \tilde{a}\rangle_\Omega - \langle \mathscr{A}^*P(t)\tilde{a}, a\rangle_\Omega$$
$$+ \langle (P(t)a)^-_{\partial\Omega}, (A^-)^2(P(t)\tilde{a})^-_{\partial\Omega}\rangle_{\partial\Omega} \quad \text{a.e. } t \in T,$$

and

$$\text{(b)} \qquad \frac{d}{dt}\langle r(t), a\rangle_\Omega = -\langle \mathscr{A}^*r(t), a\rangle_\Omega + \langle A^-r^-_{\partial\Omega}(t), A^-(P(t)a)^-_{\partial\Omega}\rangle_{\partial\Omega}$$
$$- \langle f(t), P(t)a\rangle_\Omega \quad \text{a.e. } t \in T.$$

Proof. Consider (a). We first show that $t \mapsto P(t) : T \to \mathscr{L}(L^2)$ is Lipshitz continuous with respect to $\|\cdot\|_{\mathscr{L}(L^2)}$. This is an exercise in breaking up (with the help of identity (8.1)) $\langle a, (P(t+\delta) - P(t))\tilde{a}\rangle_\Omega$ into a sum of terms to which the estimate (8.2) is applicable. Thus $t \mapsto \langle a, P(t)\tilde{a}\rangle_\Omega$ is, in particular, absolutely continuous for each a, $\tilde{a} \in L^2(\Omega)$. We conclude by using the identity (8.3) to prove that $t \mapsto \langle a, P(t)\tilde{a}\rangle_\Omega$ is differentiable from the right at every $t \in [0, t_1[$ to the value stated. It follows that $\langle a, P(t)\tilde{a}\rangle_\Omega$ is a.e. differentiable to the value stated for every a, $\tilde{a} \in L^2$.

(b) is similarly shown.

Referring back to Theorems (7.1), (7.2), we see that it remains to establish that P, r are unique within the specified class.

Conclusion of proofs of Theorems 7.1, 7.2 (Uniqueness of P, r). Let P, r be functions satisfying conditions (a)–(d) of Theorems 7.1, 7.2. Write $\mathscr{P}(t)$ for the affine map $a \overset{\mathscr{P}(t)}{\mapsto} P(t)a + r(t)$, each $t \in T$.

Step 1. We show by Picard iteration that for $a \in L^2(\Omega)$, there is a unique map $t \mapsto y(t): T \to L^2(\Omega)$ such that $(t, x) \to (y(t))(x)$ defines a strong solution to the mixed problem

$$\frac{\partial y}{\partial t} = \mathscr{A} y,$$

$$My_\Sigma = (A^-)\{t \mapsto (\mathscr{P}(t)y(t))^-_{\partial\Omega}\},$$

$$y(0) = a.$$

Step 2. Take $y(t)$ as in Step 1. For each $t \in T$, define $p(t) \in L^2(\Omega)$ by $p(t) = \mathscr{P}(t)y(t)$. We verify that

(i) $t \mapsto \kappa^{-1}p(t)$ defines an element in $L^\infty(T; H^1(\Omega; \mathbb{R}^n))$,

(ii) $t \mapsto p(t)$ defines an element in $H^1(T; L^2(\Omega; \mathbb{R}^n))$ with

$$D_t p(t) = -\mathscr{A}^* p(t),$$

(iii) $M^* p_{\partial\Omega}(t) = 0$, all $t \in T$.

Part (i) follows simply from the strong continuity of $t \mapsto y(t)$ and the assumed strong continuity from the right and boundedness of $t \mapsto \kappa^{-1}P(t)$: $T \to \mathscr{L}(L^2; H^1)$. It is less straightforward to prove (ii); in outline we show that $t \mapsto p(t)$: $T \to L^2$ is weakly differentiable to $-\mathscr{A}^* p(t)$ *everywhere* on $[0, t_1[$ from the right. This involves developing an identity[7] similar to (8.3) for $p(t)$ as defined here. Since $-\mathscr{A}^*$ is a first order operator and $p(t) \in L^\infty(T; H^1)$, we have that $\{t \mapsto -A^* p(t)\} \in L^\infty(T; L^2)$. But by a refinement of a result in [8] (see [12]), a square summable function $t \mapsto q(t)$ which is *everywhere*[8] weakly differentiable from the right to a square summable function lies in $H^1(T; L^2(\Omega))$ with $D_t q(t) = \partial_t^+ q(t)$ (∂_t^+ = weak right derivative). This establishes (ii). Part (iii) is immediate from the assumptions on $\mathscr{P}$.

Step 3. Define $u^* \in L^2(\Sigma; \mathbb{R}^n)$ as $u^*(t) = A^-(\mathscr{P}(t)y(t))^-_{\partial\Omega}$ a.e. $t \in T$. We next show that

$$(8.5) \quad \langle (y - z)(t_1), R(y^u - y)(t_1)\rangle_\Omega + \langle u - u^*, u^*\rangle_\Sigma = 0, \qquad u \in L^2(\Sigma; \mathbb{R}^n).$$

Thus y is identified as the strong solution to the state equation corresponding to the control u^*. Equation (8.5) will be recognized as the variational equality characterizing the optimal control.

We deduce (8.5) from the properties of $p(t)$ established in Step 2; the crucial step (see [11] for details) is in justifying the use of the divergence theorem as in the proof of Proposition 5.1 *even though* we do not know a priori that $p(t)$ is a strong solution to the adjoint problem, or indeed even defines an element in $\mathscr{F}^*$.

Step 4. It is immediate from (8.5) that u^* is the optimal control. We must still do some work, however, to establish that $\mathscr{P}(t)$ (not merely $t \mapsto A^-(\mathscr{P}(t)y(t))^-_{\partial\Omega}$) is uniquely defined. This is accomplished by a Holmgren-type argument (see [11] for details). The proof is completed by noting that $\mathscr{P}(t)$ has the unique representation

$$\mathscr{P}(t)a = P(t)a + r(t), \qquad a \in L^2(\Omega).$$

[7] It is here that we require $P(t) = P^*(t)$.

[8] It is precisely because we cannot relax this condition to a.e. differentiability that we need to hypothesize that $t \mapsto \kappa^{-1}P(t)$ is strongly continuous from the right.

9. Concluding remarks. The Introduction indicates some respects in which the present study is incomplete. Most notably we should like to synthesize the boundary control in feedback form for such cost functions as

$$J(u)=\int_T \langle y^u(t), Q(t)y^u(t)\rangle_\Omega\, dt+\langle y^u(t_1), Ry^u(t_1)\rangle_\Omega+\langle u, u\rangle_\Sigma.$$

In this situation, results in [9] would indicate that we should replace (7.3) by

$$\frac{d}{dt}P(x,x',t)=-\mathscr{A}_x^*P(x,x',t)-P(x,x',t)\mathscr{A}_{x'}-Q(x,x',t)$$
$$+\int_{\partial\Omega} P(x,\sigma,t)^-(A^-)^2P(\sigma,x',t)^-\,d\sigma, \tag{7.3'}$$

where $Q(x,x',t)$ is the kernel of $Q(t)$. However, (7.3') is not meaningful as it stands[9] because with no assurance that $P(t)$ maps $\mathscr{D}^n$ into H^1, neither the last term in (7.3') nor the boundary condition $M^*P(t)=0$ is well-defined.

Acknowledgment. The authors gratefully acknowledge the assistance of Professor S. K. Mitter.

REFERENCES

[1] R. W. CARROLL, *Abstract Methods in Partial Differential Equations*, Harper and Row, New York, 1969.
[2] R. HERSH, *Mixed problems in several variables*, J. Math. Mech., 12 (1963), pp. 317–334.
[3] T. L. JOHNSON, *Optimal control of first order distributed systems*, M.I.T. Electronic Systems Laboratory Rep. 482, Cambridge, Mass., 1972.
[4] H. O. KREISS, *Initial boundary value problems for hyperbolic systems*, Comm. Pure Appl. Math., 13 (1970), pp. 277–298.
[5] J. L. LIONS, *Optimal control of systems governed by partial differential equations*, transl. by S. K. Mitter, Springer, Berlin, 1971.
[6] J. L. LIONS AND E. MAGENES, *Non-homogeneous boundary value problems and applications*, 1, transl. by F. Kenneth, Springer, Berlin, 1972.
[7] J. RAUCH, *L^2 is a continuable initial condition for Kreiss' mixed problems*, Comm. Pure Appl. Math., 25 (1972), pp. 265–285.
[8] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
[9] D. L. RUSSELL, *Quadratic performance criteria in boundary control of linear symmetric hyperbolic systems*, this Journal, 11 (1973), pp. 475–509.
[10] L. SCHWARTZ, *Theorie des Noyaux*, Proc. Internat. Congress of Mathematicians, 1950, pp. 220–230.
[11] R. B. VINTER, *Optimal control of non-symmetric hyperbolic systems in n variables on the half-space*, Imperial College Tech. Rep. 74/63, London, England, 1974.
[12] ———, *Some results concerning perturbed evolution equations with applications to delay systems*, Imperial College Tech. Rep. 74/62, London, England, 1974.

[9] We would also need to attach meaning to (6.3).

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

ORDERABLE SET FUNCTIONS AND CONTINUITY. II: SET FUNCTIONS WITH INFINITELY MANY NULL POINTS*

URIEL G. ROTHBLUM†

Abstract. A set function (which is not necessarily additive) on a measurable space I is called orderable if for each measurable order $\mathcal{R}$ there is a measure $\varphi^{\mathcal{R}}v$ on I such that for all initial segments J, $(\varphi^{\mathcal{R}}v)(J) = v(J)$. Properties of orderable set functions v which have infinitely many null points are investigated in this paper. We show that such set functions are continuous and that a set A is v-null if and only if $|\varphi^{\mathcal{R}}v|(A) = 0$ for all measurable orders $\mathcal{R}$. A characterization of orderable nonatomic set functions as well as a characterization of weakly continuous set functions which have a mixing value are given. It is also shown that if a set function is weakly continuous with respect to a measure, then it is weakly equivalent to some measure.

1. Introduction. Let $\mathcal{R}$ be an order on a measurable space I. An *initial segment* of $\mathcal{R}$ is a set of the form $\{t \in I | s\mathcal{R}t\}$. The order $\mathcal{R}$ is *measurable* if the σ-field generated by the initial segments of $\mathcal{R}$ is the σ-field of all measurable sets. A (not necessarily additive) set function v on I is *orderable* if for each measurable order $\mathcal{R}$ there is a measure $\varphi^{\mathcal{R}}v$ such that for all subsets J of I that are initial segments in the order $\mathcal{R}$, we have $(\varphi^{\mathcal{R}}v)(J) = v(J)$.

To understand orderability intuitively, think of I as consisting of an (inhomogeneous) liquid, and of $v(S)$ as representing some (not necessarily additive) measure of the "worth" of a particular part S of I. Think of this liquid as flowing from one place to another, the drops arriving in the order $\mathcal{R}$. As it arrives, each drop of the liquid contributes to (or detracts from) the worth of that portion of the liquid already at the destination. Intuitively, $(\varphi^{\mathcal{R}}v)(S)$ is the total increment contributed in this way by all the drops in a set S. Since v is in general not additive, $\varphi^{\mathcal{R}}v$ will depend strongly on $\mathcal{R}$; and in fact, it may not even exist for all $\mathcal{R}$. Orderable v's are those for which it does. The reader is referred to [2, Chap. III] for an explanation of how these notions are motivated by game-theoretic considerations

This is one of a series of studies (cf. [1], [2], [6]) in which orderability and various continuity notions of set functions are investigated and related to each other.

A subset A of I is called v-null if $v(S \setminus A) = v(S)$ for all subsets S of I. A point t in I is v-null if $\{t\}$ is v-null. The set of set functions which have infinitely many null points is denoted INP.[1] Properties of orderable set functions in INP are investigated in this paper.[2] It is shown (Theorem 1) that such set functions are continuous;[3] i.e., for an increasing (or decreasing) sequence $\{B_i\}$ of measurable sets whose union (or intersection) is B, $\lim_{i\to\infty} v(B_i) = v(B)$. It is also shown

* Received by the editors August 14, 1975.

† School of Organization and Management, Yale University, New Haven, Connecticut 06520. This work was supported by the National Science Foundation under Grant GS-3269 at the Institute for Mathematical Studies in the Social Sciences, Stanford University, Stanford, California, and under Grant GP-37069 at the Courant Institute of Mathematical Sciences, New York, New York.

[1] In the game theoretical motivation INP means infinitely many null players.

[2] We remark that the assumption that $v \in$ INP appeared in a footnote in [2, § 2] as a condition under which any value of a set function gives zero to null sets.

(Theorem 2) that for orderable set functions in INP, $A \subseteq I$ is v-null if and only if A is $\varphi^{\mathcal{R}}v$-null for all measurable orders $\mathcal{R}$. These, and some other properties which we shall prove for orderable set functions in INP, seem to be essential for any solution concept in the game theoretical motivation.

A set function v is nonatomic if there exists no S which is not v-null, such that for every $T \subseteq S$ either T or $S \setminus T$ is v-null. It is shown (Theorem 3 of this paper due to Aumann) that an orderable set function is nonatomic if and only if every $t \in I$ is v-null. Using this result we shall prove that an orderable set function is nonatomic if and only if $\varphi^{\mathcal{R}}v$ is nonatomic for every measurable order $\mathcal{R}$.

Another result (§ 6) is as follows: we say that v is *weakly continuous* with respect to a measure μ (see [6]) if

$$\mu(A) = 0 \Rightarrow A \text{ is } v\text{-null}, \tag{1.1}$$

and that v is *weakly equivalent* to μ if

$$\mu(A) = 0 \Leftrightarrow A \text{ is } v\text{-null}. \tag{1.2}$$

The result (Theorem 4) then says that an orderable set function is weakly continuous with respect to some measure if and only if it is weakly equivalent to some measure (not necessarily the same one). A Lebesgue decomposition of orderable set functions which are weakly continuous with respect to some measure follows as an immediate corollary.

Finally (Theorem 5), we characterize set functions in MIX (see [2, Chap. 2]) which are weakly continuous with respect to some nonatomic measure.

2. Notations and definitions. Composition will usually be denoted by $\circ$; thus if f is defined on the range of μ, then the function whose value on S is $f(\mu(S))$ will be denoted $f \circ \mu$. Set theoretic subtraction will be denoted by $\setminus$. A *measure* is a σ-additive real-valued set function defined on a field, which vanishes on $\varnothing$. The total variation of a measure μ on a measurable set S is denoted $|\mu|(S)$. Absolute continuity between measures will be denoted by $\ll$ (see [3]).

We next summarize some definitions, conventions and results from [2].

Let $(I, \mathcal{C})$ be the measurable space consisting of the unit interval and the Borel subsets.[4] A *set function* is a real valued function v on $\mathcal{C}$ such that $v(\varnothing) = 0$. A set $S \in \mathcal{C}$ is *null* (or v-null) if $v(T \setminus S) = v(T)$ for all $T \in \mathcal{C}$. An *atom* of a set function v is a nonnull measurable set S, such that for every measurable set $T \subseteq S$, either T or $S \setminus T$ is v-null. If v has no atoms it is called *nonatomic*. Restricted to measures, this definition coincides with the usual concept on nonatomicity of measures.[5] A set function is *monotonic* if $T \subseteq S$ implies $v(T) \leqq v(S)$. The difference between two monotonic set functions is said to be *of bounded variation*. The set of all set functions of bounded variation forms a linear space, which is called

---

[3] Cf. [2, Ex. 33.11].

[4] This is assumed for simplicity only. All the results remain true if $(I, \mathcal{C})$ is only assumed to be a countably generated and separated Borel space.

[5] The definition of nonatomicity in this paper is different from the one used in [5]. It coincides with that of [2]. Theorem 3 of this paper shows that the two definitions coincide for orderable set functions.

BV. The linear subspace of BV consisting of all totally finite measures on $(I, \mathscr{C})$ is denoted M. The linear subspace of M consisting of nonatomic measures is denoted NA. The set of monotonic elements in M (resp., NA) will be denoted M^+ (resp., NA^+).

An *order* on the underlying space I is a relation $\mathscr{R}$ on R that is transitive, irreflexive, and complete.[6] Let "$s \overset{\mathscr{R}}{\geqq} t$" denote "$s\mathscr{R}t$ or $s = t$". If for A, $B \subseteq I$ it holds that $x\mathscr{R}y$ whenever $x \in A$, $y \in B$ we will write $A\mathscr{R}B$. If A contains a single element z, we write $B\mathscr{R}z$ (resp., $z\mathscr{R}B$) rather than $B\mathscr{R}\{z\}$ (resp., $\{z\}\mathscr{R}B$). An *initial segment* is a set of the form $I(s, \mathscr{R}) = \{t | s\mathscr{R}t\}$ where $s \in I$. An *initial set* is a set J which fulfills the condition $s \in J$, $s\mathscr{R}s'$ implies $s' \in J$. The entire space and the empty set will also be considered as initial segments, and as such will be denoted $I(\infty, \mathscr{R})$, $I(-\infty, \mathscr{R})$ respectively; it will be understood $\infty\mathscr{R}s\mathscr{R} - \infty$ for each $s \in I$ and we will denote $\{-\infty\} \cup I \cup \{\infty\}$ by $\bar{I}$. (Formally we extend $\mathscr{R}$ to $\bar{I}$. This however is a notational device; we are not adding anything to the underlying space, and all set functions and measures continue to be defined on subsets of I only.) For $s, x \in I$ let $E(s, \mathscr{R}) = \{t | t\mathscr{R}s\}$ be called a *final set* and let $[s, x] = \{t | x \overset{\mathscr{R}}{\geqq} t \overset{\mathscr{R}}{\geqq} s\}$ be called a *closed order* interval.

Denote by $F(\mathscr{R})$ the σ-field generated by all the initial segments. An order $\mathscr{R}$ is *measurable* if $F(\mathscr{R}) = \mathscr{C}$. A subset Q of I will be called *$\mathscr{R}$-dense* if for all $s, t \in I$ such that $s\mathscr{R}t$ there is a member $q \in Q$ such that $s \overset{\mathscr{R}}{\geqq} q \overset{\mathscr{R}}{\geqq} t$. By Lemma 12.5 of [2], for any measurable order $\mathscr{R}$ there exists a denumerable $\mathscr{R}$-dense set. A set function v is called *orderable* if for each measureable order $\mathscr{R}$ there is a measure $\varphi^{\mathscr{R}}v$ such that for all initial segments $I(s, \mathscr{R})$, we have

$$(\varphi^{\mathscr{R}}v)(I(s, \mathscr{R})) = v(I(s, \mathscr{R})). \tag{2.1}$$

Since (2.1) determines $\varphi^{\mathscr{R}}v$ on all the initial segments, and by the measurability of $\mathscr{R}$ the initial segments generate $\mathscr{R}$, it follows that there can be at most one measure $\varphi^{\mathscr{R}}v$ satisfying (2.1). Thus for orderable set functions there is exactly one measure $\varphi^{\mathscr{R}}v$ satisfying (2.1). The set of all orderable set functions in BV will be denoted ORD.

Let $v, w \in \mathrm{BV}$; then v is said to be *weakly continuous with respect to w* (written $v \underset{w}{\lessdot} w$) [6, §§ 3 and 4] if for any $S \in \mathscr{C}$,

$$S \text{ is } w\text{-null} \Rightarrow S \text{ is } v\text{-null}. \tag{2.2}$$

Note that if $v \underset{w}{\lessdot} u$ and $u \underset{w}{\lessdot} w$, where $v, u, w \in \mathrm{BV}$ then $v \underset{w}{\lessdot} w$. Of course if $v = f \circ \mu \in \mathrm{BV}$ where $\mu \in M^+$ and f maps the range of μ into the reals, then $v \underset{w}{\lessdot} \mu$. A set function in BV is said to be *weakly continuous* if there is a measure $\mu \in \mathrm{NA}^+$ such that $v \underset{w}{\lessdot} \mu$. The set of all weakly continuous set functions is a linear subspace of BV [6, Prop. 4.2] which is denoted WC.

3. Continuity of set functioning in INP ∩ ORD.

LEMMA 1. *Let $\{B_n\}$, $n \geqq 1$, be an increasing sequence of measurable sets, with $B = \bigcup_{n=1}^{\infty} B_n$. Then there exists a measurable order $\mathscr{R}$ such that B and all the B_n's are $\mathscr{R}$-initial sets.*

---

[6] A relation $\mathscr{R}$ is complete if for all $s, t \in I$ one and only one of the three statements $s\mathscr{R}t$, $t\mathscr{R}s$, $s = t$ holds. We shall interpret "$s\mathscr{R}t$" as "s is greater than t."

Proof. Let $\mathcal{R}_1$ be a measurable order such that $(I \setminus B)\mathcal{R}(B \setminus B_1)\mathcal{R}B_1$. The existence of $\mathcal{R}_1$ follows from [6, Cor. 5.2.]. Define, inductively, a sequence of measureable orders by

$$x\mathcal{R}_{n+1}y \Leftrightarrow \begin{cases} x, y \in B_n \text{ and } x\mathcal{R}_n y, \text{ or} \\ x, y \in B_{n+1} \setminus B_n \text{ and } x\mathcal{R}_n y, \text{ or} \\ x, y \in I \setminus B_{n+1} \text{ and } x\mathcal{R}_n y, \text{ or} \\ y \in B_n \text{ and } x \in I \setminus B_n, \text{ or} \\ y \in B_{n+1} \setminus B_n \text{ and } x \in I \setminus B_{n+1}. \end{cases}$$

This means that $B_{n+1} \setminus B_n$ is put just beyond B_n, $I \setminus B_{n+1}$ beyond $B_{n+1} \setminus B_n$, and the order $\mathcal{R}_n$ is preserved on B_n, $B_{n+1} \setminus B_n$ and $I \setminus B_{n+1}$. By [6, Lem. 5.1] all the orders $\mathcal{R}_n$ are measurable; moreover, for all $n \geqq 1$, and $1 \leqq i \leqq n$, B_i and B are $\mathcal{R}$-initial sets. Of course, for $m \geqq n$ all $\mathcal{R}_m$ coincide on $I \setminus (B \setminus B_n)$. Let us define an order $\mathcal{R}$ as follows: Let s, t be in I; then there clearly exists an n such that $s, t \notin B \setminus B_n$; let $s\mathcal{R}t$ if and only if $s\mathcal{R}_m t$ for all $m \geqq n$. Clearly $\mathcal{R}$ is well defined and all the B_i's are $\mathcal{R}$-initial sets. We shall now show that $\mathcal{R}$ is a measurable order, i.e., $F(\mathcal{R}) = \mathcal{C}$. The direction $F(\mathcal{R}) \subseteq \mathcal{C}$ is trivial. To verify that $\mathcal{C} \subseteq F(\mathcal{R})$ note first that $\mathcal{R}$ has a denumerable dense set, e.g., the union of the denumerable $\mathcal{R}_n$-dense sets. This implies that all $\mathcal{R}$-initial sets are in $F(\mathcal{R})$ (compare with the proof of Lemma 5.1 of [6]). Now, for $x \in I$, the decomposition

$$I(x, \mathcal{R}_1) = \{I(x, \mathcal{R}_1) \cap B_1\} \cup \{I(x, \mathcal{R}_1) \setminus B\} \cup \bigcup_{n=1}^{\infty} \{I(x, \mathcal{R}_1) \cap (B_{n+1} \setminus B_n)\}$$

shows that $I(x, \mathcal{R}_1) \in \mathcal{R})$. Since $\mathcal{C} = F(\mathcal{R}_1)$ this completes the proof of Lemma 1.

COROLLARY 1. *Let $\{B_n\}$, $n \geqq 1$, be an increasing sequence of measurable sets, such that $B = \bigcup_{n=1}^{\infty} B_n$. Then there exists a measurable order $\mathcal{R}$ such that B and all the B_n's are $\mathcal{R}$-initial segments.*

Proof. Without loss of generality assume that B_n's are strictly increasing. For $n \geqq 1$ let $x_n \in B_{n+1} \setminus B_n$. Define $C_{2n} = B_n \cup \{x_n\}$, $C_{2n-1} = B_n$. Applying Lemma 1 after reordering $I \setminus B$ so that it will have an $\mathcal{R}_1$-initial element completes the proof.

COROLLARY 2. *Let $v \in \mathrm{ORD}$, and let $\{B_n\}$, $n \geqq 1$, be an increasing sequence of measurable sets with $B = \bigcup_{n=1}^{\infty} B_n$. Then $v(B_n) \to v(B)$ as $n \to \infty$.*

Proof. By Corollary 1 we know the existence of a measurable order $\mathcal{R}$ such that B and all the B_n's are $\mathcal{R}$-initial segments. Hence $(\varphi^{\mathcal{R}} v)(B) = v(B)$, and for $n \geqq 1$, $(\varphi^{\mathcal{R}} v)(B_n) = v(B_n)$. Since $\varphi^{\mathcal{R}} v$ is a σ-additive measure $(\varphi^{\mathcal{R}} v)(B_n) \to (\varphi^{\mathcal{R}} v)(B)$ as $n \to \infty$, completing the proof.

COROLLARY 3. *Let $v \in \mathrm{ORD}$. Then a countable union of v-null sets is v-null.*

Proof. Let A_n, $n \geqq 1$, be v-null and let $B \in \mathcal{C}$. For every $i \geqq 1$,

$$v\Big(B \setminus \bigcup_{n=1}^{\infty} A_n\Big) = v\Big\{\Big(B \setminus \bigcup_{n=1}^{\infty} A_n\Big) \cup \bigcup_{k=1}^{i} (A_k \cap B)\Big\}.$$

Letting $i \to \infty$ and using Corollary 2 implies that

$$v\Big(B \setminus \bigcup_{n=1}^{\infty} A_n\Big) = \lim_{i\to\infty} v\Big\{\Big(B \setminus \bigcup_{n=1}^{\infty} A_n\Big) \cup \bigcup_{k=1}^{i} (A_k \cap B)\Big\}$$

$$= v\Big\{\Big(B \setminus \bigcup_{n=1}^{\infty} A_n\Big) \cup \bigcup_{k=1}^{\infty} (A_k \cap B)\Big\} = v(B).$$

Remark. If $\{B_n\}$, $n \geqq 1$, is a decreasing sequence of measurable sets we may build an order such that all the B_n's are $\mathcal{R}$-initial segments. This would be done by steps analogous to those leading to the conclusion of Lemma 1 and Corollary 1. One can easily verify that $B = \cap_{n=1}^{\infty} B_n$ is not an $\mathcal{R}$-initial segment unless there is some n such that for $i > n$ all the B_i's coincide; hence the proof used in Corollary 2 would not be sufficient to show that $v(B_n) \to v(B)$ as $n \to \infty$. Indeed, in general we cannot assure that $v(B_n) \to v(B)$ as $n \to \infty$. Let v be defined by

$$v(S) = \begin{cases} 1 & \text{if } 0 \in S \text{ and } S \neq \{0\}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Note that v is monotonic; hence $v \in \mathrm{BV}$. Moreover, $v \in \mathrm{ORD}$; indeed, for a measurable order $\mathcal{R}$ in which 0 is not an $\mathcal{R}$-smallest element, $\varphi^{\mathcal{R}} v$ is the measure which is concentrated on $\{0\}$; the same is true if 0 is an $\mathcal{R}$-smallest element and there is no $\mathcal{R}$-smallest element in $E(0, \mathcal{R})$. In the case when b is an $\mathcal{R}$-smallest element in $E(0, \mathcal{R})$, then $\varphi^{\mathcal{R}} v$ exists and is equal to the measure which is concentrated on $\{b\}$. Let now $B_n = [0, 1/n)$; then $(B_n\}$ is clearly decreasing and $\cap_{n=1}^{\infty} B_n = \{0\}$. For all $n \geqq 1$, $v(B_n) = 1$, but $v(B) = 0$; hence $\lim_{n\to\infty} v(B_n) \neq v(B)$.

LEMMA 2. *Let $v \in \mathrm{ORD}$, $\mathcal{R}$ a measurable order, and J an $\mathcal{R}$-initial set. Then J is measurable. Moreover, if J is infinite or if $v \in \mathrm{INP}$, then*

$$(\varphi^{\mathcal{R}} v)(J) = v(J). \tag{3.2}$$

Remark. The infiniteness requirement of J is a surprising condition. To verify its necessity look at the example given in (3.1) of the previous remark. Let $\mathcal{R}$ be the usual order on $[0, 1]$ which is clearly measurable; then $(\varphi^{\mathcal{R}} v)(\{0\}) = 1$ but $v(\{0\}) = 0$.

Note also that (3.2) need not hold for infinite J's if we do not assume $v \in \mathrm{ORD}$, even if we do assume that for the $\mathcal{R}$ in question, there is a σ-additive totally finite measure $\varphi^{\mathcal{R}} v$ satisfying (2.1)! For example, let

$$w = f \circ \lambda,$$

where λ is the Lebesgue measure and

$$f(x) = \begin{cases} 0, & x \leqq \frac{1}{2}, \\ 1, & x > \frac{1}{2}, \end{cases}$$

and let $\mathcal{R}$ be the usual order; it is clear that $\varphi^{\mathcal{R}} v$ exists and equals the measure concentrated on $\{1/2\}$. Hence $(\varphi^{\mathcal{R}} v)([0, 1/2]) = 1$ but $v([0, 1/2]) = 0$. It might easily be shown that $v \notin \mathrm{ORD}$. Let $\mathcal{R}'$ be the order that throws $\{1/2\}$ beyond $[0, 1]$ and coincides with the usual order on $[0, 1] \setminus \{1/2\}$. By Lemma 5.1 of [6] this order is measurable. If $\varphi^{\mathcal{R}'} v$ existed, then for $n \geqq 3$, $(\varphi^{\mathcal{R}'} v)([1/2 - 1/n,$

$1/2+1/n]_{\mathcal{R}'}) = 1$, in spite of the fact that $[1/2-1/n, 1/2+1/n]_{\mathcal{R}'}$ is a decreasing sequence with a void intersection.

Proof of the lemma. The measurability of J follows from Lemma 12.14 of [2]. If J is an $\mathcal{R}$-initial segment, then the conclusion of Lemma 2 is trivial. If J is not an $\mathcal{R}$-initial segment, let Q be a countable $\mathcal{R}$-dense set and denote $\bar{Q} = Q \cup \{-\infty\} \cup \{\infty\}$ and $\bar{J} = J \cup \{-\infty\}$. Clearly $J = \cap\{I(q, \mathcal{R}) | q \in \bar{Q} \setminus \bar{J}\}$. Since the $I(q, \mathcal{R})$'s are linearly ordered under inclusion, each finite intersection of those sets is equal to one of the $I(q, \mathcal{R})$'s. Hence we can write $J = \bigcap_{j=1}^{\infty} I(q_j, \mathcal{R})$, where $\{q_j\}$ is an $\mathcal{R}$-decreasing sequence of points in $\bar{Q} \setminus \bar{J}$, i.e., $\{I(q_j, \mathcal{R})\}$ is a decreasing sequence of sets. Note that $\{I(q_j, \mathcal{R})\}$ is not a finite sequence since J is not an $\mathcal{R}$-initial segment.

Assuming J is not finite let us choose a sequence $\{x_i | i \geqq 1\}$ of elements in J. Let $\mathcal{R}^*$ be a measurable order for which

$$(I \setminus J)\mathcal{R}^* \cdots \mathcal{R}^*\{x_n\}\mathcal{R}^* \cdots \mathcal{R}^*\{x_1\}\mathcal{R}^*(J \setminus \{x_i | i \geqq 1\})$$

and $\mathcal{R}^*$ coincides with $\mathcal{R}$ on $J \setminus \{x_i | i \geqq 1\}$ and on $I \setminus J$. Similar arguments to those used in the proof of Lemma 1 imply that $\mathcal{R}^*$ is measurable. Now

$$\begin{aligned}(\varphi^{\mathcal{R}} v)(J) = \lim_{j\to\infty} (\varphi^{\mathcal{R}} v)(I(q_j, \mathcal{R})) &= \lim_{j\to\infty} (\varphi^{\mathcal{R}^*} v)(I(q_j, \mathcal{R}^*)) \\ &= (\varphi^{\mathcal{R}^*} v)(J) = \lim_{i\to\infty} (\varphi^{\mathcal{R}^*} v)(I(x_i, \mathcal{R}^*)) \\ &= \lim_{i\to\infty} v(I(x_i, \mathcal{R}^*)) = v(J)\end{aligned}$$

We have used Corollary 2 and the fact that $I(x_i, \mathcal{R}^*)$ is an increasing sequence whose union is J.

Assume now that J is finite and that $v \in \mathrm{INP}$. Let $\{x_i\}$, $i \geqq 1$, be a sequence of v-null elements which are all in $I \setminus J$. Let $\mathcal{R}^*$ be the measurable order for which

$$I \setminus (J \cup \{x_i | i \geqq 1\})\mathcal{R}^* J \mathcal{R}^* \cdots \mathcal{R}^* x_n \mathcal{R}^* \cdots \mathcal{R}^* x_2 \mathcal{R}^* x_1$$

and which coincides with $\mathcal{R}$ on J and on $I \setminus (J \cup \{x_i | i \geqq 1\})$. Note that by Corollary 3 $\{x_i | i \geqq 1\}$ is v-null (as a countable union of v-null sets). Now,

$$\begin{aligned}(\varphi^{\mathcal{R}} v)(J) = \lim_{j\to\infty} (\varphi^{\mathcal{R}} v)(I(q_j, \mathcal{R})) &= \lim_{j\to\infty} v(I(q_j, \mathcal{R})) \\ &= \lim_{j\to\infty} v(I(q_j, \mathcal{R}) \cup \{x_i | i \geqq 1\}) \\ &= \lim_{j\to\infty} (\varphi^{\mathcal{R}^*} v)(I(q_j, \mathcal{R}) \cup \{x_i | i \geqq 1\}) \\ &= (\varphi^{\mathcal{R}^*} v)(J \cup \{x_i | i \geqq 1\}) \\ &= v(J \cup \{x_i | i \geqq 1\}) = v(J).\end{aligned}$$

We used the facts that $\{x_i | i \geqq 1\}$ is v-null, and the fact that since (for $j \geqq 1$) $I(q_j, \mathcal{R}) \cup \{x_i | i \geqq 1\}$ and $J \cup \{x_i | i \geqq 1\}$ are infinite $\mathcal{R}^*$-initial sets, v and $\varphi^{\mathcal{R}^*} v$ coincide on them. Thus the proof of Lemma 2 is completed.

LEMMA 3. *Let* $v \in \mathrm{ORD} \cap \mathrm{INP}$, *and let* $\{B_n\}$, $n \geqq 1$, *be a decreasing sequence of measurable sets whose intersections is* B. *Then* $\lim_{n\to\infty} v(B_n) = v(B)$.

Remark. Look at the remark following the proof of Corollary 3 to verify the necessity of the requirement that $v \in \mathrm{INP}$.

Proof. By similar arguments to those used in the proof of Lemma 1 there exists a measurable order $\mathscr{R}$ such that B and all the B_n's are initial sets (not necessarily initial segments). Using Lemma 2 and the σ-additivity of $\varphi^{\mathscr{R}}v$ we get that

$$\lim_{n\to\infty} v(B_n) = \lim_{n\to\infty} (\varphi^{\mathscr{R}}v)(B_n) = (\varphi^{\mathscr{R}}v)(B) = v(B),$$

which completes the proof of Lemma 3.

THEOREM 1. *If* $v \in \mathrm{ORD} \cap \mathrm{INP}$, *then* v *is continuous.*

Proof. The proof of Theorem 1 follows directly by Corollary 2 and Lemma 3.

4. Nullness of sets with respect to v and the $\varphi^{\mathscr{R}}v$'s.

LEMMA 4. *Let* $v \in \mathrm{ORD} \cap \mathrm{INP}$, *and let* A *be a* v*-null set. If* $\mathscr{R}$ *and* $\mathscr{R}'$ *are two measurable orders such that for* $s, t \in I \setminus A$

$$s\mathscr{R}t \Rightarrow s\mathscr{R}'t, \tag{4.1}$$

then

$$\varphi^{\mathscr{R}}v = \varphi^{\mathscr{R}'}v. \tag{4.2}$$

Remark. If $v \notin \mathrm{INP}$, then (4.1) becomes a very restrictive condition, since the v-null sets are only subsets of some finite set; hence $\mathscr{R}$ and $\mathscr{R}'$ are "almost" the same. However, we cannot assure (4.2) for those simple cases. Indeed, let v be defined by (3.1), and let w be defined by $w(S) = v(S \setminus \{1\})$. One can easily verify that w is orderable and that $\{1\}$ is a v-null set. Let $\mathscr{R}$ be the regular order and let $\mathscr{R}'$ be defined by

$$s\mathscr{R}'t \Leftrightarrow \begin{cases} s > t \text{ and } s, t \neq 1, \text{ or} \\ t = 1 \text{ and } s \neq 0, \text{ or} \\ t = 0 \text{ and } s = 1; \end{cases}$$

i.e., $\{1\}$ is put between $\{0\}$ and $(0, 1)$ and the usual order is preserved on $(0, 1)$. Clearly $\varphi^{\mathscr{R}'}w$ is the measure concentrated on $\{1\}$ and $\varphi^{\mathscr{R}}w$ is the measure concentrated on $\{0\}$, though (4.1) holds whenever $s, t \in [0, 1] \setminus \{1\}$, and $\{1\}$ is w-null.

Proof. It is clearly sufficient to prove our lemma for the case when $\mathscr{R}'$ "throws" A beyond $I \setminus A$ and preserves $\mathscr{R}$ on A and on $I \setminus A$. We shall first show that $\varphi^{\mathscr{R}}w$ and $\varphi^{\mathscr{R}'}w$ coincide on $\mathscr{R}$-initial segments; i.e.,

$$(\varphi^{\mathscr{R}}w)(I(x, \mathscr{R})) = (\varphi^{\mathscr{R}'}w)(I(x, \mathscr{R})) \tag{4.3}$$

for all $x \in I$.

Let $x \in I \setminus A$; then $I(x, \mathscr{R}') = I(x, \mathscr{R}) \setminus A$ and

$$\begin{aligned} (\varphi^{\mathscr{R}}v)(I(x, \mathscr{R})) &= v(I(x, \mathscr{R})) \\ &= v(I(x, \mathscr{R}) \setminus A) = v(I(x, \mathscr{R}')) \\ &= (\varphi^{\mathscr{R}'}v)(I(x, \mathscr{R})) - (\varphi^{\mathscr{R}'}v)(I(x, \mathscr{R}) \cap A). \end{aligned} \tag{4.4}$$

Now, note that $I(x, \mathcal{R}) \cap A$ is the set subtraction of two $\mathcal{R}'$-initial sets on which v coincides; hence, by Lemma 2 we get that

$$(\varphi^{\mathcal{R}'}v)(I(x, \mathcal{R}) \cap A) = 0. \tag{4.5}$$

The above and (4.4) complete the proof of (4.3) for all $x \in I \setminus A$.

Let now $x \in A$. By applying Lemma 2 on $\mathcal{R}'$ we get that

$$\begin{aligned}(\varphi^{\mathcal{R}}v)(I(x, \mathcal{R})) &= v(I(x, \mathcal{R})) \\ &= v(I(x, \mathcal{R}) \setminus A) = (\varphi^{\mathcal{R}'}v)(I(x, \mathcal{R}) \setminus A) \\ &= (\varphi^{\mathcal{R}'}v)(I(x, \mathcal{R})) - (\varphi^{\mathcal{R}'}v)(I(x, \mathcal{R}) \cap A).\end{aligned} \tag{4.6}$$

Again, $I(x, \mathcal{R}) \cap A$ is the set subtraction of two $\mathcal{R}'$-initial sets on which v coincides. By Lemma 2 it follows again that (4.5) holds, hence (4.3) follows easily from (4.6). This completes the proof of (4.3) for all $x \in I$.

The $I(x, \mathcal{R})$'s generate all the measurable sets, and $\varphi^{\mathcal{R}}v$, $\varphi^{\mathcal{R}'}v$ are two measures which coincide on all the $I(x, \mathcal{R})$'s; hence $(\varphi^{\mathcal{R}}v)(S) = (\varphi^{\mathcal{R}'}v)(S)$ for all $S \in \mathcal{C}$ as was to be proved.

THEOREM 2. *Let* $v \in \mathrm{ORD} \cap \mathrm{INP}$, *and let* $A \in \mathcal{C}$. *Then*

$$A \text{ is } v\text{-null if and only if } A \text{ is } \varphi^{\mathcal{R}}v\text{-null for all measurable orders.} \tag{4.7}$$

Remark. If $v \notin \mathrm{INP}$, then v-null sets are only subsets of some finite set. However (4.7) need not hold for these simple sets. Indeed, look at the first example preceding the proof of Lemma 4 and verify that $(\varphi^{\mathcal{R}'}w)(\{1\}) = 1$ though $\{1\}$ is v-null.

If $v \in \mathrm{INP}$ but $v \notin \mathrm{ORD}$, then the conclusion of Theorem 2 need not hold, even if we do assume that for the $\mathcal{R}$ in question there is a σ-additive totally finite measure $\varphi^{\mathcal{R}}v$ satisfying (2.1). See the second example preceding the proof of Lemma 2 and verify that $(\varphi^{\mathcal{R}}v)(\{1/2\}) = 1$ though $\{1/2\}$ is v-null.

Proof. Note that A is $\varphi^{\mathcal{R}}v$-null if and only if $|\varphi^{\mathcal{R}}v|(A) = 0$. First assume that A is $\varphi^{\mathcal{R}}v$-null for all measurable orders $\mathcal{R}$. Assume there exists a $B \in \mathcal{C}$ such that $v(B) \neq v(B \setminus A)$. By Corollary 1 there exists a measurable order $\mathcal{R}$ for which

$$(I \setminus B)\mathcal{R}(B \cap A)\mathcal{R}(B \setminus A).$$

Hence, by Lemma 2,

$$(\varphi^{\mathcal{R}}v)(B \cap A) = (\varphi^{\mathcal{R}}v)(B \setminus (B \setminus A)) = v(B) - v(B \setminus A) \neq 0.$$

The above contradicts the fact that $|\varphi^{\mathcal{R}}v|(A) \neq 0$.

Assume now that A is v-null. Let $B \subseteq A$ and let $\mathcal{R}$ be a measurable order. Let $\mathcal{R}'$ be the measurable order that "throws" B beyond $I \setminus B$ and preserves $\mathcal{R}$ on B and on $I \setminus B$. Since $I \setminus B$ is an $\mathcal{R}'$-initial set it follows by Lemma 2 that

$$\begin{aligned}(\varphi^{\mathcal{R}'}v)(B) &= (\varphi^{\mathcal{R}'}v)(I) - (\varphi^{\mathcal{R}'}v)(I \setminus B) \\ &= v(I) - v(I \setminus B) = 0.\end{aligned}$$

By Lemma 4, $\varphi^{\mathcal{R}}v = \varphi^{\mathcal{R}'}v$, hence $(\varphi^{\mathcal{R}}v)(B) = 0$. Since $(\varphi^{\mathcal{R}}v)(B) = 0$ for all $B \subseteq A$ it follows that $|\varphi^{\mathcal{R}}v|(A) = 0$ which means A is $\varphi^{\mathcal{R}}v$-null.

Remark. We point out that the requirement in Theorems 1 and 2 that the set functions in question be in INP could be removed if one changed the definition of orderability by requiring that $(\varphi^{\mathscr{R}}v)(J) = v(J)$ for initial segments and in addition for sets of the form $J = \{t | s \overset{\mathscr{R}}{=} t\}$.

5. Nonatomicity of orderable set functions.

THEOREM 3 (Aumann). *Let v be an orderable set function. Then v is nonatomic if and only if every $s \in I$ is v-null.*

Proof. If v is nonatomic then clearly $\{s\}$ is v-null for every $s \in I$ (else $\{s\}$ would be an atom).

Assume now that s is v-null for every $s \in I$. We shall prove, by contradiction, that v has no atoms. Let $E \in \mathscr{C}$ be an atom of v; i.e., E is not v-null and for every $F \subseteq E$ either F or $E \setminus F$ is v-null. Since every point is v-null it follows by Corollary 5.4 that every denumerable set is v-null; hence E is nondenumerable.

Assume first that $I \setminus E$ is nondenumerable. It is known (cf. [5, Thms. 2.8 and 2.12]) that any uncountable Borel subset of any Euclidean space, and indeed of any complete separable metric space when considered as a measurable space, is isomorphic[7] to $([0, 1], \beta)$ where β is the Borel field on $[0, 1]$. It follows directly from this theorem that there exists an isomorphism ψ of $(I, \mathscr{C})$ such that $\psi(E) = [0, 1/2)$. For every $n \geqq 1$, $1 \leqq i \leqq 2^n$ define

$$I_{n,i} = [(i-1) \cdot 2^{-n}, i \cdot 2^{-n}].$$

Note that

$$\bigcup_{i=1}^{2^n} \psi^{-1}(I_{n,i}) = E.$$

The fact that E is an atom now implies that for every $n \geqq 1$ there exists a unique $i(n)$, $1 \leqq i(n) \leqq 2^n$, such that $\psi^{-1}(I_{n,i})$ is v-null for all $i \neq i(n)$, and $\psi^{-1}(I_{n,i(n)})$ is not v-null. It immediately follows that $\{I_{n,i(n)}\}$ is a decreasing sequence of intervals. Since the length of these intervals converges to 0, $\bigcap_{n=1}^{\infty} I_{n,i(n)}$ contains at most one point.

Let $B_n = \psi^{-1}(I_{n,i(n)})$; then $\{B_n\}$ is a decreasing sequence of sets, $\bigcap_{n=1}^{\infty} B_n$ contains at most one point and for every $n \geqq 1$, $E \setminus B_n$ is v-null. By Corollary 3, $\bigcup_{n=1}^{\infty} (E \setminus B_n) = E \setminus \bigcap_{n=1}^{\infty} B_n$ is v-null. This is a contradiction to the assumptions that E is not v-null and every single point in I is v-null.

If $I \setminus E$ is denumerable, one can easily conclude that I is an atom. Repeating the previous arguments after replacing $[0, 1/2)$ by $[0, 1]$ one may similarly contradict the assumption that every $s \in I$ is v-null.

Remark. The characterization of nonatomic set functions by the requirement that every $s \in I$ is v-null holds whenever the countable union of null sets is null. The proof follows exactly like the proof of Theorem 3.

COROLLARY 4. *Let $v \in$ ORD. Then v is nonatomic if and only if $\varphi^{\mathscr{R}}v$ is nonatomic for all measurable orders $\mathscr{R}$.*

---

[7] Two measurable spaces are called isomorphic if there is a one-to-one function from one onto the other which is measurable in both directions, i.e., both it and its inverse are measurable; the mapping is called an isomorphism.

Proof. It was proved [6, Thm. 3.1] that if $v \in \mathrm{ORD}$, then every $s \in I$ is v-null if and only if for every measurable order $\mathcal{R}$ every $s \in I$ is $\varphi^{\mathcal{R}}v$-null. This and Theorem 3 immediately imply the conclusion of Corollary 4.

6. Weak equivalence and the Lebesgue decomposition. We are going to extend weak continuity of set functions in BV (see § 2) to sets of set functions. If $\mathcal{H}$ and $\mathcal{K}$ are two sets of set functions in BV and if for every $A \in \mathcal{C}$,

$$A \text{ is } w\text{-null for every } w \in \mathcal{H} \Rightarrow A \text{ is } v\text{-null for every } v \in \mathcal{K},$$

then we say that $\mathcal{K}$ is *weakly continuous with respect to* $\mathcal{H}$ and write $\mathcal{K} \underset{w}{<} \mathcal{H}$. If $\mathcal{K} \underset{w}{<} \mathcal{H}$ and $\mathcal{H} \underset{w}{<} \mathcal{K}$, then the sets $\mathcal{H}$ and $\mathcal{K}$ are called *weakly equivalent* and we shall write $\mathcal{H} \underset{w}{\sim} \mathcal{K}$. If $\mathcal{H}$ contains exactly one set function v, i.e., $\mathcal{H} = \{v\}$, the abbreviated notation $\mathcal{K} \underset{w}{<} v$, $v \underset{w}{<} \mathcal{K}$, and $\mathcal{K} \underset{w}{\sim} v$, will be employed for $\mathcal{K} \underset{w}{<} \mathcal{H}$, $\mathcal{H} \underset{w}{<} \mathcal{K}$, and $\mathcal{K} \underset{w}{\sim} \mathcal{H}$.

THEOREM 4. *Let $v \in \mathrm{ORD}$, $\mu \in M^+$, and $v \underset{w}{<} \mu$. Then there exists an $\eta \in M^+$ such that $v \underset{w}{\sim} \eta$.*

Proof. By [6, Thm. 3.3] we know that $\{\varphi^{\mathcal{R}}v | \mathcal{R} \text{ is a measurable order}\} \underset{w}{\sim} v$. Hence

$$\{|\varphi^{\mathcal{R}}v| \,|\, \mathcal{R} \text{ is a measurable order}\} \underset{w}{\sim} v \underset{w}{<} \mu. \tag{6.1}$$

By Lemma 7 of [4], a set $\mathcal{M} \subseteq M^+$ which is weakly continuous with respect to some measure $\mu \in M^+$ has a weakly equivalent countable subset. Hence, by (6.1), there exists a sequence of measurable orders $\mathcal{R}_n$ such that

$$\{|\varphi^{\mathcal{R}_n}v|\}_{n=1}^{\infty} \underset{w}{\sim} v. \tag{6.2}$$

Let $\eta = \sum_{n=1}^{\infty} 2^{-n}|\varphi^{\mathcal{R}_n}v|$. Since $\|\varphi^{\mathcal{R}}v\|$ is uniformly bounded in $\mathcal{R}$ (see [2, Prop. 12.8]), it follows that η is totally finite, i.e., $\eta \in M$. Obviously,

$$\eta \underset{w}{\sim} \{|\varphi^{\mathcal{R}_n}v|\}_{n=1}^{\infty},$$

and the desired result follows directly from (6.2).

Remark. The meaning of Theorem 4 is that if $v \in \mathrm{ORD}$ and v is weakly continuous with respect to some $\mu \in M^+$, then in the appropriate sense, there exist "minimal" measures with respect to which v is weakly continuous.

Remark. One can easily see that if μ in Theorem 4 is nonatomic then η is also nonatomic.

COROLLARY 5 (Lebesgue decomposition). *Let $v \in \mathrm{ORD}$, $\mu \in M^{\perp}$ where $v \underset{w}{\ll} \mu$. Then, for every $\xi \in M$ there exist measures $\xi^{ac}, \xi^{\perp} \in M$ and a set $A \in \mathcal{C}$ such that $\xi = \xi^{ac} + \xi^{\perp}$, $\xi^{ac} \underset{w}{\ll} v$, A is v-null and $\xi^{\perp}(I \setminus A) = 0$.*

7. A characterization of the mixing value. We start this section by defining the subspace MIX of BV and the mixing value which is defined on this subspace. These definitions were first introduced in §§ 14, 15 of [2].

Let NA^1 be the subset of NA^+ consisting of measures μ for which $\mu(I) = 1$. If $\mu \in \mathrm{NA}^1$, define a *μ-mixing sequence* to be a sequence $\{\theta_1, \theta_2, \cdots\}$ of μ-measure preserving automorphism[8] of $(I, \mathcal{C})$ such that for all $S, T \in \mathcal{C}$ we have

$$\lim_{n\to\infty} \mu(S \cap \theta_n T) = \mu(S) \cdot \mu(T).$$

[8] An automorphism of a measurable space $(I, \mathcal{C})$ is an isomorphism of that space onto itself.

Let $\mathcal{R}$ be a given order and ψ be an automorphism of $(I, \mathcal{C})$. Then $\psi\mathcal{R}$ is the order defined by $(\psi x)\psi\mathcal{R}(\psi y)$ if and only if $x\mathcal{R}y$. Obviously $\psi\mathcal{R}$ is measurable if and only if $\mathcal{R}$ is. For orderable set functions v and measurable orders $\mathcal{R}$ we shall be interested in measures of the form $\varphi^{\psi\mathcal{R}}v$; since this notation will occasionally be cumbersome, we will sometimes write $\varphi(v, \psi\mathcal{R})$ rather than $\varphi^{\psi\mathcal{R}}v$. No confusion should occur.

Let $v \in \mathrm{ORD}$. A set function φv is the *mixing value* of v if there is a measure μ_v in NA^1 such that for $\mu \in \mathrm{NA}^1$ with $\mu_v \ll \mu$,

(7.1) for all μ-mixing sequences $\{\theta_1, \theta_2, \cdots\}$, for all measurable orders $\mathcal{R}$, and for all coalitions S, we have

$$\varphi(v, \theta_n\mathcal{R})(S) \to (\varphi v)(S) \quad \text{as } n \to \infty.$$

The mixing value, if it exists, is clearly unique. The set of all members of ORD that have a mixing value is denoted MIX. The following is an immediate corollary of Theorem 2. It also follows from footnote 4 in [2, § 2, p. 18].

PROPOSITION 1. *Let $v \in \mathrm{MIX} \cap \mathrm{INP}$ and let φv be its mixing value. If A is v-null then $|\varphi v|(A) = 0$.*

In [2, Prop. C.1] we find a characterization of set functions in MIX which are absolutely continuous (see § 5 of [2] for the definition of absolute continuity). We next give a similar characterization of weakly continuous set functions in MIX. Namely:

THEOREM 5. *Let $v \in \mathrm{ORD} \cap \mathrm{WC}$. A necessary and sufficient condition that the set function φv be the mixing value of v is that for all μ in NA^1 with $v \overset{w}{\leq} \mu$, condition (7.1) holds.*

In order to prove Theorem 5 we need the following lemma:

LEMMA 5. *Let $v \in \mathrm{ORD}$, $\mu \in \mathrm{NA}^1$, θ_n, $n = 1, 2, \cdots$, be a μ-mixing sequence, $\mathcal{R}$ be a measurable order, and let $\tau \in \mathrm{NA}^1$. If $v \overset{w}{\leq} \mu$, then there exist a measure $\xi \in \mathrm{NA}^1$ and a ξ-mixing sequence $\{\psi_1, \psi_2, \cdots\}$, such that $\mu + \tau$ is absolutely continuous with respect to ξ and*

$$\varphi(v, \psi_k^{-1}\mathcal{R}) = \varphi(v, \theta_k\mathcal{R}).$$

Proof. The proof follows exactly as the proof of Lemma C.18 in [2], with the only exception that at the end of the proof one should refer to Lemma 4 rather than to Lemma C.14 of [2].

Proof of Theorem 5. The proof follows from Lemma 5 exactly as Proposition C.1 in [2] follows from Lemma C.18.

Remark. By Proposition 4.2 in [1] it follows that for absolutely continuous set functions, weak continuity and absolute continuity with respect to elements in NA^1 coincide. This implies that Theorem 5 generalizes Proposition C.1 of [2].

REFERENCES

[1] R. J. AUMANN AND U. G. ROTHBLUM, *Orderable set functions and continuity. III: Orderability and absolute continuity*, this Journal, 15 (1977), pp. 156–162.

[2] R. J. AUMANN AND L. S. SHAPLEY, *Values of Non-Atomic Games*, Princeton University Press, Princeton, 1974.

[3] N. DUNFORD AND J. R. SCHWARZ, *Linear Operators. Part I*, Interscience, New York, 1958.

[4] P. R. HALMOS AND L. J. SAVAGE, *Applications of the Radon–Nykodyn theorem to the theory of sufficient statistics*, Ann. Math. Statist., 20 (1949), pp. 225–234.

[5] K. R. PARTHASARTHY, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

[6] U. G. ROTHBLUM, *On orderable set functions and continuity. I*, Israel J. of Math., 16 (1973), pp. 375–397.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

ORDERABLE SET FUNCTIONS AND CONTINUITY. III: ORDERABILITY AND ABSOLUTE CONTINUITY*

ROBERT J. AUMANN† AND URIEL G. ROTHBLUM‡

Abstract. The concepts of orderability and absolute continuity of set functions were introduced by Aumann and Shapley (1974). They showed that every absolutely continuous set function is orderable. The main result of this paper is to show that the converse is false.

1. Introduction. This paper is one of a series of studies (cf. [1], [5], [6]) in which orderability and various continuity notions for set functions are investigated and related to each other. Throughout we assume familiarity with the concepts summarized in § 2 of [6]. Our main result (§ 5) concerns the *absolute continuity* of set functions (see [1, § 5] or § 2 of this paper). In [1, Prop. 12.8] it was shown that every absolutely continuous set function is orderable; here (§ 5) we construct an example to show that the converse is false. The example is a function of two nonatomic measures, and is in a sense "simplest possible": In § 4 we show that for functions of a single nonatomic measure, orderability and absolute continuity are equivalent.

2. Notations and definitions. We refer the reader to § 2 of [6] for a summary of some notations and definitions from [1] and [5] that will be used in this paper. Familiarity with the above section will be assumed throughout our discussions.

For x in the Euclidean space E^n, $\|x\|$ will always mean the summing norm, i.e., $\|x\| = \sum_{i=1}^n |x_i|$. If $x, y \in E^n$, write $x \leqq y$ if $x_i \leqq y_i$ for all i. If μ is a vector measure $(\mu_1, \cdots, \mu_n)$, then $\sum\mu$ will denote $\sum_{i=1}^n \mu_i$.

We next summarize some definitions and conventions from [1] which were not used in [6] and will be needed in this paper. The norm on BV is the *variation norm*, defined by

$$\|v\| = \inf\{u(I) + w(I) \mid u - w = v, \text{ where } u \text{ and } w \text{ are monotonic}\}.$$

A *chain* is a nondecreasing sequence of sets of the form $\varnothing = S_0 \subset S_1 \subset \cdots \subset S_n = I$. A *link* of this chain is a pair of successive elements. A *subchain* is a set of links. A chain will be identified with the subchain consisting of all links. If v is a set function and Λ is a subchain of a chain, then the *variation of v over* Λ is defined by $\|V\|_\Lambda = \sum|v(S_i) - v(S_{i-1})|$, where the sum ranges over $\{i \mid \{S_{i-1}, S_i\} \in \Lambda\}$. For a fixed Λ, $\|\cdot\|_\Lambda$ is a pseudonorm on BV, i.e., it enjoys all the properties of a norm except $\|v\|_\Lambda = 0 \Rightarrow v = 0$. It is known (see [1, Prop. 4.1]) that for every $v \in \mathrm{BV}$, $\|v\| = \sup \|v\|_\Lambda$, where the supremum is taken over all subchains Λ. It is also known that

* Received by the editors August 14, 1975. This work was supported by the National Science Foundation under Grant GS-3269 at the Institute for Mathematical Studies in the Social Sciences, Stanford University, Stanford, California.

† Institute of Mathematics, Hebrew University, Jerusalem, Israel, and Institute for Mathematical Studies in the Social Sciences, Stanford University, Stanford, California 94035.

‡ School of Organization and Management, Yale University, New Haven, Connecticut 06520.

the linear subspaces M, NA, WC and ORD are closed subspaces of BV [5, Prop. 4.2 and 4.3].

A set function v is said to be *absolutely continuous with respect to* a set function w (written $v \ll w$) [1, p. 35] if for every $\varepsilon > 0$ there is a $\delta > 0$ such that for every chain Ω and every subchain Λ of Ω, $\|w\|_\Lambda \leqq \delta$ implies $\|v\|_\Lambda \leqq \varepsilon$. Note that this relation is transitive, and that if v and w are measures, it coincides with the usual notion of absolute continuity. A set function is *absolutely continuous* if there is a measure $\mu \in \mathrm{NA}^+$ such that $v \ll \mu$. The set of all absolutely continuous set functions forms a closed linear subspace of BV [1, Prop. 5.2], denoted AC. Finally, pNA denotes the closed subspace of BV spanned by all powers of nonatomic measures.

3. Weak continuity and absolute continuity. A real-valued function on a subset of E^n is said to be *monotonically absolutely continuous* if for every $\varepsilon > 0$ there is a $\delta > 0$ such that if $x_1 \leqq y_1 \leqq x_2 \leqq \cdots \leqq x_n \leqq y_n$, then

$$\sum_{i=1}^{n} \|y_i - x_i\| \leqq \delta \Rightarrow \sum_{i=1}^{n} |f(y_i) - f(x_i)| \leqq \varepsilon.$$

If the domain of f is one-dimensional, then monotonic absolute continuity coincides with the usual absolute continuity.

PROPOSITION 1. *Let μ be an n-dimensional σ-additive measure whose components are in NA^+ and are mutually singular. Let f be a real-valued function on the range of μ with $f(0) = 0$. Let $v = f \circ \mu$. Then $v \ll \sum \mu \Leftrightarrow f$ is monotonically absolutely continuous.*

Proof. The direction $\Leftarrow$ is obvious. To prove the direction $\Rightarrow$, recall Lyapunov's theorem [4], according to which the range of a nonatomic σ-additive vector measure is convex and compact. From this and the mutual singularity it follows that if $x_1 \leqq y_1 \leqq x_2 \leqq \cdots \leqq x_n \leqq y_n$, then there exist $S_1, T_1, \cdots, S_n, T_n$ in $\mathscr{C}$ such that $\mu(S_i) = x_i$, $\mu(T_i) = y_i$, and $S_1 \subseteq T_1 \subseteq \cdots \subseteq S_n \subseteq T_n$, completing the proof of Proposition 1.

PROPOSITION 2. *Let $v \in \mathrm{BV}$ and $\mu, \xi \in M^+$. If $v \ll \xi$, then $v \underset{w}{\leqslant} \mu$ if and only if $v \ll \mu$.*

Proof. Sufficiency of the condition is obvious. To see the necessity, let $\xi = \xi^{ac} + \xi^{\perp}$ be the Lebesgue decomposition of ξ with respect to μ, i.e., $\xi^{\perp}$ and ξ^{ac} are nonnegative measures such that $\xi^{ac} \leqq \mu$ and $\xi^{\perp} \perp \mu$ [3, Thm. C, p. 134]. Let $A \in \mathscr{C}$ be such that $\xi^{\perp}(A) = 0$ and $\mu(I \setminus A) = 0$.

We shall show that $v \ll \xi^{ac}$, and since $\xi^{ac} \ll \mu$ it will follow that $v \ll \mu$. Let $\delta > 0$ correspond to a given ε in accordance with the absolute continuity $v \ll \xi$; i.e.,

$$\text{for any subchain } \Lambda, \qquad \|\xi\|_\Lambda \leqq \delta \Rightarrow \|v\|_\Lambda \leqq \varepsilon. \tag{3.1}$$

We shall prove that $v \ll \xi^{ac}$ by showing that

$$\text{for any subchain } \Lambda, \qquad \|\xi^{ac}\|_\Lambda \leqq \delta \Rightarrow \|v\|_\Lambda \leqq \varepsilon. \tag{3.2}$$

If we intersect each set in each link of Λ with A then we get a subchain Λ^* such that $\|\xi\|_{\Lambda^*} = \|\xi^{ac}\|_{\Lambda^*} \leqq \delta$, and therefore by (3.1), $\|v\|_{\Lambda^*} \leqq \varepsilon$. But because $v \underset{w}{\leqslant} \mu$ and $\mu(I \setminus A) = 0$, it follows that $\|v\|_\Lambda = \|v\|_{\Lambda^*} \leqq \varepsilon$. This proves (3.2).

COROLLARY 1. *Let* $\mu = (\mu_1, \mu_2, \cdots, \mu_n)$ *be an* n*-dimensional vector of measures in* NA^+. *Let* f *be a real-valued function on the range of* μ, *such that* $v = f \circ \mu \in \mathrm{BV}$. *Then* $v \in \mathrm{AC}$ *if and only if* $v \ll \sum \mu$.

Proof. Sufficiency of the condition is obvious. To verify the necessity note that $f \circ \mu \underset{w}{\leq} \sum \mu$ and use Proposition 2.

COROLLARY 2.[1] *Let* $v = f \circ \mu$, *where* $\mu \in \mathrm{NA}^+$; *then* $v \in \mathrm{pNA}$ *if and only if* $v \in \mathrm{AC}$.

Proof. The fact that $\mathrm{pNA} \subseteq \mathrm{AC}$ has been proved in [1, Cor. 5.3]. Now let $f \circ \mu \in \mathrm{AC}$. Then by Corollary 1, $f \circ \mu \ll \mu$, and hence by Proposition 1 and Theorem C in [1], $f \circ \mu \in \mathrm{pNA}$.

COROLLARY 3. *The inclusions* $\mathrm{BV} \supseteq \mathrm{WC} \supseteq \mathrm{AC}$ *are strict.*

Proof. The unanimity game v defined by

$$v(S) = \begin{cases} 1, & S = I, \\ 0, & \text{otherwise}, \end{cases}$$

shows that $\mathrm{BV} \neq \mathrm{WC}$. Next, let λ be Lebesgue measure, and let g be the Cantor function, which is not absolutely continuous; then $g \circ \lambda \in \mathrm{WC}$, and by Propositions 1 and 2, $g \circ \lambda \notin \mathrm{AC}$.

4. Ordered absolute continuity. Let $\mathcal{R}$ be a measurable order. A chain $\varnothing = S_0 \subseteq S_1 \subseteq \cdots \subseteq S_m = I$ is called an $\mathcal{R}$*-chain* if all the S_i are $\mathcal{R}$-initial segments. Note that an $\mathcal{R}$-chain is defined by a finite sequence of elements in I, $\infty \overset{\mathcal{R}}{=} s_m \overset{\mathcal{R}}{=} \cdots \overset{\mathcal{R}}{=} s_1 \overset{\mathcal{R}}{=} s_0 = -\infty$, such that $I(s_i, \mathcal{R}) = S_i$.

If v and w are in BV, then v is said to be *ordered absolutely continuous with respect to* w (written $v \underset{o}{\ll} w$), if for every measurable order $\mathcal{R}$ and $\varepsilon > 0$ there exists a $\delta > 0$ such that for every $\mathcal{R}$-chain Ω and every subchain Λ of Ω, $\|w\|_\Lambda \leqq \delta$ implies $\|v\|_\Lambda \leqq \varepsilon$. Note that the relation is transitive.

PROPOSITION 3. *Let* $v \in \mathrm{BV}$, $\mu \in M^+$.[2] *Then* v *is ordered absolutely continuous with respect to* μ *if and only if* $v \in \mathrm{ORD}$ *and* $v \underset{w}{\leq} \mu$.

Proof. First assume that v is ordered absolutely continuous. It is easily verified that this implies $v \underset{w}{\leq} \mu$. Using the argument of the proof of Proposition 12.8 of [1] we obtain that[3] $v \in \mathrm{ORD}$. This completes the proof of one direction.

To prove the second direction, let us assume $v \underset{w}{\leq} \mu$ and $v \in \mathrm{ORD}$. By [5, Thm. 3.2], we know that $v \underset{w}{\leq} \mu$ implies that $\varphi^{\mathcal{R}} v \underset{w}{\leq} \mu$ for all measurable orders $\mathcal{R}$. Recall that weak continuity and absolute continuity between members of M coincide [2, § III. 4.3, p. 131]; hence $\varphi^{\mathcal{R}} v \ll \mu$ for all measurable orders $\mathcal{R}$. But then it follows that $v \underset{o}{\ll} \mu$.

A set is said to be *ordered absolutely continuous* if there is a measure $\mu \in \mathrm{NA}^+$ such that v is ordered absolutely continuous with respect to μ. The set of all ordered absolutely continuous functions in BV is denoted OAC.

---

[1] Cf. [1, Thm. C].

[2] One may extend this theorem and require only $\mu \in M$, and not $\mu \in M^+$. This would slightly complicate the proof.

[3] In Proposition 12.8 of [1] one assumes $v \ll \mu$ and obtains in addition to $v \in \mathrm{ORD}$, also that $\varphi^{\mathcal{R}} v \ll \mu$ uniformly in $\mathcal{R}$. Here we assume only $v \underset{o}{\ll} \mu$, and can also obtain $\varphi^{\mathcal{R}} v \ll \mu$, but not uniformly.

COROLLARY 4. ORD $\cap$ WC = OAC

COROLLARY 5. OAC *is a closed linear subspace of* BV.

Remark. One may conjecture that if $v \in \mathrm{ORD}$ and every point in I is v-null then there exists a measure $\mu \in \mathrm{NA}^+$ such that $v \lesssim_w \mu$. If this is true then clearly it should yield that OAC equals the set of all set functions in ORD for which every point is null.

PROPOSITION 4. *Let* $v = f \circ \mu$, *where* $\mu \in \mathrm{NA}^+$; *then* $v \ll \mu$ *if and only if* $v \lessapprox_o \mu$.

Proof. If $v \ll \mu$, then trivially $v \lessapprox_o \mu$. Assume now that $v \lessapprox_o \mu$. By Proposition 3, $v \in \mathrm{ORD}$ and $v \lesssim_w \mu$. Let $\mathscr{R}$ be an arbitrary fixed measurable order, then by [5, Thm. 3.2], $\varphi^{\mathscr{R}} v \lesssim_w \mu$. Since weak continuity and absolute continuity between totally finite measures coincide, it follows that $\varphi^{\mathscr{R}} v \ll \mu$. For a given ε, let δ be given in accordance with the absolute continuity $\varphi^{\mathscr{R}} v \ll \mu$; i.e., for every subchain Λ,

$$\|\mu\|_\Lambda \leqq \delta \Rightarrow \|\varphi^{\mathscr{R}} v\|_\Lambda \leqq \varepsilon. \tag{4.1}$$

We shall show that $v \ll \mu$ by showing that for every subchain Λ,

$$\|\mu\|_\Lambda \leqq \delta \Rightarrow \|v\|_\Lambda \leqq \varepsilon. \tag{4.2}$$

Let Λ be a subchain satisfying $\|\mu\|_\Lambda \leqq \delta$ whose links are $\{S_j, T_j | 1 \leqq j \leqq m\}$, where $\varnothing \subseteq S_1 \subseteq T_1 \subseteq S_2 \subseteq \cdots \subseteq S_m \subseteq T_m \subseteq I$. Let

$$\bar{S}_j = \cap\{I(s, \mathscr{R}) | s \in I, \mu(I(s, \mathscr{R})) > \mu(S_j)\},$$

$$\bar{T}_j = \cap\{I(s, \mathscr{R}) | s \in I, \mu(I(s, \mathscr{R})) > \mu(T_j)\}.$$

By [1, Lem. 12.15] it follows that for $1 \leqq j \leqq m$, $\bar{S}_j$, $\bar{T}_j$ are measurable and that

$$\mu(\bar{S}_j) = \mu(S_j) \quad \text{and} \quad \mu(\bar{T}_j) = \mu(T_j). \tag{4.3}$$

Note also that $\bar{S}_j$ and $\bar{T}_j$ are $\mathscr{R}$-initial sets; hence, by [6, Lem. 2], it follows that for $1 \leqq j \leqq m$,

$$(\varphi^{\mathscr{R}} v)(\bar{T}_j) = v(\bar{T}_j) \quad \text{and} \quad (\varphi^{\mathscr{R}} v)(\bar{S}_j) = v(\bar{S}_j). \tag{4.4}$$

Let $\bar{\Omega}$ be the chain $\varnothing \subseteq \bar{S}_1 \subseteq \bar{T}_1 \subseteq \bar{S}_2 \subseteq \cdots \subseteq \bar{S}_m \subseteq \bar{T}_m \subseteq I$ and let $\bar{\Lambda}$ be a subchain of $\bar{\Omega}$ whose links are $\{\bar{S}_j, \bar{T}_j\}$, $1 \leqq j \leqq m$. Note that (4.3) implies that $\|\mu\|_{\bar{\Lambda}} = \|\mu\|_\Lambda \leqq \delta$. Hence, by (4.1), $\|\varphi^{\mathscr{R}} v\|_{\bar{\Lambda}} \leqq \varepsilon$, and therefore (4.4) and (4.3) imply that

$$\begin{aligned} \varepsilon \geqq \|\varphi^{\mathscr{R}} v\|_{\bar{\Lambda}} = \|v\|_{\bar{\Lambda}} &= \sum_{j=1}^{m} |f(\mu(\bar{T}_j)) - f(\mu(\bar{S}_j))| \\ &= \sum_{j=1}^{m} |f(\mu(T_j)) - f(\mu(S_j))| = \|v\|_\Lambda. \end{aligned}$$

We have established (4.2), thus completing the proof of Proposition 4.

COROLLARY 6. *Let* $v = f \circ \mu$ *where* $\mu \in \mathrm{NA}^+$; *then*

$$v \in \mathrm{AC} \Leftrightarrow v \ll \mu \Leftrightarrow v \in \mathrm{OAC} \Leftrightarrow v \lessapprox_o \mu \Leftrightarrow v \in \mathrm{ORD} \Leftrightarrow v \in \mathrm{pNA}.$$

Proof. The above follows from Proposition 2, Corollary 2, Proposition 3 and Proposition 4.

Remark. It clearly follows from Corollary 6 that if we wish to construct an example of the form $v = f \circ \mu$ that is in ORD$\setminus$AC, then μ has to be at least two-dimensional.

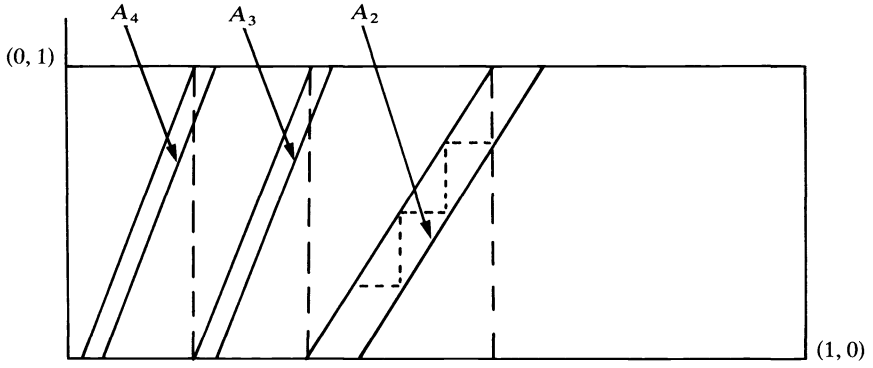
5. ORD includes AC strictly. It was proved in [1, Prop. 12.8] that ORD $\supseteq$ AC. We are now going to construct an example of a set function in ORD that is not in AC.[4] The example that we are going to describe appears, in a different context, at the beginning of § 9 in [1]. For each $k \geqq 2$ let $A_k \subset [0, 1]^2$ be the parallelogram whose vertices are: $(2^{-k}, 0)$, $(2^{-k}+4^{-k}, 0)$, $(2^{-k+1}+4^{-k}, 1)$ and $(2^{-k+1}, 1)$ (see Fig. 1). Define a nondecreasing continuous function f on the square such that for $x \in A_k$,

$$f(x) = f(x_1, x_2) = 2^k x_1 + 2^{-k+1} - 1;$$

for x between A_k and A_{k-1},

$$f(x) = 2^{-k+1} + x_2 + \frac{(x_1 - 2 \cdot 4^{-k})}{(1 - 2^{-k} + x_2)};$$

for x to the right of A_2 let f be defined by the same formula that defines f on A_2, i.e., $f(x) = 4x_1 - 1/2$; and finally for $x_1 = 0$ let $f(x) = x_2$. Let μ be any 2-dimensional vector measure on $(I, \mathscr{C})$, whose range if $[0, 1]^2$. We shall show that $v = f \circ \mu \in$ ORD$\setminus$AC.



To show that $v \notin$ AC, let

$$(2^{-k}, 0) = x_1^k \leqq x_2^k \leqq \cdots \leqq x_n^k = (2^{-k+1}, 1)$$

be a "staircase" sequence of points in A_n, i.e., each point differs from the preceding one in one coordinate only (see Fig. 1). On the vertical segments of this sequence, f does not change; all the change is concentrated on the horizontal segments. But the total length of the horizontal segments goes to 0, whereas the total change in f is 1. Therefore f is not monotonically absolutely continuous,

[4] One can easily see that by "smoothing" our example one can get a set function in MIX [1, § 13] that is not in AC.

therefore $f \circ \mu$ is not absolutely continuous with respect to $\Sigma\mu$ (Proposition 1), and therefore $f \circ \mu \notin \mathrm{AC}$ (Corollary 1).

Let us now prove that $v \in \mathrm{ORD}$. Set $\mu = \mu_1 + \mu_2$. We shall show that v is ordered absolutely continuous with respect to μ, and then use Proposition 3. Let $\mathcal{R}$ be a fixed measurable order. For a given $\varepsilon > 0$ we may choose a $1 > \delta_1 > 0$ such that

$$\|x - y\| \leqq \delta_1 \Rightarrow |f(x) - f(y)| \leqq \varepsilon/2. \tag{5.1}$$

This is possible because of the uniform continuity of f in $[0, 1]^2$.

Let J_1 denote the intersection of all $\mathcal{R}$-initial segments of μ_1-measure > 0. By [1, Lem. 12.15] it follows that J_1 is measurable and $\mu_1(J_1) = 0$. Let J denote the intersection of all $\mathcal{R}$-initial segments of μ-measure $> \mu(J_1) + \delta_1$. By the same lemma[5] we mentioned before, it follows that J is measurable and $\mu(J) = \mu(J_1) + \delta_1$, therefore $J \supseteq J_1$. Finally, observe that $\|\mu(J) - \mu(J_1)\| = \delta_1$; hence by (5.1) it follows that $|v(J) - v(J_1)| \leqq \varepsilon/2$.

Now let p be an integer $\geqq 2$ such that $2^{p-1} \geqq 1/\mu_1(J)$. Note that p depends only on $\mathcal{R}$ and ε. One can easily verify that f fulfills a Lipschitz condition on $\{x \in [0, 1]^2 | x_1 \geqq \mu_1(J)\}$ with constant 2^p, i.e., $\|f(y) - f(x)\| \leqq 2^p \|x - y\|$; this implies that if S, $T \in \mathscr{C}$ and $J \subseteq S \subseteq T$, then $\|v(T) - v(S)\| \leqq 2^p \{\mu(T) - \mu(S)\}$. Define $\delta = \min\{\delta_1, (2^p + 1)^{-1}\varepsilon/2\}$ and note that δ depends only on $\mathcal{R}$ and ε.

Let Λ be a subchain of an $\mathcal{R}$-chain Ω, with links $\{S_i, T_i\}$ $(1 \leqq i \leqq n)$, where $\varnothing \subseteq S_1 \subseteq T_1 \subseteq S_2 \subseteq \cdots \subseteq S_n \subseteq T_n \subseteq I$. By definition of $\mathcal{R}$-chain, S_i and T_i are $\mathcal{R}$-initial segments $(1 \leqq i \leqq n)$. We shall show that $\|\mu\|_\Lambda \leqq \delta$ implies $\|v\|_\Lambda \leqq \varepsilon$, which implies that v is ordered absolutely continuous with respect to μ, and hence by Proposition 4 that $v \in \mathrm{ORD}$.

Let $\|\mu\|_\Lambda \leqq \delta$, i.e., $\|\mu\|_\Lambda = \sum_{i=1}^n \{\mu(T_i) - \mu(S_i)\} \leqq \delta$. Without loss of generality we may assume that if $T_i \supseteq J$, then $S_i \supseteq J$; otherwise split $\{S_i, T_i\}$ into two links $\{S_i, J\}$ and $\{J, T_i\}$. Similarly we may assume that if $S_i \subseteq J_1$, then $T_i \subseteq J_i$. Note that since μ and v are monotonic, $\|\mu\|_\Lambda$ and $\|v\|_\Lambda$ remain unchanged. Let

$$I_1 = \{1 \leqq i \leqq n | T_i \subseteq J_1\},$$

$$I_2 = \{1 \leqq i \leqq n | J \subseteq S_i\},$$

$$I_3 = \{1 \leqq i \leqq n | J_1 \subseteq S_i \subseteq T_i \subseteq J\}.$$

I_1, I_2 and I_3 are disjoint, and by our previous assumption $I_1 \cup I_2 \cup I_3 = \{1, 2, \cdots, n\}$. Now

$$\|v\|_\Lambda = \sum_{i=1}^{n} |v(T_i) - v(S_i)| = \sum_{l=1}^{3} \sum_{i \in I_l} \{v(T_i) - v(S_i)\}$$

$$\leqq \sum_{i \in I_1} \{\mu_2(T_i) - \mu_2(S_i)\} + \sum_{i \in I_2} \{2^p(\mu(T_i) - \mu(S_i)\} + v(J) - v(J_1)$$

$$\leqq \delta + 2^p \cdot \delta + \frac{\varepsilon}{2} \leqq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

[5] The lemma must be modified to apply to measures μ in NA^+ for which $u(I) \neq 1$. Note that $\mu(J_1) + \delta_1 < \mu(I)$.

This completes the proof that $v \in \mathrm{ORD} \setminus \mathrm{AC}$. Hence we have shown

(5.2) ORD includes AC strictly.

REFERENCES

[1] R. J. AUMANN AND L. S. SHAPLEY, *Values of Non-Atomic Games*, Princeton University Press, Princeton, NJ, 1974.

[2] N. DUNFORD AND J. R. SCHWARZ, *Linear Operators, Part I*, Interscience, New York, 1958.

[3] P. R HALMOS, *Measure Theory*, Van Nostrand, Princeton, 1950.

[4] A. LYAPUNOV, *Sur les fonctions-vecteurs complèment additives*, Bull. Acad. Sci. U.S.S.R. Ser. Math., 4 (1940), pp. 465–478.

[5] U. G. ROTHBLUM, *On orderable set functions and continuity. I*, Israel J. Math., 16 (1973), pp. 375–397.

[6] ———, *Orderable set functions and continuity. II: Set functions with infinitely many null points*, this Journal, 15 (1977), pp. 144–155.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

ON THE STABILITY OF NONAUTONOMOUS DIFFERENTIAL EQUATIONS $\dot{x} = [A + B(t)]x$, WITH SKEW SYMMETRIC MATRIX $B(t)$*

A. P. MORGAN† AND K. S. NARENDRA‡

Abstract. In this paper we characterize (in Theorem 1) the uniform asymptotic stability of equations of the form

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \left[\begin{array}{c|c} A(t) & -B(t) \\ \hline B(t) & 0 \end{array}\right] \begin{bmatrix} x \\ y \end{bmatrix}$$

(where $A(t)+A(t)^T$ is negative definite) in terms of the "richness" of $B(t)$. The equation is uniformly asymptotically stable if and only if $B(t)$ is sufficiently rich. We actually obtain stability results for a much broader class of systems (Theorems 2 and 3) whose behavior is similar to the one above. Such systems have come up recently in some adaptive control problems.

1. Introduction. The purpose of this paper is to characterize the uniform asymptotic stability of certain nonautonomous linear systems of the form

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \left[\begin{array}{c|c} A & -B^T \\ \hline C & 0 \end{array}\right] \begin{bmatrix} x \\ y \end{bmatrix}$$

where $A = A(t)$ is a time varying stable $n \times n$ matrix and $B = B(t)$, $C = C(t)$ are time varying $m \times n$ matrices. Such equations arise in connection with questions of adaptive identification and control as described in Narendra and Kudva [5].

Theorem 1 below is illustrative of the type of result we have obtained. It is a corollary to the more general Theorems 2 and 3. We state and discuss these results in § 2 giving examples and some indication of proofs, including the presentation of a key lemma.

A result concerning (nonuniform) asymptotic stability has also been obtained, and this is stated in § 3. In § 4 we discuss in more detail the control applications of this work, which are summarized as Theorems 4 and 5. Section 5 contains the longer proofs.

Previous work on the stability properties of this type of system has been done by Yuan and Wonham [7]. They found sufficient conditions for asymptotic stability in the case that the system can be put in the form

$$\dot{e} = Ee + \Phi x + \Psi u,$$
$$\dot{\Phi} = -\Gamma e x^T,$$
$$\dot{\Psi} = -\Gamma e u^T.$$

(See § 4, Theorem 4 for more details.) Anderson in [1] considered some almost periodic cases, obtaining sufficient conditions for uniform asymptotic stability.

* Received by the editors October 3, 1975, and in revised form April 17, 1976. The research reported in this document was sponsored in part by support extended to Yale University by the U. S. Office of Naval Research under Contract N00014-67-A-0097-0020.

† Department of Mathematics, University of Miami, Coral Gables, Florida. Now at Medical College of Georgia, Augusta, Georgia 30902.

‡ Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520.

2. Statement of main theorems. The following Theorem 1 gives a complete characterization of uniform asymptotic stability when $A+A^T$ is negative definite and $C=B$. It is a corollary to Theorems 2 and 3, which we will state after a discussion of Theorem 1.

First, however, we establish some notation and state several definitions. The $n \times n$ time varying matrix $A=A(t)$ is called "stable" if the system $\dot{x}=A(t)x$ is uniformly asymptotically stable. The length of $x \in R^n$ is denoted "$|x|$". If A is an $n \times n$ matrix, "$|A|$" denotes the uniform norm of A derived from $|x|$.

"$P(t)$ is positive definite" means that there exist positive constants α and β such that $\alpha x^T x \leqq x^T P(t) x \leqq \beta x^T x$ for all $x \in R^n$ and all t. "$Q(t)$ is negative definite" means $-Q(t)$ is positive definite. If A is constant, then $\dot{x}=Ax$ is stable if and only if A is negative definite.

The equilibrium $x=0$ of the differential equation $\dot{x}=f(x,t)$ is uniformly asymptotically stable (u.a.s.) if it is uniformly stable and for some $\varepsilon_1>0$ and all $\varepsilon_2>0$ there is a $T=T(\varepsilon_1,\varepsilon_2)>0$ such that if $x(t)$ is a solution and $|x(t_0)| \leqq \varepsilon_1$, then $|x(t)| \leqq \varepsilon_2$ for all $t \geqq t_0+T$. If T depends on t_0, then $x \equiv 0$ is (nonuniformly) asymptotically stable (a.s.).

THEOREM 1. *Let $A=A(t)$ be an $n \times n$ matrix of bounded piecewise continuous functions such that $A+A^T$ is negative definite. Let $B(t)$ be an $n \times m$ matrix of bounded piecewise continuous functions. Then the system*

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \left[\begin{array}{c|c} A & -B^T \\ \hline B & 0 \end{array}\right] \begin{bmatrix} x \\ y \end{bmatrix} \tag{1}$$

is u.a.s. if and only if there are positive numbers T_0, ε_0, and δ_0 such that given $t_1 \geqq 0$ and a unit vector $w \in R^n$, there is a $t_2 \in [t_1, t_1+T_0]$ such that

$$\left| \int_{t_2}^{t_2+\delta_0} B(\tau)^T w \, d\tau \right| \geqq \varepsilon_0.$$

COROLLARY 1. *If $B(t)$ is smooth, $|\dot{B}(t)|$ is uniformly bounded, and there are real numbers $a>0$ and b such that*

$$\int_{t_1}^{t_2} |B(\tau)^T w| \, d\tau \geqq a(t_2-t_1)+b$$

for all unit $w \in R^m$ and all $t_2 \geqq t_1 \geqq 0$, then (1) is u.a.s.

COROLLARY 2. *If (1) is u.a.s., then there are real numbers $a>0$ and b such that*

$$\int_{t_1}^{t_2} |B(\tau)^T w| \, d\tau \geqq a(t_2-t_1)+b$$

for all unit vectors $w \in R^m$ and all $t_2 \geqq t_1 \geqq 0$.

The proof of Corollary 2 follows at once from the theorem, because the integral condition of the theorem clearly implies the integral condition of the corollary. The proof of Corollary 1 follows from the theorem, because, under the additional hypotheses on $B(t)$, it is easy to show that the integral condition of the theorem is equivalent to the integral condition of the corollary. The example below shows that, without additional hypotheses on $B(t)$, the integral condition of Corollary 1 can hold without the condition of the theorem being true.

The condition given in Theorem 1 is a "richness" condition for $B(t)$. It says that for any unit direction w, $B(t)^T w$ is "periodically" large. The condition requires that there be a fixed length of time, T_0, such that $B(t)$ "points in all directions" as t takes on values in any interval of length T_0. Also it requires that $B(t)$ maintain sufficient length. However, it requires even more than this, since the condition of Corollary 2,

$$\int_{t_1}^{t_2} |B(\tau)^T w|\, d\tau \geqq a(t_2 - t_1) + b,$$

is not sufficient. It is therefore apparent that, for fixed w, the sign changes that the components of $B(t)^T w$ go through are also significant. (See the example below.)

It is immediate that the condition there be positive numbers T_0 and ε_0 such that

$$\left| \int_t^{t+T_0} B(\tau)^T w\, d\tau \right| \geqq \varepsilon_0$$

for all unit $w \in R^m$ and $t \geqq 0$ is sufficient but not necessary for (1) to be u.a.s. (In this case δ_0 can be chosen arbitrarily and does not depend on w.)

The following example illustrates some of the above comments. Let $a_k = \sum_{n=1}^{k} (1/n)$, and define a square wave function with increasing frequency $u(t)$: $[0, \infty] \to R^1$ by

$$u(t) = \begin{cases} 1 \text{ if } t \in \left[a_k,\ a_k + \dfrac{1}{2(n+1)} \right], \\ -1 \quad \text{if } t \in \left[a_k + \dfrac{1}{2(n+1)},\ a_{k+1} \right]. \end{cases}$$

See Fig. 1. Then it follows from the theorem that the two-dimensional system

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\alpha & -u(t) \\ u(t) & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

(where α is a positive number) is not u.a.s. Note, however, that

$$\int_{t_1}^{t_2} |u(\tau)|\, d\tau = (t_2 - t_1).$$

Thus the necessary condition of Corollary 2 is not sufficient. Also, compare this with the following comments on the category PS^*.

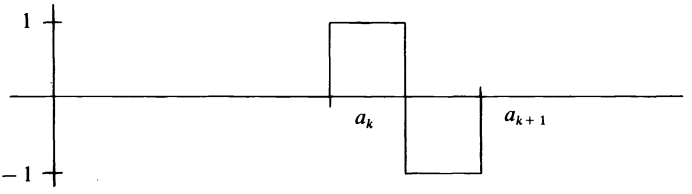


FIG. 1

We should point out that Corollary 1 can be generalized as follows. Instead of requiring that $B(t)$ be smooth and $|\dot{B}(t)|$ be bounded, we make the somewhat less restrictive assumption that the components of $B(t)$ be contained in the set PS^*, defined as a convenient category of input functions by Yuan and Wonham in [7].

PS^* is the set of all piecewise smooth functions $g:[0,\infty)\to R^1$ that are uniformly bounded, whose derivatives are uniformly bounded (where defined), and for which the intervals over which g is smooth do not shrink to 0.

For example, an input function g defined to be constant on intervals (a_n, a_{n+1}) where $a_{n+1}-a_n$ is bounded below as $n\to\infty$ is in PS^*.

Theorem 1 is an immediate corollary to Theorems 2 and 3 below, which are our main results. First, we recall the following.

By a theorem of Krasovskii, the uniform asymptotic stability of $\dot{x}=A(t)x$ implies that given any continuous symmetric positive definite $Q(t)$, there exists a continuous symmetric positive definite $P(t)$ such that $\dot{P}+PA+A^TP=-Q$. (See Narendra and Taylor [6, p. 62], or Halanay [2, p. 44].)

THEOREM 2. *Let $A(t)$ be a stable $n\times n$ matrix of bounded piecewise continuous functions. Let $P(t)$ be a symmetric positive definite matrix of bounded continuous functions such that $\dot{P}+PA+A^TP$ is negative definite. (Many such P exist, by our comment above.) Let $B(t)$ be an $n\times m$ matrix of bounded piecewise continuous functions.*

Assume that there exist positive numbers T_0, ε_0, and δ_0 such that given $t_1\geqq 0$ and a unit vector $w\in R^m$, there is a $t_2\in[t_1, t_1+T_0]$ such that

$$\left|\int_{t_2}^{t_2+\delta_0} B(\tau)^T w\, d\tau\right| \geqq \varepsilon_0.$$

Then the system

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix}=\left[\begin{array}{c|c} A & -B^T \\ \hline B\cdot P & 0\end{array}\right]\begin{bmatrix}x\\ y\end{bmatrix} \tag{2}$$

is u.a.s.

Corollary 1 and the comments about Yuan and Wonham's PS^* in the discussion following it hold exactly as written in this case.

THEOREM 3. *Let $A(t)$ be a stable $n\times n$ matrix of bounded piecewise continuous functions. Let $B(t)$ and $C(t)$ be $n\times m$ matrices of bounded piecewise continuous functions. Suppose that the system*

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix}=\left[\begin{array}{c|c} A & -B^T \\ \hline C & 0\end{array}\right]\begin{bmatrix}x\\ y\end{bmatrix} \tag{3}$$

is u.a.s. Then there are positive real numbers T_0, δ_0, and ε_0 such that given $t_1\geqq 0$ and a unit vector w, there is a $t_2\in[t_1, t_1+T_0]$ such that

$$\left|\int_{t_2}^{t_2+\delta_0} B(\tau)^T w\, d\tau\right| \geqq \varepsilon_0.$$

Corollary 2 holds exactly as written in this case also.

The comments made after the statement of Theorem 1 apply to Theorems 2 and 3. The condition which is necessary and sufficient for u.a.s. is a "richness" condition for $B(t)$, which however involves a subtlety concerning the sign changes of the components of $B(t)^T w$ as $t \to \infty$.

The following is a key observation, used in the proof of Theorem 2. Consider equation (2), and assume that the hypothesis of Theorem 2 holds. We shall use the notation $z(t) = [x(t), y(t)]^T$ from now on.

LEMMA 1. *Let ε_1 and δ be given positive numbers. Then there is a $T = T(\varepsilon_1, \delta)$ such that if $z(t)$ is a solution of* (2) *and $|z(t_1)| \leqq \varepsilon_1$, then there exists some $t_2 \in [t_1, t_1 + T]$ such that $|y(t_2)| \leqq \delta$.*

The lemma says that if $B(t)$ is sufficiently rich, then, for any solution $z(t) = [x(t), y(t)]^T$, $y(t)$ gets "periodically" small.

We shall now outline how the proof of Theorem 2 follows from the lemma. This is written out in detail in § 5. Define the Lyapunov function $V(z, t) = V([x, y]^T, t) = x^T P(t) x + y^T y$. Then $\dot{V}(z, t) \leqq -k|x|^2$ where k is some positive number. Thus if $|y|$ is small (as it must be periodically, by the lemma) and $|z|$ is not, then $|x|$ is not. In this case, $|\dot{V}|$ is large and $V(z, t)$ is decreasing. Since $\dot{V} \leqq 0$ we have uniform stability, and the observations of the previous two sentences show that $|z|$ is periodically getting smaller and smaller. Uniform asymptotic stability follows.

The proof of the lemma can be easily derived from the following two sublemmas.

SUBLEMMA 1. *Let $\varepsilon_1 > \varepsilon_2 > 0$. Then there is an $n = n(\varepsilon_1, \varepsilon_2)$ such that if $z(t) = [x(t), y(t)]^T$ is a solution of* (2) *with $|z(t_1)| \leqq \varepsilon_1$ and $S = \{t \in [t_1, \infty) \mid |x(t)| > \varepsilon_2\}$, then $\mu(S) \leqq n$ where μ denotes Lebesgue measure.*

This sublemma holds without any restriction on $B(t)$. It states that there is a uniform limit on the amount of time a solution starting inside the ε_1 ball can remain outside the ε_2 ball. It therefore implies that if $z(t)$ is any solution of (2), then $x(t) \to 0$. It also implies the following. Given $\varepsilon_1 > \varepsilon_2 > 0$, there is a $T > 0$ such that if $z(t)$ is a solution of (2) with $|z(t_1)| \leqq \varepsilon_1$, then there is a $t_2 \in [t_1, t_1 + T]$ such that $|x(t_2)| \leqq \varepsilon_2$.

It is not the case that $x(t) \to 0$ uniformly in initial times t_0 without any restriction on $B(t)$. In other words, it is not the case that given $\varepsilon_1 > \varepsilon_2 > 0$, there is a $T > 0$ such that if $(x(t), y(t))$ is a solution with $|x(t_0)| \leqq \varepsilon_1$, then $|x(t)| \leqq \varepsilon_2$ for all $t \geqq t_0 + T$. For example, let a_n be a sequence with $a_0 = 0$, $a_{n+1} > a_n$, and $a_{n+1} - a_n \to \infty$ as $n \to \infty$. Define

$$b(t) = \begin{cases} 1 & \text{if } t \in [a_n, 1 + a_n], \\ 0 & \text{otherwise.} \end{cases}$$

Then, it is easy to see that solutions to

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -1 & -b(t) \\ b(t) & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{4}$$

with initial position $x = 0$, $y = 1$, and initial times $t_0 = 1 + a_n$ have the property that it takes longer and longer for $x(t)$ to go to 0 as $n \to \infty$.

Assume that the hypothesis of Theorem 2 holds. Then we have

SUBLEMMA 2. *Let $\delta > 0$ and $\varepsilon_1 > 0$ be given. Then there exist positive numbers ε and T such that if $z(t)$ is a solution of (2) with $|z(t_1)| \leqq \varepsilon_1$ and if $|y(t)| \geqq \delta$ for $t \in [t_1, t_1 + T]$, then there is a $t_2 \in [t_1, t_1 + T]$ such that $|x(t_2)| \geqq \varepsilon$.*

Thus, if $B(t)$ is "rich" and $|y(t)|$ is large, then $|x(t)|$ must be periodically large. The lemma is established from the two sublemmas as follows. If $y(t)$ is not periodically small, then the sublemmas imply that $x(t)$ is periodically both large and small. But Sublemma 1 puts an upper bound on this type of behavior. The details of the proof of the lemma are in § 5.

3. Nonuniform asymptotic stability. The following proposition is a nonuniform version of Theorem 2. Its proof appears in § 5.

PROPOSITION. *Let $A(t)$ be a stable $n \times n$ matrix of bounded piecewise continuous functions. Let $P(t)$ be a symmetric positive definite matrix of bounded continuous functions such that $\dot{P} + AP + A^T P$ is negative definite. Let $B(t)$ be an $n \times m$ matrix of bounded piecewise continuous functions.*

Assume there exist positive numbers ε_0 and δ_0 such that given a unit vector $w \in R^m$, there is a sequence $t_n \to \infty$ such that

$$\left| \int_{t_n}^{t_n + \delta_0} B^T(\tau) w \, d\tau \right| \geqq \varepsilon_0 \quad \text{for all } n.$$

Then the system

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \left[\begin{array}{c|c} A & -B^T \\ \hline BP & 0 \end{array}\right] \begin{bmatrix} x \\ y \end{bmatrix}$$

is asymptotically stable.

Remark. We can easily see that if there is a nonzero $w \in R^m$ such that

$$\left| \int_0^{\infty} B^T(\tau) w \, d\tau \right| < \infty,$$

then the system is not asymptotically stable. It is not unreasonable to conjecture that the sufficient condition of the proposition is also necessary for asymptotic stability.

4. Applications to control theory. The type of equations discussed in earlier sections have come up recently in connection with control problems dealing with the adaptive observer. It also appears reasonable to assume that questions regarding the uniform asymptotic stability of similar nonautonomous equations will increasingly occur in adaptive control problems where parameters of the systems can be adjusted at the discretion of the designer, i.e., parts of the vector differential equation can be chosen. In this section we characterize the uniform asymptotic stability of two types of equations which arose in the context of identification. (See Narendra and Kudva [5], for details. Also compare Yuan and Wonham [7].)

THEOREM 4. *Consider the system*

$$\begin{aligned} \dot{e} &= Ee + \Phi x + \Psi u, \\ \dot{\Phi} &= -\Gamma e x^T, \\ \dot{\Psi} &= -\Gamma e u^T, \end{aligned} \tag{5}$$

where E is a stable $n \times n$ constant matrix, $e \in R^n$, Φ is $n \times n$, Ψ is $n \times m$, Γ is a symmetric positive definite matrix such that $\Gamma E + E^T \Gamma$ is stable, and $x : [t_0, \infty] \to R^n$, $u : [t_0, \infty) \to R^m$ are piecewise continuous, uniformly bounded vector valued functions.

Then (5) *is u.a.s. if and only if there are positive constants T_0, ε_0, δ_0 such that given $t_1 \geqq 0$ and a unit vector*

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in R^n \times R^m = R^{n+m},$$

there is a $t_2 \in [t_1, t_1 + T_0]$ such that

$$\left| \int_{t_2}^{t_2+\delta_0} [x(\tau)^T, u(\tau)^T] w \, d\tau \right| \geqq \varepsilon_0.$$

If $|\dot{x}(t)|$ and $|\dot{u}(t)|$ are defined and bounded, then the above condition can be replaced by:

there exist $a > 0$ and b such that

$$\int_{t_1}^{t_2} |[x(\tau)^T, u(\tau)^T] w| \, d\tau \geqq a(t_2 - t_1) + b$$

for all unit $w \in R^n \times R^m$ and all $t_2 \geqq t_1$. This completes the statement of Theorem 4.

In the context of identification, $\dot{x} = Ax + Bu$ where A is a constant stable matrix and b is a constant matrix. Thus $\dot{x}$ is always bounded.

This theorem follows at once from Theorems 2 and 3 and their corollaries. Note also the comments in § 2 which allow us to assume "$u(t) \in PS^*$" in place of "$|\dot{u}(t)|$ bounded".

The next theorem concerns a type of equation which also arises in identification schemes. (See Narendra and Kudva [5, p. 553]. Also see Anderson [1, p. 2.20].)

THEOREM 5. *Let A be a stable $n \times m$ constant matrix, and let P be a positive definite symmetric matrix such that $PA + A^T P$ is stable. Assume that there exist nonzero vectors d and h such that $Pd = h$. Let $v(t)$ be a piecewise continuous bounded vector valued function. Then the system*

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \left[\begin{array}{c|c} A & h \cdot v(t)^T \\ \hline -v(t) \cdot d^T & 0 \end{array}\right] \begin{bmatrix} x \\ y \end{bmatrix}$$

is u.a.s. if and only if there are positive constants T_0, ε_0, and δ_0 such that if $t_1 \geqq 0$ and w is a unit vector, then there is a $t_2 \in [t_1, t_1 + T_0]$ such that

$$\left| \int_{t_2}^{t_2+\delta_0} v(\tau)^T \cdot w \, d\tau \right| \geqq \varepsilon_0.$$

If $|\dot{v}(t)|$ is bounded or $v(t) \in PS^*$, then the above condition can be replaced by: there exist $a > 0$ and b such that

$$\int_{t_1}^{t_2} |v(\tau)^T \cdot w| \, d\tau \geqq a(t_2 - t_1) + b \quad \textit{for all unit } w.$$

This theorem is immediate from Theorems 2, 3, and corollaries. (Of course, we could also easily derive a version of Theorem 5 with A, h, d, and P time varying. However, these are constant in the application cited.)

Note that, by the Kalman–Yakubovich lemma, the conditions

(a) $PA + A^T P$ stable for some positive definite symmetric P, and

(b) $Pd = h$ for some d, h,

are equivalent to the condition that the transfer matrix $H(s) \equiv h^T(sI - A)^{-1}d$ be positive real. (Narendra and Taylor [6, p. 49].)

5. Proofs of theorems. We shall present proofs to Theorem 2, Lemma 1 and the sublemmas, Theorem 3, and the proposition. We also state and prove a Lemma 2, used in the proof of Theorem 3. The proofs of the corollaries and the comment in § 2 about PS^* are routine and therefore omitted.

It will be convenient to use the notation

$$z = \begin{bmatrix} x \\ y \end{bmatrix} = [x, y]^T \in R^{n+m}.$$

Also, for convenience, we assume $|A(t)| \leqq 1$ and $|B(t)| \leqq 1$ for all t.

Proof of Theorem 2. 1. By hypothesis we may choose positive constants α, β, a, b such that

$$\alpha x^T x \leqq x^T P(t) x \leqq \beta x^T x$$

and

$$a x^T x \leqq x^T Q(t) x \leqq b x^T x \quad \text{for all } x,$$

where $-Q(t) \equiv \dot{P}(t) + P(t)A(t) + A(t)^T P(t)$. Without loss of generality, assume $\beta \geqq 1$, $\alpha \leqq 1$, and $a \leqq 1$.

Define $V(z, t) = x^T P(t) x + y^T y$. Then

$$\dot{V}(z, t) = 2x^T(\dot{P}(t) + P(t)A(t) + A(t)^T P(t))x = -2x^T Q(t) x \leqq -2a x^T x.$$

If $z(t)$ is a solution of (2), then all of the above implies that $|z(t)|^2 \leqq (\beta/\alpha)|z(t_1)|^2$ for any $t \geqq t_1$.

2. We shall now show that given $\varepsilon_1 > \varepsilon_2 > 0$ there is a γ with $0 < \gamma < 1$ and an $M > 0$ such that if $z(t)$ is a solution of (2) with

$$\varepsilon_2 \leqq V(z(t), t) \leqq \varepsilon_1 \quad \text{for } t \in [t_1, t_1 + M],$$

then there is a $t_2 \in [t_1, t_1 + M]$ such that $V(z(t_2), t_2) \leqq \gamma \cdot V(z(t_1), t_1)$.

Since $V(z(t), t)$ is nonincreasing, this implies uniform asymptotic stability. The above fact follows routinely from the lemma and the relation $\dot{V}(z(t), t) \leqq -2a x(t)^T x(t)$. However, for completeness, we shall write out the details.

Choose positive numbers c_1 and c_2 so that $1 - c_1 > 0$, $\sqrt{(1-c_1)/\beta} - 2c_2\sqrt{\beta/\alpha} > 0$ and $0 < 2ac_2(\sqrt{(1-c_1)/\beta} - 2c_2\sqrt{\beta/\alpha})^2 < 1$. (Say $c_1 = 3/4$ and $c_2 = \alpha/8\beta$.) Use the lemma to obtain T when $\varepsilon = \varepsilon_1$ and $\delta = \sqrt{\varepsilon_2 \cdot c_1}$.

Define $\gamma = 1-[2ac_2(\sqrt{(1-c_1)}/\sqrt{\beta}-2c_2\sqrt{\beta}/\alpha)^2]$ and $M = T+c_2$. We shall show that for this γ and M our result holds.

We want $0<\gamma<1$. But this is clear from the choice of c_1 and c_2. Let $t_2' \in [t_1, t_1+T]$ be such that $|y(t_2')| \leqq \delta = \sqrt{\varepsilon_2 \cdot c_1}$. If $V(z(t_2'), t_2') \leqq \varepsilon_2$, we are done. Assume $V(z(t_2'), t_2') \geqq \varepsilon_2$. Then

$$V(z(t_2'), t_2') = x(t_2')^T P(t_2') x(t_2') + |y(t_2')|^2$$

implies

$$\beta |x(t_2')|^2 \geqq V(z(t_2'), t_2') - \delta \geqq V(z(t_2'), t_2')(1-c_1).$$

Now $\dot{x} = Ax - B^T y$ gives, for any $t \geqq t_2'$,

$$|x(t_2')| - |x(t)| \leqq |x(t) - x(t_2')| \leqq \int_{t_2'}^{t} |A(\tau)x(\tau) - B(\tau)y(\tau)|\, d\tau$$

$$\leqq (1+1)(\sqrt{\beta}/\sqrt{\alpha})|z(t_2')|(t-t_2')$$

since we have assumed $|A(\tau)| \leqq 1$ and $|B(\tau)| \leqq 1$ for all τ.

If we let $t_2 = t_2' + c_2$, then we see that

$$|x(t)| \geqq |x(t_2')| - 2(t_2 - t_2')(\sqrt{\beta}/\sqrt{\alpha})|z(t_2')|$$

$$\geqq (\sqrt{(1-c_1)}/\sqrt{\beta})\sqrt{V(z(t_2'), t_2')} - 2c_2(\sqrt{\beta}/\sqrt{\alpha})|z(t_2')|$$

$$\geqq (\sqrt{(1-c_1)}/\sqrt{\beta})\sqrt{V(z(t_2'), t_2')} - 2c_2(\sqrt{\beta}/\sqrt{\alpha})(\sqrt{V(z(t_2'), t_2')}/\sqrt{\alpha})$$

$$\geqq \sqrt{V(z(t_2'), t_2')}(\sqrt{(1-c_1)}/\sqrt{\beta} - 2c_2\sqrt{\beta}/\alpha)$$

for all $t \in [t_2', t_2]$. Then

$$V(z(t_2'), t_2') - V(z(t_2), t_2) = \int_{t_2'}^{t_2} -\dot{V}(z(\tau), \tau)\, d\tau$$

$$\geqq 2a \int_{t_2'}^{t_2} |x(\tau)|^2\, d\tau$$

$$\geqq 2a \cdot c_2 \cdot V(z(t_2'), t_2') \cdot (\sqrt{(1-c_1)}/\sqrt{\beta} - 2c_2\sqrt{\beta}/\alpha)^2.$$

Thus $V(z(t_2), t_2) \leqq V(z(t_2'), t_2') \cdot \gamma$, and we are done.

Proof of the lemma. Let $\delta > 0$. By the comments after the statement of Sublemma 1 and by Sublemma 2, the assumption for some solution $z(t)$ that $|y(t)| \geqq \delta$ implies that there is an $\varepsilon > 0$ such that $|x(t)|$ is repeatedly both less than $\varepsilon/2$ and greater than ε. Now this eventually leads to a contradiction with Sublemma 1, when we let $\varepsilon_1 = |z(t_1)|$ and $\varepsilon_2 = \varepsilon/2$. Since all these results are uniform, we conclude that $|y(t)| \leqq \delta$ repeatedly (uniformly). This yields the lemma.

Proof of Sublemma 1. This is immediate from the relation $\dot{V}(z, t) \leqq -2ax^T x$. We can choose $n(\varepsilon_1, \varepsilon_2) = \varepsilon_1^2/2a\varepsilon_2^2$.

Proof of Sublemma 2. 1. By hypothesis, we have T_0, ε_0, and δ_0 given, obeying the condition in the statement of Theorem 2. Let $z(t)$ be a solution with initial condition $|z(t_1)| \leqq \varepsilon_1$. Suppose that $|y(t) \geqq \delta$ for all $t \in [t_1, t_1 + T]$ where $T \equiv T_0 + \delta_0$.

2. Now $\dot{x} = Ax - B^T y$ implies, for any $t \geqq t_1$, that

$$x(t+\delta_0) = x(t) + \int_t^{t+\delta} A(\tau)x(\tau) - B(\tau)^T y(\tau)\, d\tau,$$

which gives

$$|x(t+\delta_0)| \geqq \left| \int_t^{t+\delta_0} B(\tau)^T y\, d\tau \right| - \left| x(t) + \int_t^{t+\delta_0} A(\tau)x(\tau)\, d\tau \right|.$$

We shall see below that we can make the second term arbitrarily small and the first term relatively large by appropriate choices of t and ε. This will prove the result.

3. We have $\dot{y}(\tau) = B(\tau)P(\tau)x(\tau)$. Thus, "when x is small, y is flat." More precisely, given T' and M' positive constants, there is a $\theta > 0$ such that if $z(\tau) = [x(\tau), y(\tau)]^T$ is a solution to (2) with $|x(\tau)| \leqq \theta$ for all $\tau \in [t_1, t_1 + T']$, then $|y(\tau) - y(t_1)| \leqq \varepsilon'$ for all $\tau \in [t_1, t_1 + T]$.

Let $\varepsilon' = \varepsilon_0\delta/(2\delta_0)$, $M' = \varepsilon_1$, and $T' = T = T_0 + \delta_0$, and fix θ for these choices.

4. Define $\varepsilon = \min\{\delta\varepsilon_0/8, \delta\varepsilon_0/8\delta_0, \theta\}$. We shall now show that the sublemma holds for this choice of T and ε. If $|x(t_2)| \geqq \varepsilon$ for some $t_2 \in [t_1, t_1 + T]$, we are done. Assume $|x(t)| \leqq \varepsilon$ for all $t \in [t_1, t_1 + T]$. Then $|x(t) + \int_t^{t+\delta_0} A(\tau)x(\tau)\, d\tau| \leqq \varepsilon + \varepsilon \cdot \delta_0 \leqq \varepsilon_0\delta/4$ for any $t \in [t_1, t_1 + T_0]$. (We have assumed $|A(\tau)| \leqq 1$ and $|B(\tau)| \leqq 1$ for all τ.)

By hypothesis there is a $t' \in [t_1, t_1 + T_0]$ such that $|\int_{t'}^{t'+\delta_0} B(\tau)^T w\, d\tau| \geqq \varepsilon_0$ where $w \equiv y(t_1)/|y(t_1)|$. But

$$\left| \int_{t'}^{t'+\delta_0} B(\tau)^T (w|y(t_1)| - y(\tau))\, d\tau \right| \leqq \int_{t'}^{t'+\delta_0} |y(t_1) - y(\tau)|\, d\tau$$

$$\leqq \delta_0 \cdot \frac{\varepsilon_0\delta}{2\delta_0} = \frac{\varepsilon_0\delta}{2},$$

because $|x(\tau)| \leqq \varepsilon \leqq \theta$ for $\tau \in [t_1, t_1 + T]$. (See part 3 above.) Therefore

$$|y(t_1)| \left| \int_{t'}^{t'+\delta_0} B(\tau)^T w\, ds \right| - \left| \int_{t'}^{t'+\delta_0} B(\tau)^T y(\tau)\, d\tau \right| \leqq \frac{\varepsilon_0\delta}{2},$$

implying

$$\left| \int_{t'}^{t'+\delta_0} B(\tau)^T y(\tau)\, dz \right| \geqq \varepsilon_0\delta - \frac{\varepsilon_0\delta}{2} = \frac{\varepsilon_0\delta}{2}.$$

Thus

$$|x(t'+\delta_0)| \geqq \frac{\varepsilon_0\delta}{2} - \frac{\varepsilon_0\delta}{4} = \frac{\varepsilon_0\delta}{4} > \varepsilon.$$

This completes the proof of Sublemma 2.

Proof of Theorem 3. 1. For convenience, we define

$$E=\left[\begin{array}{c|c} A & -B^T \\ \hline C & 0 \end{array}\right] \quad \text{and} \quad F=\left[\begin{array}{c|c} 0 & B^T \\ \hline 0 & 0 \end{array}\right].$$

Thus (3) is $\dot{z}=E(t)z$.

2. Since $\dot{z}=E(t)z$ is u.a.s., we may use Kraskovskii's theorem (mentioned before the statement of Theorem 2) to conclude that there is a continuously differentiable bounded positive definite symmetric matrix $P(t)$ such that

$$\dot{P}+PE+E^TP=-I.$$

Thus we have $|\dot{P}(t)|\leqq k_0$ for some constant k_0. By positive definiteness, we have constants α and β with

$$\alpha z^Tz\leqq z^TP(t)z\leqq\beta z^Tz$$

for all $z\in R^{n+m}$ and all $t\geqq 0$. We lose no generality assuming $|P(t)|\leqq 1$.

3. We now consider the perturbed system

$$\dot{z}=E(t)z+F(t)z. \tag{3'}$$

If w is a unit vector in R^m, the $z_0(t)=[0, w]^T$ is a (constant) solution to (3'). Letting $V(z, t)=z^TP(t)z$, we have

$$\dot{V}_{3'}=\dot{V}_3+z^T[PF+F^TP]z,$$

implying

$$z_0^T\dot{P}(t)z_0=-z_0^Tz_0+z_0^T[PF+F^TP]z_0,$$

implying

$$\int_{t_0}^{t} z_0^T\dot{P}(\tau)z_0\,d\tau+\int_{t_0}^{t} z_0^Tz_0d\tau=2\int_{t_0}^{t} z_0^T[P(\tau)F(\tau)]z_0\,d\tau.$$

This gives

$$2\int_{t_0}^{t} w^TP_0(\tau)B(\tau)^Tw\,d\tau=(t-t_0)+z_0^TP(t)z_0-z_0^TP(t_0)z_0,$$

where P_0 is a submatrix of P. Thus

$$\left|\int_{t_0}^{t} w^TP_0(\tau)B(\tau)^T w\,d\tau\right|\geqq\tfrac{1}{2}(t-t_0)-\beta \tag{6}$$

for all $t\geqq t_0\geqq 0$.

4. We now apply Lemma 2 to (6) above. (Lemma 2 is stated and proved following this proof.) We conclude that there are positive constants δ_1, ε_1, and T_1 such that if $t_1\geqq 0$, then there is $t_2\in[t_1, t_1+T_1]$ such that

$$\left|\int_{t_2}^{t_2+\delta} w^TP_0(\tau)B(\tau)^Tw\,d\tau\right|\geqq\varepsilon_1\delta$$

for all δ with $0<\delta\leqq\delta_1$.

5. Now $|\dot{P}(\tau)| \leqq k_0$ implies that

$$|P(\tau) - P(t_2)| \leqq k_0 |\tau - t_2|$$

for any $\tau \geqq t_2 \geqq 0$. Thus

$$\left| \int_{t_2}^{t_2+\delta} [w^T P_0(\tau) B^T(\tau) w - w^T P_0(t_2) B^T(\tau) w] \, d\tau \right|$$

$$\leqq \int_{t_2}^{t_2+\delta} |P_0(\tau) - P_0(t_2)| \, d\tau \leqq \frac{k_0}{2} \delta^2.$$

Therefore

$$\left| \int_{t_2}^{t_2+\delta} w^T P_0(\tau) B^T(\tau) w \, d\tau \right| - \frac{k_0}{2} \delta^2 \leqq \left| \int_{t_2}^{t_2+\delta} w^T P_0(t_2) B^T(\tau) w \, d\tau \right|$$

$$\leqq \left| w^T P_0(t_2) \int_{t_2}^{t_2+\delta} B^T(\tau) w \, d\tau \right|$$

$$\leqq |w^T P_0(t_2)| \left| \int_{t_2}^{t_2+\delta} B^T(\tau) w \, d\tau \right|$$

$$\leqq \left| \int_{t_2}^{t_2+\delta} B^T(\tau) w \, d\tau \right|.$$

Now, choosing $t_2 \in [t_1, t_1 + T_1]$ as in part 4 above, we have

$$\left| \int_{t_2}^{t_2+\delta} B^T(\tau) w \, d\tau \right| \geqq \varepsilon_1 \delta - \frac{k_0}{2} \delta^2 \quad \text{for all } \delta \text{ with } 0 < \delta < \delta_1.$$

Thus it is clear that there is a δ_0 with $\delta_1 \geqq \delta_0 > 0$ such that

$$\left| \int_{t_2}^{t_2+\delta_0} B(\tau)^T w \, d\tau \right| \geqq \varepsilon_1 \delta_0 - \frac{k_0}{2} \delta_0^2 > 0.$$

Clearly δ_0 does not depend on the choice of w. Letting $T_0 \equiv T_1$, $\delta_0 \equiv \delta_0$, and $\varepsilon_0 \equiv \varepsilon_1 \delta_0 - k_0 / 2\delta_0^2$ we are done.

LEMMA 2. *Let $f: [0, \infty) \to R^1$ be piecewise continuous and bounded. Assume there are constants $a > 0$ and $b > 0$ such that*

$$\left| \int_{t_0}^{t} f(\tau) \, d\tau \right| \geqq a(t - t_0) - b$$

for all $t \geqq t_0 \geqq 0$.

Then there are positive constants δ_1, ε_1, and T_1 such that if $t_1 \geqq 0$, then there is $t_2 \in [t_1, t_1 + T_1]$ such that

$$\left| \int_{t_2}^{t_2+\delta} f(\tau) \, d\tau \right| \geqq \varepsilon_1 \delta$$

for all δ with $0 \leqq \delta \leqq \delta_1$.

Proof. Define $\delta_1 = 2b/a$, $\varepsilon_1 = a/8$, and $T_1 = 2b/a$. We have

$$\left| \int_{t_1}^{t_1+T_1} f(\tau) \, d\tau \right| \geqq b$$

for any $t_1 \geqq 0$. Suppose that

$$\left| \int_{t_2}^{t_2+\delta} f(\tau)\, d\tau \right| \leqq \delta \cdot \varepsilon_1$$

for each $t_2 \in [t_1, t_1 + T_1]$. Then we may choose $t_1 < t_2 < \cdots < t_{n+1} = t_1 + T_1$ subdividing $[t_1, t_1 + T_1]$ with $t_k - t_{k-1} = \delta$ except that $0 < t_{n+1} - t_n \leqq \delta$. Note then that we have $T_1/\delta \leqq n+1 \leqq (T_1+\delta)/\delta$ and $\delta \leqq T_1$. Thus

$$b \leqq \left| \int_{t_1}^{t_1+T_1} f(\tau)\, d\tau \right| \leqq \sum_{m=1}^{n} \left| \int_{t_m}^{t_{m+1}} f(\tau)\, d\tau \right| \leqq \sum_{m=1}^{n} \delta \cdot \varepsilon_1$$

$$= n \cdot \delta \cdot \varepsilon_1 \leqq \frac{(T_1+\delta)}{\delta} \cdot \delta \cdot \varepsilon_1 \leqq \frac{b}{2}.$$

This being a contradiction, we are done.

Proof of the proposition. 1. As in the proof of Theorem 2, we have a Lyapunov function $V(x, y, t)$ with $\dot{V}(x, y, t) \leqq -a|x|^2$ for some $a > 0$. It follows that if $(x(t), y(t))$ is a solution, then

$$\int_0^{\infty} |x(\tau)|\, d\tau < \infty.$$

This immediately implies that if $(x(t), y(t))$ is a solution, then there is a constant vector $w \in R^m$ such that $(x(t), y(t)) \to (0, w)$ as $t \to \infty$. (Compare LaSalle [3, Thm. D].)

2. Now $\dot{x} = Ax - B^T y$ implies that

$$x(t+\delta_0) - x(t) = \int_t^{t+\delta_0} A(\tau)x(\tau)\, d\tau - \int_t^{t+\delta_0} B^T(\tau)y(\tau)\, d\tau$$

for any t. It follows that

$$\left| \int_t^{t+\delta_0} B^T(\tau)y(\tau)\, d\tau \right| \to 0 \quad \text{as } t \to \infty.$$

Since $y(\tau) \to w$ as $\tau \to \infty$, this yields

$$\left| \int_t^{t+\delta_0} B^T(\tau)w\, d\tau \right| \to 0 \quad \text{as } t \to \infty.$$

If $w \neq 0$, this would contradict our hypothesis. This completes the proof of the proposition.

Acknowledgment. We would like to thank Professor J. P. La Salle for several very useful suggestions and for his encouragment of this work.

REFERENCES

[1] B. ANDERSON, *Multivariable adaptive identification*, preprint, Dept. of Electrical Engineering, Univ. of Newcastle, New South Wales, 2368, Australia, 1974.

[2] A. HALANAY, *Differential Equations*, Academic Press, New York 1966.

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

[4] A. P. MORGAN AND K. S. NARENDRA, *On the uniform asymptotic stability of certain nonautonomous linear differential equations*, this Journal, 15 (1977), pp. 5–24.

[5] K. S. NARENDRA AND P. KUDVA, *Stable adaptive schemes for system identification and control, Parts I and II*, IEEE Trans. Systems, Man, and Cybernetics, SMC-4 (1974), pp. 542–560.

[6] K. S. NARENDRA AND J. TAYLOR, *Frequency Domain Criteria for Absolute Stability*, Academic Press, New York, 1973.

[7] J. S. C. YUAN AND W. M. WONHAM, *Asymptotic identification using stability criteria. Part I: Probing signal description*, Control Systems Rep. 7422, Univ. of Toronto, Toronto, Canada, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 1, January 1977

A "SIMPLEST POSSIBLE" PROPERTY OF THE GENERALIZED ROUTH–HURWITZ CONDITIONS*

B. D. O. ANDERSON† AND E. I. JURY‡

Abstract. To decide whether a prescribed complex polynomial has all its zeros with negative real parts, there are available many tests involving the checking of rational or polynomial inequalities in the coefficients. It is shown that the generalized Routh–Hurwitz conditions are from a certain point of view not replaceable by simpler conditions.

1. Introduction. The problem of deciding when a prescribed polynomial with real or complex coefficients is such that all its zeros have negative real parts has been studied in the early work of Hermite [1], if not earlier by Cauchy, who was interested in stating procedures for counting the number of roots of a polynomial in a half plane. Of course, much has been done since that time, and the majority of known results are collected in [2], [3] and [4].

Let x be a real vector whose entries are the coefficients of a real polynomial, or the real and imaginary parts of a complex polynomial. Most results are of the following form: a prescribed polynomial has all zeros with negative real parts (in brief, is Hurwitz) if and only if $p_j(x) > 0, j = 1, 2, \cdots, J$, where the $p_j(\cdot)$ are either polynomial or rational in the components of x. For example, the Hermite test [1], generalized Routh–Hurwitz test [2] and Schwarz test [5] associated with a complex polynomial are all of this type.

Two comments on these stability conditions are relevant. First, it is possible to conceive a minor extension of this type of condition, which we illustrate by example. In lieu of the quantities $p_j(x)$, consider the quantities $(x_1 - x_2)^2 p_1(x)$, $p_2(x), \cdots, p_J(x)$. These have the property that they are nonnegative for all Hurwitz polynomials, and positive for almost all; with a suitable topology in the space of vectors x, a polynomial has the property that almost all polynomials in a small neighborhood of it satisfy the strict inequalities, and conversely if for almost all polynomials in a small neighborhood of a prescribed polynomial the inequalities hold strictly, the prescribed polynomial must be Hurwitz.

Stability conditions allowing this restricted nonnegativity replacement of pure positivity will be called "restricted nonnegativity" conditions, in contrast to the "pure positivity" condition of the second paragraph of the section. Of course, a "pure positivity" condition is a special "restricted nonnegativity" condition.

The second comment on the type of conditions considered is that one can replace rational conditions by polynomial ones: if $p_1(x) = q_1(x)/r_1(x)$, with q_1, r_1 relatively prime polynomials in the components of x, then $p_1 > 0$ if and only if $q_1 r_1 > 0$. Any results applicable to the class of stability conditions involving only polynomials then, in fact, apply to stability conditions involving rational functions.

* Received by the editors February 26, 1976. This research was sponsored by the Joint Services Electronics Program Contract F44620-71-C-0087 to the University of California, Berkeley, and the Australian Research Grants Committee.

† Department of Electrical Engineering, University of Newcastle, New South Wales 2308, Australia.

‡ Department of Electrical Engineering and the Electronics Research Laboratory, University of California, Berkeley, California 94720.

In this paper, we examine the class of stability conditions involving polynomials, including conditions of the "restricted nonnegativity" type. Our main result is that the generalized Routh–Hurwitz conditions are the simplest set of conditions for a complex polynomial to be Hurwitz, in two respects: no other set has fewer inequalities, and no other set is such that the sum of the degrees of the inequalities is lower than the sum of the degrees of the Routh–Hurwitz inequalities.

In § 2 we review the statement of the generalized Routh–Hurwitz condition, and in § 3 we establish that one at least of the Routh–Hurwitz inequalities is, in a certain sense, contained in an arbitrary set of polynomials defining a stability condition. Section 4 is devoted to proving the main result, and § 5 contains concluding remarks.

2. Generalized Routh–Hurwitz conditions. Let $f(z)$ be an nth degree polynomial with complex coefficients and with

$$f(jz)=b_0z^n+b_iz^{n-1}+\cdots+b_n+j(a_0z^n+a_1z^{n-1}+\cdots+a_n). \tag{1}$$

The a_i, b_j are real.

Define the $2n\times 2n$ matrix

$$H=\begin{bmatrix} a_0 & a_1 & \cdot & \cdot & \cdot & a_{2n-1} \\ b_0 & b_1 & \cdot & \cdot & \cdot & b_{2n-1} \\ 0 & a_0 & a_1 & \cdot & \cdot & a_{2n-2} \\ 0 & b_0 & b_1 & \cdot & \cdot & b_{2n-2} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & & & \cdot \end{bmatrix}$$

where $a_m, b_m=0$ for $m>n$. Let $\Delta_1, \Delta_2, \cdots, \Delta_n$ denote the leading principal minors of H of dimension $2, 4, \cdots, 2n$. Then it is known [2], [3], [4] that $f(z)$ has all its zeros inside $\mathrm{Re}[z]<0$ if and only if $\Delta_1>0, \Delta_2>0, \cdots, \Delta_n>0$.

We remark that Δ_n is readily recognized as the resultant [6], [7] of the two polynomials $A(z)=a_0z^n+a_1z^{n-1}+\cdots+a_n$ and $B(z)=b_0z^n+b_1z^{n-1}+\cdots+b_n$. In the sequel, we shall use two important properties of the resultant. First, viewed as a multivariable polynomial in $a_0, a_1, \cdots, a_n, b_0, \cdots, b_n$, it is prime; see [6, p. 87]. Second, when the coefficients of $A(z)$, $B(z)$ take particular numerical values with a_0, b_0 not both zero, the value taken by Δ_n is zero if and only if $A(z)$ and $B(z)$ have a nontrivial common factor. If $\Delta_{n-1}\neq 0$, the greatest common divisor of the two polynomials is of degree 1; if $\Delta_{n-1}=0, \Delta_{n-2}\neq 0$, it is of degree 2, and so on [7, p. 150].

If the two polynomials have a greatest common divisor of degree 1, it is of the form $z+\bar{\omega}_0$ for $\bar{\omega}_0$ real. It follows then that $f(jz)$ is zero when $z=-\bar{\omega}_0$, i.e. $-j\bar{\omega}_0$ is a zero of $f(z)$. Conversely, if $-j\bar{\omega}_0$ is a zero of $f(z)$, $z+\bar{\omega}_0$ is a common factor of $A(z)$ and $B(z)$ and $\Delta_n=0$.

3. A preliminary result. Notational convention to be used in this and the next section is as follows. The symbol x will be shorthand for the $(2n+2)$-vector $(a_0, a_1, \cdots, a_n, b_0, \cdots, b_n)$. An overbar on a coefficient or vector of coefficients will denote a particular value of that coefficient or vector; the symbol f, sometimes with superscripts or subscripts, will denote a polynomial in z with indeterminate coefficients, and when an overbar is used, a polynomial in z with coefficients taking particular values. Unfortunately at times we have to be slightly flexible in the use of this latter convention.

Let $q_k(x)$ for $k = 1, 2, \cdots, K$ be a set of real multivariable polynomials with the following properties. If $q_k(\bar{x}) > 0$ for all k, then $\Delta_j(\bar{x}) > 0$ for $j = 1, 2, \cdots, n$; if $\Delta_j(\bar{x}) > 0$ for all j, then $q_k(\bar{x}) \geqq 0$ for all k, and for almost all x in a suitably small neighborhood of $\bar{x}$, $q_k(x) > 0$. Evidently the q_k provide a set of polynomials constituting a tool for checking the Hurwitz character of a prescribed polynomial.

The main result of this section is as follows.

PROPOSITION 1. *With $q_k(\cdot)$ as described above, $\Delta_n(x)$ divides $\prod_{k=1}^{K} q_k(x)$ in $R[x]$.*

The proof will proceed with the aid of several lemmas. The overall strategy is to show first that if $\bar{x}$ is such that $\Delta_i(\bar{x}) > 0$ for $i \leqq n-1$ and $\Delta_n(\bar{x}) = 0$, then $\prod q_k(\bar{x}) = 0$. Then we show that for any $\bar{x}$ such that $\Delta_n(\bar{x}) = 0$, we must have $\prod q_k(\bar{x}) = 0$. The proposition is then a consequence of the fact that $\Delta_n(x)$ is prime.

LEMMA 1. *Let $f(z) = \hat{f}(z)(z + j\omega_0)$ where ω_0 is a real indeterminate, and $\hat{f}(z)$ is an $(n-1)$-st degree polynomial with indeterminate (complex) coefficients. Let $\hat{x}$, $\hat{\Delta}$ refer to the coefficients and generalized Hurwitz determinants associated with $\hat{f}$. Then*

$$\Delta_i(x) = \hat{\Delta}_i(\hat{x}), \qquad i \leqq n-1,$$
$$\Delta_n(x) = 0.$$

Proof. That $\Delta_n(x) = 0$ was pointed out at the end of the previous section. To establish that $\Delta_i(x) = \hat{\Delta}_i(\hat{x})$ observe that with $\hat{f}(jz) = \hat{b}_0 z^{n-1} + \cdots + \hat{b}_{n-1} + j(\hat{a}_0 z^{n-1} + \cdots + \hat{a}_{n-1})$, one has

$$\begin{aligned} f(jz) \doteq & -[\hat{a}_0 z^n + (\hat{a}_1 + \hat{a}_0\omega_0) z^{n-1} + \cdots + (\hat{a}_{n-1} + \hat{a}_{n-2}\omega_0) z + \hat{a}_{n-1}\omega_0] \\ & + j[\hat{b}_0 z^n + (\hat{b}_1 + \hat{b}_0\omega_0) z^{n-1} + \cdots + (\hat{b}_{n-1} + \hat{b}_{n-2}\omega_0) z + \hat{b}_{n-1}\omega_0]. \end{aligned}$$

Then

$\Delta_i(x) =$

$$\begin{bmatrix} \hat{b}_0 & (\hat{b}_1 + \hat{b}_0\omega_0) & (\hat{b}_2 + \hat{b}_1\omega_0) & \cdots & (\hat{b}_{n-1} + \hat{b}_{n-2}\omega_0) & \hat{b}_{n-1}\omega_0 & 0 & \cdots \\ -\hat{a}_0 & -(\hat{a}_1 + \hat{a}_0\omega_0) & -(\hat{a}_2 + \hat{a}_1\omega_0) & \cdots & -(\hat{a}_{n-1} + \hat{a}_{n-2}\omega_0) & -\hat{a}_{n-1}\omega_0 & 0 & \cdots \\ 0 & \hat{b}_0 & (\hat{b}_1 + \hat{b}_0\omega_0) & \cdots & (\hat{b}_{n-2} + \hat{b}_{n-3}\omega_0) & \cdot & \cdot & \cdot \\ 0 & -\hat{a}_0 & -(\hat{a}_1 + \hat{a}_0\omega_0) & \cdots & -(\hat{a}_{n-2} + \hat{a}_{n-3}\omega_0) & \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots & & \vdots & & & \end{bmatrix}$$

The result follows by subtracting ω_0 times the first column from the second, ω_0 times the second column from the third, and so on, and then interchanging the first and second rows, third and fourth rows, and so on. □

The prime use of Lemma 1 is simply to establish Lemma 2, which brings us to the first major step in proving Proposition 1.

LEMMA 2. *Let $\bar{x}$ be such that $\Delta_n(\bar{x})=0$ and $\Delta_i(\bar{x})>0$ for $i\leqq n-1$. Then $\prod q_k(\bar{x})=0$.*

Proof. By the remarks at the end of § 2, $\Delta_n(\bar{x})=0$ and $\Delta_{n-1}(\bar{x})\neq 0$ imply that the polynomial $\bar{f}(z)$ with coefficients $\bar{x}$ can be written as $\bar{f}(z)=\hat{\bar{f}}(z)(z+j\bar{\omega}_0)$ for some real $\bar{\omega}_0$ and some $(n-1)$st degree polynomial $\hat{\bar{f}}(z)$. Since $\Delta_i(\bar{x})>0$ for $i\leqq n-1$, by Lemma 1 it follows that $\hat{\bar{f}}(z)$ is Hurwitz. Set $\bar{f}_m(z)=\hat{\bar{f}}(z)(z+m^{-1}+j\bar{\omega}_0)$ for $m=1, 2, \cdots$. Then $\bar{x}_m\to\bar{x}$ as $m\to\infty$ and so $q_k(\bar{x}_m)\to q_k(\bar{x})$ as $m\to\infty$. Because $\bar{f}_m$ is Hurwitz, $0\leqq q_k(\bar{x}_m)$, so that $0\leqq q_k(\bar{x})$. If $\prod q_k(\bar{x})\neq 0$, then $q_k(\bar{x})>0$, contradicting the non-Hurwitz nature of $\bar{f}$. Therefore $\prod q_k(\bar{x})=0$. □

Our next goal is to show that $\Delta_n(\bar{x})=0$ implies $\prod q_k(\bar{x})=0$, irrespective of the signs of $\Delta_i(\bar{x})$ for $i\leqq n-1$. To do this we shall make use of some algebraic geometry ideas; we also make crucial appeal to the primeness of $\Delta_n(x)$.

LEMMA 3. *Let $\mathcal{S}$ be the set of multivariable polynomials $s(\cdot)$ in x such that any $\bar{x}$ for which $\Delta_n(\bar{x})=0$ and $\Delta_i(\bar{x})>0$ for $i\leqq n-1$ causes $s(\bar{x})=0$. Then $\mathcal{S}$ is an ideal in $R[x]$.*

The proof is a trivial application of the definition of an ideal [6]. Note that $\prod q_k(\cdot)$ by Lemma 2 and $\Delta_n(\cdot)$ by definition lie in $\mathcal{S}$.

Recall also that any polynomial ideal has a finite basis [7, p. 142]. Let this basis be $g_1(\cdot), \cdots, g_r(\cdot)$. Associated with the ideal $\mathcal{S}$ is a variety S, which is the set of x for which $g_i(x)=0$, $i=1, \cdots, r$. Any variety may be decomposed as the union of a finite number of irreducible varieties[1] [8, p. 9]; thus $S=S_1\cup\cdots\cup S_r$. Each S_i is of a certain dimension d_i; the tangent space at any x_i on S_i has dimension $\leqq d_i$, and for almost all x_i has dimension d_i [8, pp. 84–88].

We may define a second variety V as simply the zero set of $\Delta_n(\cdot)$. Because Δ_n is prime, V is irreducible and of dimension $2n+1$, since x is a $(2n+2)$-vector [8, p. 25].

LEMMA 4. *With varieties S and V as defined above, $S=V$.*

Proof. Choose an $\bar{x}$ for which $\Delta_n(\bar{x})=0$ and $\Delta_i(\bar{x})>0$ for $i\leqq n-1$. It is clear that a neighborhood around $\bar{x}$ will intersect $\Delta_n(x)=0$ in a $(2n+1)$-dimensional submanifold; the continuity of the $\Delta_i(x)$ with x ensures that if the neighborhood is sufficiently small, $\Delta_i>0$ on the submanifold. By Lemma 3, any $s\in\mathcal{S}$ is zero on this submanifold, and so, by the definition of S, the points of the submanifold lie on S. Hence there exist points on S with a tangent space of dimension $2n+1$. Therefore, in the decomposition of S into irreducible varieties, at least one variety, say S_1, has dimension $2n+1$.

Now because $\Delta_n\in S$, we have $\Delta_n(x)=\sum a_i(x)g_i(x)$ for all x and accordingly, $x\in S$ implies $x\in V$, i.e., $S\subset V$. Hence $S_1\subset V$. Since S_1 and V have the same dimension and are both irreducible as noted above, it follows (see [8, p. 23]), that $S_1=V$. Hence $S=V$.

---

[1] The results in [8] which we appeal to, though generally stated for projective varieties, all extend to affine varieties by standard devices as outlined particularly well in [9].

The proof of the proposition is now almost immediate. As noted following Lemma 3, $\prod q_k$ lies in S, i.e., $\prod q_k$ vanishes on S. Since $S = V$, this means that $\prod q_k(\bar{x}) = 0$ whenever $\Delta_n(\bar{x}) = 0$. Because Δ_n is prime, Δ_n divides $\prod q_k$.

4. Minimality of the generalized Routh–Hurwitz conditions. Associated with an nth degree complex polynomial there are n generalized Routh–Hurwitz conditions of degree $2, 4, \cdots, 2n$ in the coefficients of the polynomial. The sum of these degrees is $n(n+1)$. Our aim in this section is to show that it is not possible to reduce these numbers of n and $n(n+1)$ by working with some alternative set of polynomial inequalities. In case $n = 1$, the claim is immediate. To establish the result for arbitrary n, we shall proceed by induction.

Before stating and proving the main results, we make some preliminary remarks and definitions. With notation as in the previous section, define polynomials $p_k(x)$ by $q_k(x) = [\Delta_n(x)]^{\alpha_k} p_k(x)$ where the integer α_k is maximal. By Proposition 1 and the primeness of $\Delta_n(x)$, at least one of the $q_k(x)$ is divisible by $\Delta_n(\cdot)$, and so at least one α_k is positive.

Now suppose that $f(z)$ is of the form $\hat{f}(z)(z + j\omega_0)$ where $\hat{f}(z)$ is of degree $n-1$ with indeterminate coefficients collected in a real $2n$-vector $\hat{x}$, and ω_0 is a real indeterminate. Then x is defined by $\hat{x}$ and ω_0, and $\Delta_n(\hat{x}, \omega_0) = 0$. However, $p_k(\hat{x}, \omega_0)$ cannot be the zero polynomial, for otherwise arguments along the lines of the last section would imply that $p_k(\cdot)$ is divisible by $\Delta_n(\cdot)$. Select $\bar{\omega}_0$ such that $p_k(\hat{x}, \bar{\omega}_0)$ is not identically zero, and define $\tilde{p}_k(\hat{x}) = p_k(\hat{x}, \bar{\omega}_0)$. The definitions of $\bar{\omega}_0$, p_k and $\tilde{p}_k$ will be used in the proofs of Theorem 1 and 2 below.

THEOREM 1. *Let $f(jz)$ be the n-th degree polynomial given in* (1), *and let $x = (a_0, a_1, \cdots, a_n, b_0, \cdots, b_n)$. Suppose that $q_k(\cdot)$, $k = 1, 2, \cdots, K$, are real polynomials such that $q_k(\bar{x}) > 0$ for all k implies $\bar{f}$ is Hurwitz, and such that $\bar{f}$ Hurwitz implies $q_k(\bar{x}) \geqq 0$ and $q_k(x) > 0$ for all k and almost all x in a sufficiently small neighborhood of $\bar{x}$. Then $\sum \delta[q_k] \geqq n(n+1)$. Here, $\delta[q_k]$ denotes the degree of $q_k(\cdot)$.*

To prove the result, we shall proceed via a sequence of intermediate lemmas, beginning with the following extension of Lemma 1.

LEMMA 5. *Let $f_{\pm m}(z) = \hat{f}(z)(z \pm m^{-1} + j\bar{\omega}_0)$ where $m = 1, 2, \cdots$, and $\hat{f}(z)$ is an $(n-1)$-st degree complex polynomial with indeterminate coefficients. Let $f(z) = \lim_{m\to\infty} f_{\pm m}(z)$, and let $x, x_{\pm m}, \hat{x}, \Delta_i$, etc., be obviously defined. Suppose that for some specialization $\bar{\hat{x}}$ of $\hat{x}$, $\hat{\Delta}_i(\bar{\hat{x}}) \neq 0$, $i \leqq n-1$. Then for suitably large m, $\Delta_n(\bar{x}_{+m})$ and $\Delta_n(\bar{x}_{-m})$ have opposite signs.*

Proof. Since $\hat{\Delta}_{n-1}(\bar{\hat{x}}) \neq 0$, $\bar{\hat{f}}(z)$ has no pure imaginary zeros. Therefore $\bar{f}_{\pm m}(z)$ has no pure imaginary zeros and the number of right half plane zeros is given by the variations in sign in the sequence $1, \Delta_1(\bar{x}_{\pm m}), \cdots, \Delta_n(\bar{x}_{\pm m})$; see [2, p. 249]. By Lemma 1 and continuity, for m sufficiently large and $i \leqq n-1$, $\Delta_i(\bar{x}_{\pm m})$ approximates and therefore has the same sign as $\hat{\Delta}_i(\bar{\hat{x}}) = \Delta_i(\bar{x})$. Accordingly, since $f_{-m}(z)$ has one more zero in $\mathrm{Re}[z] > 0$ than $f_{+m}(z)$, $\Delta_n(\bar{x}_{+m})$ and $\Delta_n(\bar{x}_{-m})$ must have different signs. □

We remark that, strictly, the above proof does not use the fact that $\Delta_i(\hat{x}) \neq 0$ for $i < n-1$, since procedures are available for modifying the variations in sign formula to cope with the vanishing of intermediate Hurwitz determinants [2].

LEMMA 6. *Let $f(z)=\hat{f}(z)(z+j\bar{\omega}_0)$ and let the polynomial $\tilde{p}_k(\hat{x})$ be defined from $q_k(x)$ as described earlier. Then if $\bar{\hat{f}}$ is Hurwitz, $\tilde{p}_k(\bar{\hat{x}})\geqq 0$ for $k=1,2,\cdots,K$ with strict inequality for almost all $\hat{x}$ sufficiently close to $\bar{\hat{x}}$. Conversely, if $\tilde{p}_k(\bar{\hat{x}})>0$, $\bar{\hat{f}}$ is Hurwitz.*

Proof. Suppose $\bar{\hat{f}}$ is Hurwitz. Then $\bar{f}_m(z)=\bar{\hat{f}}(z)(z+m^{-1}+j\bar{\omega}_0)$ is Hurwitz, implying $q_k(\bar{x}_m)>0$ for $k=1,2,\cdots,K$ and $\Delta_n(\bar{x}_m)>0$. Therefore $p_k(\bar{x}_m)>0$ for $k=1,2,\cdots,K$. Letting $m\to\infty$ yields $\tilde{p}_k(\bar{\hat{x}})=p_k(\bar{\hat{x}},\bar{\omega}_0)\geqq 0$ for $k=1,2,\cdots,K$. Since no $\tilde{p}_k(\cdot)$ is identically zero, we must have strict inequality for almost all $\hat{x}$ sufficiently close to $\bar{\hat{x}}$.

Conversely, suppose $\tilde{p}_k(\bar{\hat{x}})>0$ for $k=1,2,\cdots,K$. Then for sufficiently large m, $p_k(\bar{x}_m)>0$. If $\Delta_n(\bar{x}_m)>0$, then $q_k(\bar{x}_m)>0$ for $k=1,2,\cdots,K$, implying $\bar{f}_m$ and therefore $\hat{f}$ are Hurwitz. It remains therefore to rule out the possibilities that $\Delta_n(\bar{x}_m)<0$ or $\Delta_n(\bar{x}_m)=0$. Assuming the former, we see from Lemma 5 that $\Delta_n(\bar{x}_{-m})>0$ while also $p_k(\bar{x}_{-m})>0$ for m sufficiently large. Then $q_k(\bar{x}_{-m})>0$ for $k=1,2,\cdots,K$, which contradicts the non-Hurwitz character of $\bar{f}_{-m}(z)=\hat{\bar{f}}(z)(z-m^{-1}+j\bar{\omega}_0)$. If $\Delta_n(\bar{x}_m)=0$, then $\hat{\Delta}_{n-1}(\bar{\hat{x}})=0$. However, for almost all $\hat{x}$ in a neighborhood of $\bar{\hat{x}}$, we still have $\tilde{p}_k(\hat{x})>0$ while $\hat{\Delta}_{n-1}(\hat{x})\neq 0$ and thus $\Delta_n(x_m)\neq 0$. Then $\hat{f}$ is Hurwitz, and it follows easily that $\bar{\hat{f}}$ must be Hurwitz. □

Lemma 6 and the earlier definitions show how to pass from a set of stability conditions for nth degree complex polynomials to a set for $(n-1)$st degree complex polynomials. This is the key to establishing Theorem 1.

Proof of Theorem 1. By the induction hypothesis, $\sum_k \delta[\tilde{p}_k(\hat{x})]\geqq(n-1)n$. Now it is easily established that $\sum_k \delta[p_k(x)]=\sum \delta[\tilde{p}_k(\hat{x})]$ from the definitions, while also $\sum_k \delta[q_k(x)]=\sum_k \alpha_k\,\delta[\Delta_n(x)]+\sum_k \delta[p_k(x)]$. Since $\alpha_k\geqq 0$ with at least one α_k positive and $\delta[\Delta_n(x)]=2n$, this gives $\sum_k \delta[q_k(x)]\geqq 2n+(n-1)n=(n+1)n$. This completes the induction.

The second main result relates to the number of inequalities.

THEOREM 2. *With the same hypothesis as Theorem* 1, $K\geqq n$.

The proof will again proceed via a number of lemmas.

LEMMA 7. *With the integer α_k as defined earlier, at least one α_k is odd for some k.*

Proof. Let $\bar{f}_{-m}(z)=\bar{\hat{f}}(z)(z-m^{-1}+j\bar{\omega}_0)$, with $\bar{\hat{f}}(z)$ Hurwitz and such that $\tilde{p}_k(\bar{\hat{x}})>0$ for all k. Then since $\lim_{m\to\infty} p_k(\bar{x}_{-m})=\tilde{p}_k(\bar{\hat{x}})$, for sufficiently large m, $p_k(\bar{x}_{-m})>0$ for all k, while also $\Delta_n(\bar{x}_{-m})\neq 0$. If the α_k are all even, this implies that $q_k(\bar{x}_{-m})>0$ for all k, a contradiction of the fact that $\bar{f}_{-m}(z)$ is not Hurwitz.

LEMMA 8. *With the q_k reordered so that $\alpha_1,\cdots,\alpha_s$ are odd and $\alpha_{s+1},\cdots,\alpha_k$ are even and possibly zero,*

$$\{q_k(x)>0 \text{ for } k=1,2,\cdots,K\}$$
$$\leftrightarrow\{\Delta_n p_1>0, p_1p_2>0,\cdots,p_1p_s>0, p_{s+1}>0,\cdots,p_K>0\}$$

Proof. Both inequality sets are clearly equivalent to $\Delta_n p_1>0,\cdots,\Delta_n p_s>0, p_{s+1}>0,\cdots,p_K>0$. □

Now suppose that the $\tilde{p}_k(\hat{x})$ are as defined earlier. Also define $\tilde{q}_1(\hat{x})=\tilde{p}_1(\hat{x})\tilde{p}_2(\hat{x}),\cdots,\tilde{q}_{s-1}(\hat{x})=\tilde{p}_1(\hat{x})\tilde{p}_s(\hat{x}),\ \tilde{q}_s(\hat{x})=\tilde{p}_{s+1}(\hat{x}),\cdots,\tilde{q}_{K-1}(\hat{x})=\tilde{p}_K(\hat{x})$.

LEMMA 9. *Let $f(z)=\hat{f}(z)(z+j\bar{\omega}_0)$, with the $\tilde{q}_k(\hat{x})$ defined as above. Then if $\bar{\hat{f}}$ is Hurwitz, $\tilde{q}_k(\bar{\hat{x}})\geqq 0$ for $k=1,2,\cdots,K-1$ with strict inequality for almost all $\hat{x}$ sufficiently close to $\bar{\hat{x}}$. Conversely, if $\tilde{q}_k(\bar{\hat{x}})>0$, for $k=1,2,\cdots,K-1$, $\bar{\hat{f}}$ is Hurwitz.*

Proof. Suppose $\bar{\bar{f}}$ is Hurwitz. Then $\bar{f}_m(z) = \bar{\bar{f}}(z)(z + m^{-1} + j\bar{\omega}_0)$ is Hurwitz. Using Lemma 8, we see this implies that $p_1(\bar{x}_m)p_2(\bar{x}_m) > 0, \cdots, p_1(\bar{x}_m)p_s(\bar{x}_m) > 0$, $p_{s+1}(\bar{x}_m) > 0, \cdots, p_K(\bar{x}_m) > 0$. Letting $m \to \infty$ and using the definition of the $\tilde{q}_k(\hat{x})$ establishes that $\tilde{q}_k(\bar{\hat{x}}) \geqq 0$. Since no $\tilde{p}_k(\hat{x})$ is identically zero, no $\tilde{q}_k(\hat{x})$ can be; therefore strict inequality holds for almost all $\hat{x}$ sufficiently close to $\bar{\hat{x}}$.

Conversely, let $\tilde{q}_k(\bar{\hat{x}}) > 0$. For suitably large m, $p_1(\bar{x}_{\pm m})p_2(\bar{x}_{\pm m}) > 0, \cdots, p_1(\bar{x}_{\pm m})p_s(\bar{x}_{\pm m}) > 0$, $p_{s+1}(\bar{x}_{\pm m}) > 0, \cdots, p_K(\bar{x}_{\pm m}) > 0$. Assume temporarily that $\tilde{p}_1(\bar{\hat{x}})\hat{\Delta}_{n-1}(\hat{x})$ is not zero. [If $s > 1$, $\tilde{p}_1(\bar{\hat{x}})$ is guaranteed to be nonzero, since $0 \neq \tilde{q}_1(\bar{\hat{x}}) = \tilde{p}_1(\bar{\hat{x}})p_2(\bar{\hat{x}})$.] Then $p_1(\bar{x}_{\pm m})$ have the same sign and, by Lemma 5, $\Delta_n(\bar{x}_{\pm m})$ have opposite signs. Therefore $\Delta_n(\bar{x}_{\pm m})p_1(\bar{x}_{\pm m})$ have opposite signs. If $\Delta_n(\bar{x}_{-m})p_1(\bar{x}_{-m}) > 0$, this, in conjunction with the inequalities $p_1(\bar{x}_{\pm m})p_2(\bar{x}_{\pm m}) > 0$, etc., implies by Lemma 8 that $\bar{f}_{-m}$ is Hurwitz, which is impossible. Therefore $\Delta_n(\bar{x}_m)p_1(\bar{x}_m) > 0$ and using Lemma 8, $\bar{f}_{+m}$ is seen to be Hurwitz. Therefore $\hat{f}$ is Hurwitz.

It remains to consider the case where one or both of $\tilde{p}_1(\bar{\hat{x}})$ and $\hat{\Delta}_{n-1}(\bar{\hat{x}})$ are zero. For almost all $\hat{x}$ in a sufficiently small neighborhood of $\bar{\hat{x}}$, $\tilde{p}_1(\hat{x})\hat{\Delta}_{n-1}(\hat{x})$ must be nonzero while $\tilde{q}_k(\hat{x}) > 0$. Therefore $\hat{f}$ is Hurwitz by the argument of the preceding paragraph. Then $\bar{\hat{f}}$ must be Hurwitz. □

It is now easy to complete the proof of Theorem 2. Applying the induction hypothesis to the inequality set $\tilde{q}_k(\hat{x}) > 0$ associated with $\hat{f}$ yields $K - 1 \geqq n - 1$. Therefore $K \geqq n$, as required.

We remark that there seems no direct way of combining the proofs of Theorems 1 and 2. Both theorems are proved by deriving from an inequality set associated with an nth degree polynomial a second set associated with an $(n-1)$st degree polynomial. The second set differs between the two theorems, as that set appropriate for proving the degree property is inappropriate for proving the number-of-inequalities property, and vice versa.

5. Conclusions and remarks. We have shown that the generalized Routh–Hurwitz conditions are the simplest set of polynomial inequalities defining the Hurwitz property of a complex polynomial, in the sense that no other set can contain fewer inequalities nor have a "total" degree smaller than that of the generalized Routh–Hurwitz conditions.

The question of what are the simplest set of polynomial inequalities defining the Hurwitz property of a *real* polynomial has not been tackled. Though the set of all real polynomials is obtainable by specializing certain coefficients in (1) to be zero, it does not of course follow that by making corresponding specializations in the generalized Routh–Hurwitz conditions, one obtains a set of "simplest possible" inequalities for real Hurwitz polynomials. Indeed this is demonstrably not the case, because this procedure recovers the standard Hurwitz test, and the Liénard–Chipart test is certainly simpler in terms of degree [2]. Work by the authors has come close to establishing that the Liénard–Chipart criteria are the simplest set of conditions for real polynomials to be Hurwitz, as might be expected; a full proof however is still lacking.

The same sort of results as those obtained in this paper appear to follow for "unit-circle" stability. More precisely, the Schur–Cohn inequalities as set out in [4; see pp. 28, 29] would appear to be the simplest possible in the two senses dealt with above.

REFERENCES

[1] C. HERMITE, *Sur le nombre des racines d'une équation algébrique comprise entre des limites données*, J. Reine Angew. Math., 52 (1854), pp. 39–51; *Oeuvres*, 1, pp. 397–414.
[2] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1959.
[3] M. MARDEN, *Geometry of Polynomials*, American Mathematical Society, Providence, R.I., 1966.
[4] E. I. JURY, *Inners and Stability of Dynamic Systems*, John Wiley, New York, 1974.
[5] H. R. SCHWARZ, *Ein Verfahren zur Stabilitätsfrage bei Matrizen-Eigenwertprobleme*, Z. Angew. Math. Phys., 7 (1956), pp. 473–500.
[6] B. L. VAN DER WAERDEN, *Modern Algebra*, vol. I, Frederick Ungar, New York, 1953.
[7] W. V. D. HODGE AND D. PEDOE, *Methods of Algebraic Geometry*, vol. 1, Cambridge University Press, Cambridge, 1968.
[8] ———, *Methods of Algebraic Geometry*, vol. 2, Cambridge University Press, Cambridge, 1968.
[9] S. LEFSCHETZ, *Algebraic Geometry*, Oxford University Press, London, 1953.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 2, February 1977

A GENERAL THEORY OF OBSERVATION AND CONTROL*

SZYMON DOLECKI† AND DAVID L. RUSSELL‡

Abstract. This report explores the duality relationships between observation and control in an abstract Banach space setting. Preservation of observability and controllability in the presence of certain perturbations is studied in the context of differential equations in Banach space. Some attention is also given to the problem of optimal reconstruction of system states from observations.

1. Introduction. The duality relationship between observation and control has been recognized in the literature for at least two decades and, in one form or another, probably goes back much further. In the optimal control literature, which is much too extensive to reference here, the maximum principle characterizes the optimal control by its relationship to a functional of, or observation on, solutions of an adjoint system. This duality relationship is, of course, closely related to the duality results familiar to students of linear and nonlinear programming. It is not this theory which we treat in this paper but rather those duality relationships which relate controllability of a system to the observability of an adjoint system. This theory is just an adaptation of the theorems of functional analysis which relate the range of an operator to the null space of the adjoint operator. It is inevitable therefore that much of this material should already be available in the general mathematical literature if one has the patience to seek it out. It appears to the authors, however, that the precise outlines of this theory as it applies to various types of control processes are not generally appreciated. For example, the controllability theory of parabolic partial differential equations has been developed [11], [14], [15], [25], [26] simultaneously with the completely equivalent theory of state observation of such processes [7], [8], [27], [28], [30], [31], [38], [39]. It is very likely that this duplication has occurred in many other situations and will occur again, for example, in connection with control and observation theories for delay-differential systems [2], [3], which are only now being vigorously developed. For this reason and others, we present here a duality theory of wide applicability which summarizes known results of functional analysis in the context of control and observation theories and extends those results in directions indicated by applications of those theories.

The outlines of this duality theory have already been given by the first author in [9], a preliminary version of this paper.

In the present article X, Y and Z stand for Banach spaces of general type. If A is an operator defined on a subspace of one of these spaces its domain will be denoted by $\mathscr{D}(A)$. Its range, also in such a space, is denoted by $\mathscr{R}(A)$ and its kernel is signified by ker (A).

* Received by the editors August 25, 1975. Supported in part by the Office of Naval Research under Contract NR 041-404 and in part by the United States Army under Contracts DA-31-124-ARO-D-462 and DAHC04-75-C-0024.

† Instytut Matematyczny, Polskiej Akademii Nauk, Sniadeckich 8, Warsaw, Poland. Part of this author's work was carried out while visiting the University of Wisconsin, Madison.

‡ Department of Mathematics and Mathematics Research Center, University of Wisconsin—Madison, Madison, Wisconsin 53706.

We are primarily concerned with an abstract linear system

$$\begin{array}{ccc} X \supset \mathcal{D}(C) & \xrightarrow{C} & Y \\ & \searrow^{F} & \\ & Z, & \end{array} \tag{1.1}$$

where $C: X \to Y$ is linear with dense domain $\mathcal{D}(C)$ and $F: X \to Z$ is linear and bounded. We shall also give particular consideration to the system

$$X \supset \mathcal{D}(C) \xrightarrow{C} Y, \tag{1.2}$$

a specialization of (1.1) obtained by taking F to be the identity operator on X. Structures of this sort have been studied in [30] and [31].

Along with (1.1) and (1.2) we consider two types of dual systems. The first type is

$$\begin{array}{ccc} X^* & \xleftarrow{C^*} & \mathcal{D}(C^*) \subset Y^* \\ & \nwarrow^{F^*} & \\ & Z^*, & \end{array} \tag{1.3}$$

or, when F is the identity on X,

$$X^* \xleftarrow{C^*} \mathcal{D}(C^*) \subset Y^*, \tag{1.4}$$

obtained by letting X^*, Y^*, Z^* denote the conjugate spaces of X, Y and Z, respectively, and by taking C^*, F^* to be the adjoint operators for C and F. (These are, of course, defined in the Hilbert space sense if X, Y, Z are Hilbert spaces.) The second type is obtained by letting $\|\cdot\|_F$ be the pseudo-norm on X (norm on $X_F = X/\ker(F)$) defined by

$$\|x\|_F \equiv \|Fx\|_Z, \qquad x \in X.$$

(We remark that this is the weakest topology on X for which F remains continuous.) We then introduce X^F, the space of linear functionals on X which are continuous with respect to $\|\cdot\|_F$ and consider the F-dual system

$$X^F \xleftarrow{C^F} \mathcal{D}(C^F) \subset Y^*. \tag{1.5}$$

Here the superscript F denotes conjugate spaces and adjoint operators with respect to the pseudo-norm $\|\cdot\|_F$. Thus $\mathcal{D}(C^F)$ consists of precisely those $g \in Y^*$ for which the linear functional f defined on $\mathcal{D}(C)$ by

$$f(x) = g(Cx), \qquad x \in \mathcal{D}(C),$$

is continuous with respect to $\|\cdot\|_F$ (and hence defines, without ambiguity, a continuous linear functional f on X_F), and for such g we have

$$C^F g = f.$$

It is important that we should relate these abstract systems to observation and control systems in the usually understood sense. Consider, therefore, the following example.

Let $S(t)$ be a strongly continuous semigroup of bounded operators defined on a reflexive Banach space W for $t \geqq 0$ and let $S(t)^*$ denote the adjoint semigroup, also strongly continuous. For $w_0 \in W$ we define $w(\cdot) \in \mathscr{C}[0, T; W]$ by

$$w(t) = S(t)w_0 \tag{1.6}$$

and call it the *trajectory* of w_0. For such a trajectory $w(\cdot)$ we construct an observation

$$w(\cdot) = Hw(\cdot) = HS(\cdot)w_0 \in \mathscr{C}[0, T; V] \subseteq L^p[0, T; V], \qquad 1 < p < \infty. \tag{1.7}$$

Here V is reflexive and H: $W \to V$ is the *observing operator*. The operator H may be continuous or, as in the case of boundary observation on solutions of partial differential equations, may be defined only on a dense subspace $D \subseteq W$ which is invariant under the action of the semigroup $S(t)$. In either event the *observation operator* C: $D \to L^p[0, T; V]$ is defined by

$$Cw_0 = HS(\cdot)w_0, \qquad w_o \in D. \tag{1.8}$$

It may occur that C extends continuously to all of W even though $HS(\cdot)$ does not.

As already implied, it is convenient to regard $\mathscr{C}[0, T; V]$ as being embedded in the Banach spaces $L^p[0, T; V]$, $1 < p < \infty$, with norm

$$\|v(\cdot)\|_{L^p[0,T;V]} = \left[\int_0^T \|v(t)\|_V^p \, dt\right]^{1/p}. \tag{1.9}$$

The dual space for $L^p[0, T; V]$ is just $L^q[0, T; V^*]$ where $1/p + 1/q = 1$, and the adjoint of C is the operator

$$C^*: L^q[0, T; V^*] \to W^*$$

defined by

$$\begin{aligned}(C^*h)w_0 &= \left(\int_0^T S(t)^*H^*h(t)\, dt\right) w_0 \\ &= \left(\int_0^T S(T-t)^*H^*f(t)\, dt\right) w_0 \qquad (h(t) = f(T-t))\end{aligned}$$

with $\mathscr{D}(C^*)$ consisting of those $h \in L^2[0, T, V^*]$ such that $(\int_0^T S(t)^*H^*h(t)\, dt)w_0$ defines a continuous linear functional on W.

In this setting C^* may be regarded as a control operator yielding the final state $y_f(T)$ for the trajectory $y_f(\cdot)$ obtained by the action of the control function f in the system

$$y_f(t) = \int_0^t S(t-s)^*H^*f(s)\, ds, \qquad t \geqq 0. \tag{1.10}$$

In many such instances the operator $F\colon X \to Z$, appearing in (1.1), will be the "final state" operator

$$F\colon W \to W, \quad Fw_0 = S(T)w_0, \quad w_0 \in W. \tag{1.11}$$

The (final state) observation problem consists of the question as to whether or not $\|Fw_0\|_W = \|S(T)w_0\|_W$ is bounded relative to $\|Cw_0\|_Y$, where $Y = L^p[0, T; V]$. When this is true there arises the possibility of the existence of a *reconstruction operator* $G\colon Y \to W$ such that

$$F = GC, \tag{1.12}$$

i.e.,

$$S(T)w_0 = G(HS(\cdot)w_0), \tag{1.13}$$

which amounts to continuous constructibility of the final state from the observations $HS(\cdot)w_0 \in L^p[0, T; V]$. The adjoint (control) problem relates to the possible solution of the equation

$$F^*y_0 + C^*h = 0 \tag{1.14}$$

for each $y_0 \in W^*$ (i.e. $\mathscr{R}(F^*) \subset \mathscr{R}(C^*)$), or

$$S(T)^*y_0 + \int_0^T S(T-t)^*H^*f(t)\,dt = 0, \tag{1.15}$$

which amounts to finding $f \in L^q[0, T; V^*]$ steering the initial state

$$y_f(0) = y_0$$

to the final state

$$y_f(T) = 0$$

with $y_f(t)$, $0 \leqq t \leqq T$, being given by (1.10).

For

$$F = S(0) = I, \qquad Fw_0 = w_0$$

one has the problem of initial state observation: the question as to whether or not $\|w_0\|_W$ is bounded relative to $\|Cw_0\|_V$. This very special case is distinguished from the general case (F a bounded operator) in Definition 1.1 below. The case $F = I$ (initial state observation) relates to what we shall call observability while the more general case is called F-observability.

We shall frequently return to examples of this sort in the sequel. For the moment we trust that this example has served to motivate the following definitions applying to the general systems (1.1)–(1.5).

DEFINITION 1.1. *The system* (1.1) *is* (*continuously*) *F-observable if there is a constant* $K \geqq 0$ *such that*

$$\|x\|_F = \|Fx\|_Z \leqq K\|Cx\|_Y, \qquad x \in \mathscr{D}(C).$$

The system (1.2) (*or, alternatively,* (1.1) *with F equal to the identity*) *is observable if*

$$\|x\|_X \leqq K\|Cx\|_Y, \qquad x \in \mathscr{D}(C). \tag{1.16}$$

The system (1.1) *is F-constructible if there is a bounded operator* $G\colon Y \to Z$ *such that*

$$F = GC.$$

DEFINITION 1.2. *Let* Γ *be a subspace of* X^*. *The system* (1.3) *is* Γ *controllable if*

$$\Gamma \subseteq \mathcal{R}(C^*).$$

We shall say that (1.3) *is* F^**-controllable if*

$$\mathcal{R}(F^*) \subseteq \mathcal{R}(C^*)$$

and (1.4) (*or* (1.3) *with* F^* *equal to the identity*) *is* (*exactly*) *controllable if*

$$X^* \subseteq \mathcal{R}(C^*).$$

The system (1.5) *is* (*exactly*) *controllable if*

$$X^F \subseteq \mathcal{R}(C^F).$$

Without giving a formal definition, we would remark that the system (1.3) could be thought of as possessing an F^*-(*open loop*) *synthesis* when there is an operator $K\colon Z^* \to Y^*$ such that

$$F^* + C^*K = 0.$$

In the context of the system (1.10), the operator K applied to an element $y_0 \in W^*$ would produce a function $h = H^*y_0$ such that the control $f(t) = h(T-t)$ steers y_0 to zero at time T. In practice H^* would often reduce to a (time varying) linear feedback relationship. This concept of (open loop) synthesis will be seen to be the natural dual of F-constructibility.

Controllability as defined in Definition 1.2 has been referred to in the literature [11], [12], [17], [35], [36] as *exact* controllability. The term *approximate controllability*, a much weaker concept which means

$$\overline{\mathcal{R}(C^*)} = X^*,$$

has been studied in [11], [12], [23], [33], [34]. The term *distinguishability* is used for the relationship

$$\ker(C) = 0.$$

All of these notions correspond to varying degrees of solvability of equations or well-posedness of certain linear operations, reformulated in the language appropriate to observation and control [20], [29], [32], [33], [34].

We have used the spaces X^*, Y^*, Z^* in Definition 1.2 to facilitate exposition of the duality relationships between observation and control. It is clear, however, that this definition can also be used in connection with spaces $\tilde{X}$, $\tilde{Y}$, $\tilde{Z}$ which are not necessarily conjugate spaces of any other Banach spaces X, Y, Z. Definition 1.1 would then be of interest in connection with $\tilde{X}^*$, $\tilde{Y}^*$ and $\tilde{Z}^*$.

It may possibly be questioned whether there is any real need for the notion of F-observability, since (with F bounded) it is always implied if the process is observable (cf. 1.16)). The following theorem shows that observability must be

considered a relatively rare property (and, as a consequence, that the weaker notion of F-observability plays a useful role).

THEOREM 1.3. *We consider the observed system* (1.6), (1.7) *with* $S(t)$ *a strongly continuous semigroup of bounded operators for* $t>0$ *and* H *bounded. If we have initial state observability for some time* $T>0$, *i.e.*,

$$K\|HS(\cdot)w_0\|_{L^p[0,T;V]} \geqq \|w_0\|_W, \qquad w_0 \in W, \tag{1.17}$$

for some p, $1 \leqq p < \infty$, *and if for each* $t \leqq 0$ *the range of* $S(t)$ *is dense in* W, *then* $S(t)$ *can be extended to a strongly continuous group of bounded operators on* $-\infty < t < \infty$.

Remark. Clearly the condition $\overline{\mathcal{R}(S(t))} = w$ can be replaced by the condition

$$S^*(t)\tilde{w}_0 = 0 \Rightarrow \tilde{w}_0 = 0 \quad \text{in } W^*,$$

which is sometimes easier to verify.

Proof of Theorem 1.3. The strong continuity of $S(t)$ for $t \geqq 0$ together with the uniform boundedness principle shows that for any finite interval $[0, \tau]$ there is a positive number $M(\tau)$ such that

$$\|S(t)\| \leqq M(\tau), \qquad t \in [0, \tau]. \tag{1.18}$$

We clearly have $M(\tau)$ nondecreasing with increasing τ.

Suppose the observability result (1.17) is true and suppose also that for every real γ and every $\tau_0 > 0$ there is an element $w_0 = w_0(\gamma, \tau_0) \in W$, $w_0 \neq 0$, such that

$$\|S(\tau_0)w_0\|_W < e^{-\gamma\tau_0}\|w_0\|_W. \tag{1.19}$$

We shall see that these two assumptions are inconsistent.

If (1.19) is true, then for $\tau_0 \leqq t \leqq T$,

$$\begin{aligned} \|HS(t)w_0\|_V &= \|HS(t-\tau_0)S(\tau_0)w_0\|_V \\ &\leqq \|H\|\,\|S(t-\tau_0)\|\,e^{-\gamma\tau_0}\|w_0\|_W \leqq \|H\|M(T)\,e^{-\gamma\tau_0}\|w_0\|_W, \end{aligned} \tag{1.20}$$

where $M(T)$ is defined by (1.18), and for $0 \leqq t \leqq \tau_0$

$$\|HS(t)w_0\|_V \leqq \|H\|M(\tau_0)\|w_0\|_W. \tag{1.21}$$

Let $\gamma \to -\infty$ and $\tau_0 \to 0$ in a sequence of values γ_k, $\tau_{0,k}$ such that

$$\lim_{k\to\infty} \gamma_k \tau_{0,k} = -\infty.$$

Then it is easy to see from (1.20) and (1.21) that

$$\lim_{k\to\infty} \frac{\|HS(\cdot)w_{0,k}\|_{L^p[0,T;V]}}{\|w_{0,k}\|_W} = \lim_{k\to\infty} \frac{\|Cw_{0,k}\|_{L^p[0,T;V]}}{\|w_{0,k}\|_W} = 0$$

and initial state observability, (1.17), is contradicted.

Hence, if we have initial state observability, there must be a real number γ and a positive number τ_0 such that

$$\|S(\tau_0)w_0\|_W \geqq e^{-\gamma\tau_0}\|w_0\|_W, \qquad w_o \in W.$$

Every $t \geqq 0$ can be represented as

$$t = k\tau_0 + \tau$$

where k is a nonnegative integer and $0 \leqq \tau < \tau_0$. Therefore

$$\begin{aligned} \|S(t)w_0\|_W &= \|S(\tau_0)^k S(\tau)w_0\|_W \\ &\geqq e^{-k\gamma\tau_0}\|S(\tau)w_0\|_W \geqq \frac{e^{-k\gamma\tau_0}\|S(\tau_0)w_0\|_W}{M(\tau_0)} \\ &\geqq \frac{e^{-\gamma\tau_0}}{M(\tau_0)} e^{-\gamma t}\|w_0\|_W, \end{aligned} \tag{1.22}$$

wherein the second inequality follows from

$$\|S(\tau_0)w_0\|_W = \|S(\tau_0-\tau)S(\tau)w_0\|_W \leqq M(\tau_0)\|S(\tau)w_0\|_W.$$

Thus (1.22) shows

$$\|S(t)w_0\|_W \geqq \hat{M} e^{-\gamma t}\|w_0\|_W \tag{1.23}$$

for some $\hat{M} > 0$.

From (1.23) it follows that for every $t \geqq 0$, $S(t)$ is boundedly invertible on its range. Since we have assumed that this range is dense in W, the bounded operator $S(t)^{-1}$ has a unique extension to the whole space W and the range of $S(t)$ must be all of W.

Denoting $S^{-1}(t)$ by $S(-t)$ we have defined $S(t)$ for all t, $-\infty < t < \infty$. It is easy to verify that $S(t)$ is a group. The strong continuity for $t \leqq 0$ follows from the fact that $\mathcal{R}(S(-T)) = \mathcal{D}(S(T)) \equiv W$ and the identity

$$S(t+\delta)y - S(t)y = S(t+T+\delta)S(-T)y - S(t+T)S(-T)y$$

with T taken $> |t|$ and δ taken sufficiently small. Thus the proof of our theorem is complete.

Anticipating Theorem 2.1 of the next section we have

COROLLARY 1.4. *Let the hypotheses of Theorem* 1.2 *hold except that* (1.17) *is replaced by the controllability condition*

$$\left\{ y_f(T) \,|\, y_f(T) = \int_0^T S(T-t)^* H^* f(t)\, dt, \quad f \in L^q[0, T, V^*], \frac{1}{p} + \frac{1}{q} = 1 \right\} = W^*.$$

Then the result that $S(t)$ can be extended to a group remains true.

Remark. Here $S(t)^*$ is the strongly continuous adjoint semigroup. The result clearly implies that $S(t)^*$ can be extended to a group also.

The condition $\mathcal{R}(S(t)) = W$ in Theorem 1.3 cannot be dispensed with. The one dimensional hyperbolic system

$$\frac{\partial w}{\partial t} + \frac{\partial w}{\partial x} = 0, \qquad 0 < x < \infty, \qquad t \geqq 0, \tag{1.24}$$

$$w(0, t) = 0, \qquad t \geqq 0, \tag{1.25}$$

with $w(x, 0) \in L^2(0, \infty) (= W)$ is easily seen to be initial state observable in any time $T > 0$ via the observing operator

$$H = I,$$

leading to the observation operator

$$Cw(x, 0) = w(\cdot, \cdot) \in L^1(0, T; L^2[0, \infty)).$$

The semigroup corresponding to (1.24), (1.25) is just right translation with zero values "fed in" at the left-hand endpoint $x = 0$. The range is not dense in W for any $t > 0$ and the semigroup cannot be extended to a group.

2. Duality theorems. We begin this section by stating two theorems relating observability and controllability (cf. Definitions 1.1, 1.2 with $F = I$, the identity operator). These are actually familiar results in functional analysis, relating the annihilation properties of C to the range of the adjoint operator C^*. The relevant theorems may be found in, e.g., [18] or [20].

THEOREM 2.1. *The system* (1.1) *is observable if and only if the system* (1.3) *is* (*exactly*) *controllable.*

Proof. A restatement in terms of the operators C and C^* appearing in (1.1), (1.3) is: $C: \mathcal{D}(C) \subseteq X \to Y$ *has a bounded inverse on* $\mathcal{R}(C) \subseteq Y$ *if and only if* $\mathcal{R}(C^*) = X^*$. This is precisely Theorem II.3.11 in [18].

The second theorem is of the same character but it applies in the case where the observation process takes place in spaces X^*, Y^* which are dual spaces while the control process takes place in the original spaces X, Y. For example, a control system

$$\dot{x} = Ax + Bu$$

with $x \in L^1[0, 1]$ is not covered by Theorem 2.1 because $L^1[0, 1]$ is not the dual space of any Banach space. It would be covered by

THEOREM 2.2. *Let* C_1 *be closed with*

$$C_1: \mathcal{D}(C_1) \subseteq \tilde{X} \to \tilde{Y}, \qquad \mathcal{D}(C_1) \text{ dense in } \tilde{X}, \tag{2.1}$$

$$C_1^*: \mathcal{D}(C_1^*) \subseteq \tilde{Y}^* \to \tilde{X}^*. \tag{2.2}$$

Then, applying Definition 1.1 *to* (2.2) *and Definition* 1.2 *to* (2.1) (*see remark near the end of* § 1) *we see that the system* (2.1) *is* (*exactly*) *controllable if and only if* (2.2) *is observable.*

Proof. The restatement is: *If* C_1 *is closed then* $C_1^*: \mathcal{D}(C_1^*) \subseteq \tilde{Y}^* \to \tilde{X}^*$ *has a bounded inverse on* $\mathcal{R}(C_1^*)$ *if and only if* $\mathcal{R}(C_1) = \tilde{Y}$. The proof of the "if" part is Theorem II.3.13 in [18] (and actually does not require the assumption that C_1 is closed). The proof of the "only if" part is Theorem II.4.3 in [18]. (See also [20].)

The principal theoretical result of this section is a generalization of Theorem 2.1 to encompass systems (1.1) and (1.5) and the notions of F-observability and F^*-controllability where F is not necessarily the identity operator. For convenience we enumerate three propositions:

(a) *System* (1.1) *is* F*-observable* (*cf. Definition* 1.1).

(b) *System* (1.5) *is* (*exactly*) *controllable* (*cf. Definition* 1.2).

(c) *System* (1.3) *is* F^**-controllable* (*cf. Definition* 1.2).

THEOREM 2.3. *The above propositions are equivalent.*

Before proving this theorem it will be convenient to establish a certain lemma. In this lemma $\tilde{F}$ refers to the mapping which carries the equivalence class,

$\tilde{x}$, of x in $X_F = X/\ker(F)$ into the element $F(x) \in Z$. (Refer to § 1 for definitions of X^F, $\|\cdot\|_F$, etc.)

LEMMA 2.4. *We have* (*cf.* (1.5))

$$(X_F)^* \xrightarrow{\mathscr{F}} X^F = \mathscr{R}(F^*),$$

where

$$\mathscr{F} = F^F(\tilde{F}^{-1})^*$$

is an isometry of $(X_F)^*$ *onto* X^F.

Proof. Since $\|\tilde{x}\|_{X_F} = \|\tilde{F}(\tilde{x})\|_Z = \|x\|_F$ it is clear that $\tilde{F}^{-1}: \mathscr{R}(F) \subseteq Z \to X_F$ and $\tilde{F}^{-1}F: X \to X_F$ have norm 1 with respect to $\|\cdot\|_F$ and $\|\cdot\|_{X_F}$. Hence $\mathscr{F}$ is bounded with norm 1. Now if $\xi \in X^F$, $\xi(x)$ has the same value for all x in an equivalence class $\tilde{x}$ and

$$\tilde{\xi}(\tilde{x}) = \tilde{\xi}(\tilde{F}^{-1}F(x)) = \xi(x)$$

defines a linear functional on X_F. Thus, $\xi = F^F(\tilde{F}^{-1})^*\tilde{\xi}$ and $\mathscr{F}$ is onto. Since

$$|\tilde{\xi}(\tilde{F}^{-1}F(x))| = |\xi(x)| \leqq \|\xi\|_{X^F}\|x\|_F = \|\xi\|_{X^F}\|\tilde{x}\|_{X_F}$$

we conclude that $\mathscr{F}^{-1}$ also has norm 1 and $\mathscr{F}$ is, therefore, an isometry.

Each element of X^F is continuous with respect to the F-topology of X and therefore also with respect to the original topology of X since F is bounded. Hence $X^F \subseteq \mathscr{R}(F^*)$. But if $\xi \in \mathscr{R}(F^*)$ we have

$$\xi = F^*\zeta, \qquad \zeta \in Z^*$$

and

$$|\xi(x)| = |\zeta(Fx)| \leqq \|\zeta\|_{Z^*}\|x\|_F$$

and we conclude $\xi \in X^F$. Hence $X^F = \mathscr{R}(F^*)$ and the proof is complete.

Proof of Theorem 2.3. From the definition of C^F in § 1 we see that

$$\mathscr{R}(C^*) \cap X^F = \mathscr{R}(C^F) \subseteq \mathscr{R}(C^*). \tag{2.3}$$

To show (b)$\Rightarrow$(c) we assume (cf. Definition 1.2) that $\mathscr{R}(C^F) = X^F$. Then Lemma 2.4 with (2.3) gives

$$\mathscr{R}(F^*) = X^F = \mathscr{R}(C^F) \subseteq \mathscr{R}(C^*)$$

so that $\mathscr{R}(C^*) \supseteq \mathscr{R}(F^*)$ and (c) is established.

To show (c)$\Rightarrow$(b) we assume $\mathscr{R}(F^*) \subseteq \mathscr{R}(C^*)$. Lemma 2.4 then gives

$$X^F \subseteq \mathscr{R}(C^*).$$

Then (2.3) implies

$$X^F = \mathscr{R}(C^*) \cap X^F = \mathscr{R}(C^*) = \mathscr{R}(C^F)$$

and we have (b).

The proof of (c)$\Rightarrow$(a) is much the same as the proof of necessity in Theorem 2.1 as it appears in [20]. We suppose that (c) is true, i.e.,

$$\mathscr{R}(F^*) \subseteq \mathscr{R}(C^*).$$

Then for each $\xi \in X^F (= \mathcal{R}(F^*)$ from Lemma 2.4) we have, for some $\eta \in Y^*$,

$$|\xi(x)| = |(C^*\eta)(x)| = |\eta(Cx)| \leqq \|\eta\|_{Y^*}\|Cx\|_Y$$

and thus

$$\Phi_x(\xi) = \frac{\xi(x)}{\|C(x)\|_Y}$$

defines, for $C(x) \neq 0$, a family of continuous linear functionals on X^F satisfying

$$|\Phi_x(\xi)| \leqq \|\eta\|_{Y^*}, \qquad \xi \in X^F (\eta \text{ depending on } \xi).$$

The Φ_x correspond, as in Lemma 2.4, to continuous linear functionals on X_F, and hence on the Banach space $\bar{X}_F$ and, applying the principle of uniform boundedness, we have, for some $K > 0$,

$$|\Phi_x(\xi)| \leqq K\|\xi\|_{X^F}$$

or

$$|\xi(x)| \leqq K\|\xi\|_{X^F}\|Cx\|_Y. \tag{2.4}$$

Since X^F and $(X_F)^*$ are (from Lemma 2.4) isomorphic, and $X_F \subseteq (X_F)^{**}$ and $\|x\|_F = \|x\|_{X_F} = \|x\|_{X_F^{**}}$ we conclude from (2.4) that

$$\|x\|_F \leqq K\|Cx\|_Y, \qquad C(x) \neq 0,$$

and then, clearly, we have (c):

$$\|x\|_F \leqq K\|Cx\|_Y, \qquad x \in \mathcal{D}(C).$$

The proof that (a) $\Rightarrow$ (c) proceeds as follows. Given a linear functional $\xi \in X^F$ we proceed to construct a linear functional η on $\mathcal{R}(C)$. If $y \in \mathcal{R}(C)$ we have

$$y = Cx, \qquad x \in \mathcal{D}(C).$$

Assuming (a), i.e.,

$$\|x\|_F \leqq K\|Cx\|_Y, \tag{2.5}$$

we see that if $y = Cx_1$ and also $y = Cx_2$, then

$$0 = \|Cx_1 - Cx_2\|_Y \geqq \|x_1 - x_2\|_F,$$

and hence,

$$|\xi(x_1) - \xi(x_2)| \leqq \|\xi\|_{X^F}\|x_1 - x_2\|_F = 0.$$

We may therefore define η unambiguously for $y \in \mathcal{R}(C)$ by

$$\eta(y) = \eta(Cx) = \xi(x)$$

and we have the estimate, using (2.5),

$$|\eta(y)| \leqq \|\xi\|_{X^F}\|x\|_F \leqq K\|\xi\|_{X^F}\|Cx\|_Y = K\|\xi\|_{X^F}\|y\|_Y,$$

and η is continuous on $\mathcal{R}(C) \subseteq Y$. Applying the Hahn–Banach theorem, there is a linear functional $\hat{\eta} \in Y^*$ which extends η and satisfies

$$|\hat{\eta}(y)| \leqq \|\xi\|_{X^F} \|y\|_Y, \qquad y \in Y.$$

Thus it remains only to show that $\hat{\eta} \in \mathcal{D}(C^F)$ and $C^F(\hat{\eta}) = \xi$. But this is clear, for if $x \in \mathcal{D}(C)$, then $Cx \in \mathcal{R}(C)$ and

$$\hat{\eta}(Cx) = \eta(Cx) = \xi(x)$$

defines an F-continuous linear functional on $\mathcal{D}(C)$; that linear functional is the restriction of ξ to $\mathcal{D}(C)$. Since $\mathcal{D}(C)$ is dense in X_F, $C^F\hat{\eta}$ is determined by its values on $\mathcal{D}(C)$. Thus $C^F\hat{\eta} = \xi$ and $\xi \in \mathcal{R}(C^F)$. Since ξ is arbitrary in X^F we have $\mathcal{R}(C^F) = X^F$ and the proof is complete.

We pass now to a discussion of the situation obtained when the control process takes place in spaces X, Y, Z which are not necessarily dual spaces while the observation process takes place in X^*, Y^*, Z^*.

THEOREM 2.5. *Consider the system*

$$\begin{array}{ccc} X \supseteq \mathcal{D}(C) & \xrightarrow{C} & Y \\ & \overset{F}{\searrow} & \uparrow \\ & & Z \end{array} \tag{2.6}$$

and suppose the system to be F-controllable, i.e., that

$$\mathcal{R}(C) \supseteq \mathcal{R}(F). \tag{2.7}$$

Then the system

$$\begin{array}{ccc} X^* & \longleftarrow & \mathcal{D}(C^*) \subseteq Y^* \\ & & \overset{F^*}{\swarrow} \\ & & Z^* \end{array} \tag{2.8}$$

is F^*-observable, i.e., for some $K > 0$,

$$\|F^*\eta\|_{Z^*} \leqq K\|C^*\eta\|_{X^*}, \qquad \eta \in \mathcal{D}(C^*). \tag{2.9}$$

Moreover, if we assume

(i) *C is closed;*

(ii) *$\mathcal{R}(C^*)$ is dense in* $\ker(C)^\perp = (X/\ker(C))^*$;

and

(iii) *$F^{-1}(\mathcal{R}(C))$ is dense in Z;*

(iv) *$\mathcal{R}(F)$ is dense in $\mathcal{R}(C)$, or, $C^{-1}(\mathcal{R}(F))$ is dense in X;*

then F^*-observability, (2.9), implies F controllability, (2.7).

Remarks. Requirement (iii) above amounts, in practice, to the assumption that a dense set of initial states can be controlled to zero, i.e. approximate null controllability. When X is reflexive, (ii) is always true for closed C. Theorem 2.3 shows, of course, that (2.9) implies (2.7) in the case of reflexive X, Y without any need for (ii), (iii), and (iv) as special assumptions.

Proof. Since $F\colon Z \to Y$ is bounded, $\ker(F)$ is closed and $Z/\ker(F)$ is a Banach space with the norm (see, e.g. [1])

$$\|\hat{z}\|_{Z/\ker(F)} = \inf_{z \in \hat{z}} \|z\|_Z,$$

$\hat{z}$ being the equivalence class containing z. Let $\hat{F}$ be the one-to-one map from $Z/\ker(F)$ onto $\mathcal{R}(F) \subseteq Y$ induced by F. Then it is clear that $\mathcal{R}(F)$ becomes a Banach space isomorphically equivalent to $Z/\ker(F)$ if we put

$$\|y\|_{\mathcal{R}(F)} = \|\hat{F}^{-1} y\|_{Z/\ker(F)}. \tag{2.10}$$

The new topology in $\mathcal{R}(F)$ is stronger than the one induced by the topology in Y so we conclude that each $\eta \in Y^*$ is an extension of some $\hat{\eta} \in \mathcal{R}(F)^*$, indeed

$$\begin{aligned} \|\hat{\eta}\|_{\mathcal{R}(F)^*} &= \sup_{\substack{y \in \mathcal{R}(F) \\ \|y\|_{\mathcal{R}(F)} = 1}} |\hat{\eta} y| = \sup_{\substack{\hat{z} \in Z/\ker(F) \\ \|\hat{z}\|_{Z/\ker(F)} = 1}} |\hat{\eta} \hat{F} \hat{z}| \\ &= \|\hat{F}^* \eta\|_{(Z/\ker(F))^*} = \|F^* \eta\|_{Z^*}. \end{aligned} \tag{2.11}$$

The last equality is established in [18] as Lemma II.4.7.

Suppose we have F-controllability, i.e.

$$\mathcal{R}(C) \supseteq \mathcal{R}(F).$$

Let C_1 be the restriction of C to the normed linear space $X_1 = C^{-1}(\mathcal{R}(F))$. Then C_1 maps X_1 onto the Banach space $\mathcal{R}(F)$ (with the topology (2.10)) and [18, Thm. II.3.13] we have

$$\|\hat{\eta}\|_{\mathcal{R}(F)^*} \leqq K \|C_1^* \hat{\eta}\|_{X_1^*}, \qquad \hat{\eta} \in \mathcal{R}(F)^*, \tag{2.12}$$

for some fixed $K > 0$. Restricting $\hat{\eta}$ to those linear functionals in $\mathcal{R}(F)^* \cap \mathcal{D}(C_1^*)$ which arise from $\eta \in Y^*$, as discussed above, we have, combining (2.11) and (2.12),

$$\|F^* \eta\|_{Z^*} \leqq K \|C_1^* \hat{\eta}\|_{X_1^*}. \tag{2.13}$$

But

$$\begin{aligned} \|C_1^* \hat{\eta}\|_{X_1^*} &= \sup_{\substack{x \in X_1 \\ \|x\|_{X_1} = 1}} |\hat{\eta} C_1 x| = \sup_{\substack{x \in X_1 \\ \|x\|_{X_1} = 1}} |\eta C x| \\ &\leqq \sup_{\substack{x \in \mathcal{D}(C) \\ \|x\|_X = 1}} |\eta C x| = \|C^* \eta\|_{X^*}. \end{aligned} \tag{2.14}$$

Combining (2.13) and (2.14) we have (2.9) and the first part of Theorem 2.5 is established.

Now we assume F^*-observability, (2.9), together with hypotheses (i)–(iv) of our theorem and we demonstrate F-controllability, (2.7). First we show that, without loss of generality, we may assume that C and F are injective, i.e., one-to-one.

Since C is assumed closed, $\hat{X} = X/\ker C$ is a Banach space with the norm

$$\|\hat{x}\|_{\hat{X}} = \inf_{x \in \hat{x}} \|x\|_X.$$

The operator $\hat{C}$, defined on $\mathscr{D}(C)/\ker(C)$ by

$$\hat{C}\hat{x} = Cx, \qquad x \in \hat{x}$$

is closed with dense domain $\mathscr{D}(\hat{C}) \subseteq \hat{X}$ and, clearly,

$$\mathscr{R}(C) = \mathscr{R}(\hat{C}).$$

Now

$$\begin{aligned} \|\hat{C}^*\eta\|_{\hat{X}^*} &= \sup_{\hat{x}\neq 0} \frac{|\eta(\hat{C}\hat{x})|}{\|\hat{x}\|_{\hat{X}}} = \sup_{\hat{x}\neq 0} \frac{|\eta Cx|}{\inf_{x\in\hat{x}} \|x\|_X} \\ &= \sup_{\hat{x}\neq 0} \sup_{x\in\hat{x}} \frac{|\eta(Cx)|}{\|x\|_X} = \|C^*\eta\|_{X^*} \end{aligned} \tag{2.15}$$

from which it is clear that $\mathscr{D}(\hat{C}^*) = \mathscr{D}(C^*)$ in Y^*. In the same way, since F bounded implies $\ker(C)$ is closed, we may define

$$\hat{F}\hat{z} = Fz, \quad z \in \hat{z} \in \hat{Z} = Z/\ker(F).$$

We see readily that $\hat{F}$ is bounded and

$$\|\hat{F}^*\eta\|_{\hat{Z}^*} = \|F^*\eta\|_{Z^*}, \qquad \eta \in Y^*.$$

The spaces X and Z may therefore be replaced by $\hat{X}, \hat{Z}$ and C, F by the one-to-one maps $\hat{C}, \hat{F}$. Thus, we may assume C and F to be injective.

Let a new topology be defined on $\mathscr{R}(F)$ by

$$\|y\|_F = \|F^{-1}y\|_Z. \tag{2.16}$$

With this norm, which yields an F-topology stronger than that induced on $\mathscr{R}(F)$ by the original topology of Y, $\mathscr{R}(F)$ becomes a Banach space which we shall denote by Y_F. Clearly $\mathscr{R}(F)^* \subseteq Y_F^*$, where $\mathscr{R}(F)^*$ denotes the dual space of $\mathscr{R}(F)$ with respect to the induced topology.

In the same way we may invest $\mathscr{R}(C)$ with a C-topology and norm $\|\cdot\|_C$ so that it becomes a Banach space Y_C. (We remark that the norms $\|\cdot\|_F$ and $\|\cdot\|_C$ may well be incommensurate.)

Let us define a subspace

$$X_1 = \overline{C^{-1}(\mathscr{R}(F))} \subseteq X$$

and an operator C_1, the restriction of C to X_1. Evidently (2.7) is equivalent to

$$\mathscr{R}(C_1) = Y_F. \tag{2.17}$$

Now $C_1: X_1 \to Y_F$ is densely defined and closed, the latter since the F topology of $\mathscr{R}(F)$ is stronger than the original induced topology. Let C_1^F denote the adjoint operator $C_1^F: Y_F^* \to X_1^*$. Using [18, Thm. II.4.3] or [20, Thm. 5.3], we see that (2.17) follows if we can establish that

$$\|\eta\|_{Y_F^*} \leqq \hat{K}\|C_1^F\eta\|_{X_1^*} \tag{2.18}$$

for all $\eta \in Y_F^*$. It becomes, therefore, a question of showing that (2.9) implies (2.18).

For $\eta \in \mathcal{R}(F)^*$ (which is the restriction to $\mathcal{R}(F)$ of some $\hat{\eta} \in Y^*$)

$$\|F^*\eta\|_{Z^*} = \sup_{z \neq 0} \frac{|\eta(Fz)|}{\|z\|_Z} = \sup_{y \in \mathcal{R}(F)} \frac{|\eta y|}{\|y\|_F} = \|\eta\|_{Y_F^*} \tag{2.19}$$

and also

$$C_1^F \eta = C_1^* \eta. \tag{2.20}$$

In general $\|C_1^*\eta\|_{X_1^*} \leqq \|C^*\hat{\eta}\|_{X^*}$ because

$$\|C^*\hat{\eta}\|_{X^*} = \sup_{\substack{x \in \mathcal{D}(C) \\ x \neq 0}} \frac{|\hat{\eta} C x|}{\|x\|_X},$$

$$\|C_1^*\eta\|_{X_1^*} = \sup_{\substack{x \in \mathcal{D}(C_1) \\ x \neq 0}} \frac{|\eta C_1 x|}{\|x\|_{X_1}} = \sup_{\substack{x \in C^{-1}(\mathcal{R}(F)) \\ x \neq 0}} \frac{|\eta C x|}{\|x\|_X},$$

and $C^{-1}(\mathcal{R}(F)) \subseteq \mathcal{D}(C)$. But our hypothesis (iv), that $C^{-1}(\mathcal{R}(F))$ is dense in X, shows that in our case the two must be equal, i.e. (cf. (2.20))

$$\|C_1^F \eta\|_{X_1^*} = \|C_1^*\eta\|_{X_1^*} = \|C^*\hat{\eta}\|_{X^*} \tag{2.21}$$

if η is the restriction to $\mathcal{R}(F)$ of $\hat{\eta} \in Y^*$. Thus (2.9), (2.18), (2.20) give, for such η,

$$\|\eta\|_{Y_F^*} = \|F^*\eta\|_{Z^*} \leqq K\|C_1^F\eta\|_{X_1^*}. \tag{2.22}$$

It remains, therefore, only to show that (2.20) may be extended to all $\eta \in Y_F^*$ (i.e., that it is not just valid for those $\eta \in \mathcal{R}(F)^*$, which are restrictions of $\hat{\eta} \in Y^*$). Let $\eta \in Y_F^*$ be such that $C_1^F\eta$ is defined. From (ii) it follows that $\mathcal{R}(C_1^*)$ is dense in X_1^*. Hence there is a sequence $\{\eta_k\} \subseteq \mathcal{R}(F)^*$ such that

$$\lim_{k \to \infty} \|C_1^*\eta_k - C_1^F\eta\|_{X_1^*} = 0. \tag{2.23}$$

From (2.18) the sequence $\{\eta_k\}$ is Cauchy in Y_F^* and there is some $\eta_0 \in Y_F^*$ such that

$$\lim_{k \to \infty} \|\eta_k - \eta_0\|_{Y_F^*} = 0.$$

Now C_1^F, being an adjoint operator, is closed so we have

$$\lim_{k \to \infty} C_1^*\eta_k = \lim_{k \to \infty} C_1^F\eta_k = C_1^F\eta_0$$

and (cf. (2.23))

$$C_1^F\eta_0 = C_1^F\eta. \tag{2.24}$$

From (iii) we now conclude that $\mathcal{R}(C_1)$ is dense in Y_F, or equivalently, that $\ker(C_1^F) = \{0\}$ and (2.24) gives

$$\eta_0 = \eta.$$

Substituting the η_k for η in (2.18) and passing to the limit we conclude that (2.18) remains valid for $\eta \in \mathcal{D}(C_1^F)$. Then (2.17) and (2.7) follow and our proof is complete.

3. Some applications. We shall begin this part of our paper by applying the results of § 2 in the case of semigroups which correspond to linear hyperbolic partial differential equations. These are rather simple from the point of view of observation since their time reversibility gives the equivalence of terminal state observation (corresponding to $F = e^{AT}$ in § 1) and initial state observation (corresponding to $F = 1$ in § 1). It is also didactically convenient to consider hyperbolic processes first because some of the results in that case have implications for parabolic processes.

Consider, then, the wave equation

$$\frac{\partial^2 w}{\partial t^2} - \sum_{k=1}^{n} \frac{\partial^2 w}{(\partial x_k)^2} = 0, \qquad x \in \Omega, \quad t \geqq 0, \tag{3.1}$$

with Ω an open, bounded, connected region in R^n with (at least piecewise) smooth boundary Γ. The almost everywhere uniquely defined unit outward normal vector to Γ at $x \in \Gamma$ will be denoted by $\nu(=\nu(x), x \in \Gamma)$ and the corresponding directional derivative by $\partial/\partial\nu$. We assume

$$\Gamma = \Gamma_0 \cup \Gamma_1$$

where Γ_1 is nonempty, the pair (Ω, Γ_1) is "star-complemented" (see [35]) and $\Gamma_0 = \Gamma - \Gamma_1$. The initial-boundary value problem consisting of (3.1) and

$$\begin{aligned} &w(x, 0) = u_0(x), \qquad \frac{\partial w}{\partial t}(x, 0) = v_0(x), \\ &u_0 \in H^1(\Omega), \qquad v_0 \in L^2(\Omega), \\ &u_0(x) \equiv 0, \qquad x \in \Gamma_0, \end{aligned} \tag{3.2}$$

$$\begin{aligned} &w(x, t) \equiv 0, \qquad x \in \Gamma_0, \quad t \geqq 0, \\ &\frac{\partial w}{\partial \nu}(x, t) \equiv 0, \qquad x \in \Gamma_1, \quad t \geqq 0, \end{aligned} \tag{3.3}$$

is known to have a unique solution $w(x, t)$ such that $w \in C[0, T; H^1(\Omega)]$, $\partial w/\partial t \in C[0, T; L^2(\Omega)]$. Indeed we have

$$\left(w(\cdot, t), \frac{\partial w}{\partial t}(\cdot, t)\right) = S(t)(u_0, v_0),$$

where $S(t)$ is a strongly continuous group of bounded operators on $H^1_{\Gamma_0}(\Omega) \oplus L^2(\Omega)$. ($H^1_{\Gamma_0}(\Omega) = \{w \in H^1(\Omega) \mid w = 0 \text{ on } \Gamma_0\}$.)

We define the observing operator H: $H^1_{\Gamma_0}(\Omega) \times L^2(\Omega) \to H^{-1/2}(\Gamma_1)$ by

$$H\left(w(\cdot, t), \frac{\partial w}{\partial t}(\cdot, t)\right) = \frac{\partial \hat{w}}{\partial t}(\cdot, t)$$

where $\hat{w}$ denotes the restriction of w to $x \in \Gamma_1$. The trace theorem [24] shows that H, as thus defined, is a bounded linear operator. The corresponding observation operator is

$$C: \mathcal{D}(C) \subseteq H^1_{\Gamma_0}(\Omega) \oplus L^2(\Omega) \to H^{1/2}(\Gamma_1 \times [0, T])$$

given by

$$C(u_0, v_0) = H\left(w(\cdot,\cdot), \frac{\partial w}{\partial t}(\cdot,\cdot)\right) = \frac{\partial \hat{w}}{\partial t}(\cdot,\cdot). \tag{3.4}$$

The trace theorem and regularity results for solutions of (3.1), (3.3) show that $\mathscr{D}(C)$ includes $H^2_{\Gamma_0}(\Omega)\oplus H^1_{\Gamma_0}(\Omega)$ at least.

It is not easy to study the observation of (3.1), (3.2), (3.3) via (3.4) by direct means. But with the introduction of an auxiliary control system and use of the results in § 2 an indirect study becomes feasible.

Consider then the dual boundary control system consisting of the equation

$$\frac{\partial^2 z}{\partial t^2} - \sum_{k=1}^{n} \frac{\partial^2 z}{(\partial x^k)^2} = 0, \tag{3.5}$$

the terminal conditions

$$z(x, T) \equiv \frac{\partial z}{\partial t}(x, T) \equiv 0, \tag{3.6}$$

and the boundary conditions

$$\begin{aligned} &z(x, t) \equiv 0, \qquad x \in \Gamma_0, \quad t \geqq 0, \\ &\frac{\partial z}{\partial \nu}(x, t) = f(x, t), \qquad x \in \Gamma_1, \quad t \geqq 0. \end{aligned} \tag{3.7}$$

Using the divergence theorem as in [33] one obtains the relationship

$$\int_{\Gamma_1\times[0,T]} \frac{\partial \hat{w}}{\partial t}(x, t) f(x, t)\, ds\, dt = -\int_\Omega \left[\frac{\partial w(x,0)}{\partial t}\frac{\partial z(x,0)}{\partial t} + \sum_{k=1}^{n} \frac{\partial w(x,0)}{\partial x^k}\frac{\partial z(x,0)}{\partial x^k}\right] dx.$$

As written, this relationship holds when $(w(x,0), (\partial w/\partial t)(x,0)) = (u_0, v_0) \in H^2_{\Gamma_0}(\Omega)\oplus H^1_{\Gamma_0}(\Omega)$, so that $\partial \hat{w}/\partial t \in H^{1/2}(\Gamma_1\times[0,T]) \subset L^2(\Gamma_1\times[0,T])$, and for control functions $f \in L^2(\Gamma_1\times[0,T])$ which steer states $(z(\cdot,0), (\partial z/\partial t)(\cdot,0)) \in H^1_{\Gamma_0}(\Omega)\times L^2(\Omega)$ to zero at time T. But the relationship extends by continuity to similar $f \in H^{-1/2}(\Gamma_1\times[0,T])$, the dual of $H^{1/2}(\Gamma_1\times[0,T])$ relative to $L^2(\Gamma_1\times[0,T])$, and can be written then as

$$\int_\Omega \left[u_0(x)\frac{\partial z(x,0)}{\partial t} + \sum_{k=1}^{n} \frac{\partial v_0(x)}{\partial x^k}\frac{\partial z(x,0)}{\partial x^k}\right] dx = -\left\langle f, \frac{\partial \hat{w}}{\partial t}\right\rangle \tag{3.8}$$

where the last symbol denotes the value of the linear functional $f \in H^{-1/2}(\Gamma_1\times[0,T])$ at the point $\partial \hat{w}/\partial t$ in $H^{1/2}(\Gamma_1\times[0,T])$.

We now introduce the Hilbert space X consisting of the pairs $(u, v) \in H^1_{\Gamma_0}(\Omega)\oplus L^2(\Omega)$ but with inner product

$$((u, v), (\hat{u}, \hat{v}))_X = \int_\Omega \left[v(x)\hat{v}(x) + \sum_{k=1}^{n} \frac{\partial u(x)}{\partial x^k}\frac{\partial \hat{u}(x)}{\partial x^k}\right] dx$$

and associated norm $\|\cdot\|_X$. It is known [24] that this norm is equivalent to the norm in $H^1_{\Gamma_0}(\Omega)\times L^2(\Omega)$. The equation (3.8) shows that

$$C: \mathscr{D}(C) \subseteq X \to H^{1/2}(\Gamma_1\times[0,T]) \equiv Y$$

has as adjoint the operator $C^*: \mathcal{D}(C^*) \subseteq H^{-1/2}(\Gamma_1 \times [0, T]) \equiv Y^* \to X^* \equiv X$ defined by

$$C^* f = -\left(z(\cdot\,, 0), \frac{\partial z}{\partial t}(\cdot\,, 0)\right),$$

where the right-hand member is the initial state steered to zero at time T when the control f is used in (3.5), (3.6), (3.7). The domain of C^* consists of precisely those f for which $(z(\cdot\,, 0), (\partial z/\partial t)(\cdot\,, 0)) \in X$.

An easy modification of the results in [35] and [34] shows that the following controllability result obtains for the system (3.5), (3.6), (3.7): if T is sufficiently large, each state

$$z(\cdot\,, 0) = z_0 \in H^1_{\Gamma_0}(\Omega),$$

$$\frac{\partial z}{\partial t}(\cdot\,, 0) = z_1 \in L^2(\Omega),$$

i.e., each state in X, can be steered to zero at time T by use of a control $f \in H^{-1/2}(\Gamma_1 \times [0, T])$, i.e. C^* maps onto X. Applying Theorem 2.1 with X and Y as indicated above, we conclude that

$$\|(u_0, v_0)\|_X \leqq K \|C(u_0, v_0)\|_{H^{1/2}(\Gamma_1 \times [0,T])} \tag{3.9}$$

for each $(u_0, v_0) \in \mathcal{D}(C)$, T sufficiently large.

One could also use Theorem 2.2 to establish the cited controllability result from (3.9) but (3.9) is not an easy inequality to establish a priori.

The control results in [16], [17] show, in the special case wherein Ω is the unit ball in R^n and $\Gamma_1 = \partial\Omega$ is the $n-1$ dimensional sphere, that the spaces $H^{1/2}(\Gamma_1 \times [0, T])$ and $H^{-1/2}(\Gamma_1 \times [0, T])$ used above can both be replaced by $L^2(\Gamma_1 \times [0, T])$ and that "T sufficiently large" in that case can be replaced by $T > 2$. Whether this special result is achievable in the more general situation described above remains an intriguing question.

The spaces $H^{1/2}(\Gamma_1 \times [0, T])$ and $H^{-1/2}(\Gamma_1 \times [0, T])$ are rather unpleasant to use in applications. Using Theorem 2.3 we are able to derive a more usable result. It is shown in [35], [34] that each initial state

$$z(\cdot\,, 0) = z_0 \in H^2_{\Gamma_0}(\Omega),$$

$$\frac{\partial z}{\partial t}(\cdot\,, 0) = z_1 \in H^1_{\Gamma_0}(\Omega)$$

can be brought to zero in time T (T as described above) using a control $f \in L^2(\Gamma_1 \times [0, T])$. (In fact, $f \in H^{1/2}(\Gamma_1 \times [0, T])$, but that is not the emphasis here.) With the aid of some theorems from [24] (see [16] for more detail) one can show that the mapping

$$\hat{z}(\cdot\,, 0) = (-\Delta)^{-1/2} z(\cdot\,, 0),$$

$$\frac{\partial \hat{z}}{\partial t}(\cdot\,, 0) = (-\Delta)^{-1/2} \frac{\partial z}{\partial t}(\cdot\,, 0), \tag{3.10}$$

where $\Delta w = \sum_{k=1}^{n} \partial^2 w/(\partial x^k)^2$ is the (negative definite) Laplace operator in Ω with Dirichlet boundary conditions on Γ_0 and Neumann boundary conditions on Γ_1, carries X onto $H^2_{\Gamma_0}(\Omega) \oplus H^1_{\Gamma_0}(\Omega)$. Denoting the map (3.10) by $F^*: X \to X$, F^* is bounded, in fact compact, and its range is $H^2_{\Gamma_0}(\Omega) \oplus H^1_{\Gamma_0}(\Omega)$, a dense subspace of X. Using the same notation as in our earlier example, the operator C^*, now with domain consisting of functions $f \in L^2(\Gamma_1 \times [0, T])$ which steer states in $H^2_{\Gamma_0}(\Omega) \oplus H^1_{\Gamma_0}(\Omega)$ to zero at time T, is such that

$$\mathcal{R}(C^*) \supseteq \mathcal{R}(F^*).$$

Theorem 2.3 then shows that

$$\|F(u_0, v_0)\|_X \leqq K \|C(u_0, v_0)\|_{L^2(\Gamma_1 \times [0,T])} \tag{3.11}$$

for those (u_0, v_0) in X mapped by C into $L^2(\Gamma_1 \times [0, T])$, a domain which includes $H^2_{\Gamma_0}(\Omega) \oplus H^1_{\Gamma_0}(\Omega)$. Since we have the identity

$$((\hat{u}_0, \hat{v}_0), (u_0, v_0))_X = ((-\Delta)^{1/2} \hat{u}_0, (-\Delta)^{1/2} u_0)_{L^2(\Omega)} + (\hat{v}_0, v_0)_{L^2(\Omega)}$$

we have

$$\begin{aligned} ((\hat{u}_0, \hat{v}_0), F(u_0, v_0))_X &= ((-\Delta)^{1/2} \hat{u}_0, (-\Delta)^{1/2} (-\Delta)^{-1/2} u_0)_{L^2(\Omega)} + (\hat{v}_0, (-\Delta)^{-1/2} v_0)_{L^2(\Omega)} \\ &= (F\hat{u}_0, \hat{v}_0), (u_0, v_0))_X, \end{aligned}$$

since $(-\Delta)^{1/2}$, $(-\Delta)^{-1/2}$ are self-adjoint and commute with each other. Thus

$$F = F^*,$$

and (3.11) gives

$$\|u_0\|^2_{L^2(\Omega)} + \|(-\Delta)^{-1/2} v_0\|^2_{L^2(\Omega)} \leqq K^2 \|C(u_0, v_0)\|^2_{L^2(\Gamma_1 \times [0,T])}$$

or

$$\|u_0\|^2_{L^2(\Omega)} + (v_0, \Delta^{-1} v_0)_{L^2(\Omega)} \leqq K^2 \|C(u_0, v_0)\|^2_{L^2(\Gamma_1 \times [0,T])}. \tag{3.12}$$

This result gives a useful lower bound for observations on solutions $w(x, t)$ whose initial states are given in terms of eigenfunctions of the Laplace operator Δ. Essentially this type of estimate was carried out in a less abstract setting in [35]. Such lower bounds are used together with a Fourier transform technique outlined in [14], [35] to obtain certain results for control and observation of the heat equation which we outline below.

We let $\Omega, \Gamma, \Gamma_0, \Gamma_1$ be defined as above and consider the parabolic process (heat equation)

$$\frac{\partial w}{\partial t} - \sum_{k=1}^{n} \frac{\partial^2 w}{(\partial x^k)^2} = 0, \qquad x \in \Omega, \quad t \geqq 0, \tag{3.13}$$

$$\begin{cases} w(x, t) = 0, & x \in \Gamma_0, \quad t \geqq 0, \\ \dfrac{\partial w}{\partial \nu}(x, t) = 0, & x \in \Gamma_1, \quad t \geqq 0, \end{cases} \tag{3.14}$$

$$w(x, 0) = w_0(x) \in L^2(\Omega). \tag{3.15}$$

We introduce the observing operator

$$Hw(\cdot, t) = \hat{w}(\cdot, t),$$

where $\hat{w}$ denotes the restriction of w to $x \in \Gamma_1$. The observation operator is then

$$C: w_0 \in L^2(\Omega) \to \hat{w} \in L^2(\Gamma_1 \times [0, \tau]) \tag{3.16}$$

for some fixed $\tau > 0$.

In conjunction with (3.13), (3.14), (3.15) we consider the controlled parabolic process

$$\frac{\partial z}{\partial t} - \sum_{k=1}^{n} \frac{\partial^2 z}{(\partial x^k)^2} = 0, \qquad x \in \Omega, \quad t \geqq 0, \tag{3.17}$$

$$\begin{cases} z(x, t) = 0, & x \in \Gamma_0, \quad t \geqq 0, \\ \dfrac{\partial z}{\partial \nu}(x, t) = f(x, t), & x \in \Gamma_1, \quad t \geqq 0, \end{cases} \tag{3.18}$$

$$z(x, 0) = 0. \tag{3.19}$$

Again use of the divergence theorem yields the result

$$\int_{\Omega} w(x, 0) z(x, \tau)\, dx = \int_{\Gamma_1 \times [0,\tau]} \hat{w}(x, t) f(x, \tau - t)\, ds\, dt$$

or

$$(w_0, z(\cdot, \tau))_{L^2(\Omega)} = (\hat{w}, h)_{L^2(\Gamma_1 \times [0,\tau])},$$

where $h(\cdot, t) = f(\cdot, \tau - t)$.

Thus if we let $X = L^2(\Omega)$, $Y = L^2(\Gamma_1 \times [0, T])$ and define C by (3.16) on the domain consisting of those w_0 for which $\hat{w} \in L^2(\Gamma_1 \times [0, \tau])$ (a domain which is easily seen to be dense in X), the dual of C is

$$C^*: h \in L^2(\Gamma_1 \times [0, \tau]) \to z(\cdot, \tau), \tag{3.20}$$

i.e., C^* takes $f(x, \tau - t)$ into the final state $z(\cdot, \tau)$ realized when the control $f(x, t)$ is used in (3.17), (3.18), (3.19).

Let the eigenvalues of the Laplace operator

$$\Delta w = \sum_{k=1}^{n} \frac{\partial^2 w}{(\partial x^k)^2}$$

with Dirichlet boundary conditions on Γ_0 and Neumann boundary conditions on Γ_1 be $-\lambda_k$, $k = 1, 2, 3, \cdots$, and let the orthonormalized eigenfunctions be $\varphi_{k,l}$, $k = 1, 2, 3, \cdots$, $l = 1, 2, \cdots, m_k$, where m_k is the multiplicity of the eigenvalue $-\lambda_k$. Then

$$C\varphi_{k,l} = e^{-\lambda_k t} \hat{\varphi}_{k,l},$$

where $\hat{\varphi}_{k,l}$ denotes the restriction of $\varphi_{k,l}$ to Γ_1. In [35] a rather involved process is described whereby the estimate (3.12) obtained for the hyperbolic system leads to

the inequality

$$K\|e^{-\lambda_k t}\hat{\varphi}_{k,l}\|_{L^2(\Gamma_1\times[0,\tau])} \geqq \|e^{-M\lambda_k^{1/2}}\varphi_{k,l}\|_{L^2(\Omega)}, \tag{3.21}$$

valid for $k = 1, 2, 3, \cdots$, $l = 1, 2, \cdots, m_k$, with K and M certain positive constants, K depending on $\tau > 0$. This means that if we define $F: X \to X$ by

$$F = \exp[-M(-\Delta)^{1/2}]$$

then we have F-observability, i.e.,

$$K\|Cw_0\|_Y \geqq \|Fw_0\|_X, \qquad w_0 \in \mathscr{D}(C) \subseteq X. \tag{3.22}$$

Now final state observability for the system (3.13), (3.14), (3.15) holds when we have (3.22) with F replaced by

$$\tilde{F} = e^{\tau\Delta}.$$

But this is implied, with K replaced by some $\tilde{K} > 0$, by (3.22) because $\exp[-M(-\Delta)^{1/2}]$ and $\exp(\tau\Delta)$ are positive self-adjoint commuting operators with respective eigenvalues $\exp(-M\lambda_k^{1/2})$ and $\exp(-\tau\lambda_k)$ and $\lim_{k\to\infty}\lambda_k = +\infty$. This final state observability result agrees with those obtained in [14], [15], [28], [29].

Application of Theorem 2.3 now shows that the range of the operator C^* includes that of F^* (or of $\tilde{F}^*$). The inclusion

$$\mathscr{R}(C^*) \supseteq \mathscr{R}(F^*) = \mathscr{R}(\exp[-M(-\Delta)^{1/2}]) \tag{3.23}$$

means (cf. (3.20)) that the set of final states $z(\cdot, \tau)$ which may be reached from $z(\cdot, 0) = 0$ by means of controls $f \in L^2(\Gamma_1 \times [0, \tau])$ includes states

$$\sum_{k,j} \alpha_{k,j}\varphi_{k,j}$$

such that

$$\sum_{k,j} \exp[2M\lambda_k^{1/2}] \cdot |\alpha_{k,j}|^2 < \infty.$$

An interpretation of this condition and an indication that, in a certain sense, it is necessary as well as sufficient, appears in [14]. Theorem 2.3 also shows, of course, that if we have (3.23) then (3.21) holds also, the converse of the result described in our discussion above.

4. Observability of perturbed systems. A well-known result (see, e.g., [22]) states that the property of observability for the system

$$\begin{aligned} \frac{dx}{dt} &= Ax, \qquad x \in R^n, \\ \omega &= Hx, \qquad \omega \in R^r, \end{aligned} \tag{4.1}$$

is stable with respect to small perturbations of the matrices A and H. Thus if (4.1) is observable, i.e.

$$\operatorname{rank}\begin{bmatrix} H \\ HA \\ \vdots \\ HA^{n-1} \end{bmatrix} = n,$$

then the perturbed system

$$\frac{dx}{dt} = (A + \tilde{A})x,$$

$$\omega = (H + \tilde{H})x$$

remains observable if $\|\tilde{A}\|$ and $\|\tilde{H}\|$ are sufficiently small.

The purpose of the present section is to extend this result to certain infinite dimensional systems. In so doing we depart from the level of generality maintained in §§ 1 and 2 and consider an observed system

$$\frac{dx}{dt} = Ax, \qquad x \in X, \tag{4.2}$$

$$\omega = Hx, \qquad \omega \in Y. \tag{4.3}$$

Here X and Y are reflexive Banach spaces. We assume that A generates a strongly continuous semigroup of bounded operators $S(t)$ on X, yielding "solutions"

$$x(t) = S(t)x_0$$

of (4.2) corresponding to initial data $x(0) = x_0$. We assume $H\colon X \to Y$ is bounded.

Consider now the system wherein we have a strongly continuous and uniformly bounded perturbation $\tilde{A}(t)\colon X \to X$,

$$\frac{d\tilde{x}}{dt} = (A + \tilde{A}(t))\tilde{x}, \quad \tilde{x} \in X, \tag{4.4}$$

$$\tilde{\omega} = H\tilde{x}, \qquad \tilde{\omega} \in Y, \tag{4.5}$$

with solutions obeying the integral equation

$$\tilde{x}(t) = S(t)x_0 + \int_0^t S(t-s)\tilde{A}(s)\tilde{x}(s)\,ds. \tag{4.6}$$

THEOREM 4.1. *Let the system* (4.1), (4.2) *be (initial state) observable in time T in the sense that for some p, $1 < p < \infty$,*

$$K\|\omega(\cdot)\|_{L^p(0,T;Y)} = K\|Hx(\cdot)\|_{L^p(0,T;Y)} \geqq \|x_0\|_X \qquad (x(0) = x_0), \tag{4.7}$$

for solutions $x(t) = S(t)x_0$ of (4.6). *Then*:

(i) *If*

$$\|\tilde{A}(t)\| \leqq M, \qquad 0 \leqq t \leqq T,$$

and M is sufficiently small, then (4.7) *remains valid for* $\tilde{\omega}$, $\tilde{x}$ *of* (4.4), (4.5) *with K replaced by a (possibly larger) constant* $\tilde{K}$;

(ii) *If we have the inequality of* (i) *for any* $M>0$ *and* $\tilde{A}(t)$ *is compact for each* $t\in[0, T]$ *and strongly differentiable with*

$$\left\|\frac{d\tilde{A}(t)}{dt}\right\|\leqq M_0, \qquad 0\leqq t\leqq T,$$

then (4.7) *continues to hold whenever* (4.4), (4.5) *is distinguishable, i.e. whenever the identity* $\tilde{\omega}(t)\equiv 0$, $t\in[0, T]$ *implies* $\tilde{x}(t)\equiv 0$, $t\in[0, T]$.

Proof. The proof of part (i) goes very quickly. From the general theory of semigroups of linear operators in Banach space we know that there is a constant $M_1>0$ and a real number λ such that

$$\|S(t)\|\leqq M_1 e^{\lambda t}, \qquad t\geqq 0. \tag{4.8}$$

Using this together with (4.6), hypothesis (i), and a variation of the Gronwall inequality, one has the estimate

$$\|\tilde{x}(t)\|_X\leqq M_1 e^{(\lambda+MM_1)t}\|x_0\|_X$$

for solutions $\tilde{x}(t)$ of (4.6) with $x(0)=x_0$. Then

$$\begin{aligned}\|\hat{\omega}(t)\|_Y &\equiv \left\|H\int_0^t S(t-s)\tilde{A}(s)\tilde{x}(s)\,ds\right|_Y \\ &\leqq M(M_1)^2\|H\|\int_0^t e^{\lambda(t-s)}e^{(\lambda+M_1M)s}\,ds\|x_0\|_X \qquad (4.9)\\ &= M(M_1)^2\|H\|e^{\lambda t}\left(\frac{e^{M_1Mt}-1}{M_1M}\right)\|x_0\|_X = M_1\|H\|e^{\lambda t}(e^{M_1Mt}-1)\|x_0\|_X\end{aligned}$$

so that

$$\|\hat{\omega}\|_{L^p[0,T;Y]}\leqq T^{1/p}M_1\|H\|e^{|\lambda|T}(e^{M_1MT}-1)\|x_0\|_X.$$

Combined with (4.7) we have, for $\tilde{\omega}=\omega+\hat{\omega}$,

$$\begin{aligned}\|\tilde{\omega}\|_{L^p[0,T;Y]} &= \|HS(\cdot)x_0+\tilde{\omega}(\cdot)\|_{L^p[0,T;Y]} \\ &\geqq\left[\frac{1}{K}-T^{1/p}M_1\|H\|e^{|\lambda|T}(e^{M_1MT}-1)\right]\|x_0\|_X\equiv L\|x_0\|_X.\end{aligned} \tag{4.10}$$

Taking M sufficiently small, we have that L is positive and (i) has been proved for $\tilde{K}=1/L$.

Passing now to the proof of (ii) we note that the observation $\tilde{C}x_0=\tilde{\omega}(\cdot)=H\tilde{x}(\cdot)$ on solutions of (4.6) can be written

$$\tilde{C}x_0=Cx_0+\hat{C}x_0,$$

where (cf. (4.6), (4.9))

$$(\hat{C}x_0)(t)=\hat{\omega}(t)=H\int_0^t S(t-s)\tilde{A}(s)x(s)\,ds. \tag{4.11}$$

Since C has a bounded inverse on its range, an extension of the Fredholm alternative theorem [10] shows that $C+\hat{C}$ is boundedly invertible on its range provided $\ker(C+\hat{C})=\{0\}$ and $\hat{C}$ is compact. The condition that $\ker(C+\hat{C})=\{0\}$ is precisely the condition of distinguishability which we have imposed on the system (4.4), (4.5) so it remains only to show that $\hat{C}$ is a compact operator.

We establish the compactness of $\hat{C}$ by showing the equicontinuity, for bounded $\|x_0\|_X$, of the functions (cf. (4.6))

$$y_{x_0}(t)=\tilde{A}(t)\tilde{x}(t)=\tilde{A}(t)\,e^{At}x_0+\tilde{A}(t)\int_0^t e^{A(t-s)}\tilde{A}(s)\tilde{x}(s)\,ds. \tag{4.12}$$

For the second term here there is little to do, for

$$\begin{aligned}\left\|\tilde{A}(t_2)\int_{t_1}^{t_2}e^{A(t_2-s)}\tilde{A}(s)\tilde{x}(s)\,ds\right\|_X &\leqq (MM_1)^2\left|\int_{t_1}^{t_2}e^{\lambda(t_2-s)}\,e^{(\lambda+MM_1)s}\,ds\right|\|x_0\|_X\\ &= MM_1\,e^{\lambda t_2}\,|e^{MM_1t_2}-e^{MM_1t_1}|\,\|x_0\|_X\\ &\leqq MM_1\,e^{(|\lambda|+MM_1)T}|t_2-t_1|\,\|x_0\|_X\end{aligned} \tag{4.13}$$

and, using the differentiability of $\tilde{A}(t)$ (hypothesis (ii)) we have

$$\begin{aligned}&\|[\tilde{A}(t_2)-\tilde{A}(t_1)]\int_0^{t_1}S(t-s)\tilde{A}(s)\tilde{x}(s)\,ds\|_X\\ &\qquad\leqq M_0|t_2-t_1|M_1\,e^{|\lambda|T}\,(e^{M_1MT}-1)\|x_0\|_X.\end{aligned} \tag{4.14}$$

The calculations in (4.13) and (4.14) are similar to the one carried out in (4.9). Combining (4.13) and (4.14) we have the equicontinuity of the second term in (4.12) for bounded $\|x_0\|_X$.

We work only a little harder with the first term in (4.12). Because A generates a strongly continuous semigroup it is known (see [10], [19]) that

$$\|(A-\mu I)^{-1}\|\leqq\frac{M_1}{\mu-\lambda},\qquad \mu>\lambda,$$

where M_1, λ are as introduced earlier. Hence for every positive integer $j>\lambda$,

$$\|E_j\|\equiv\left\|\left(I-\frac{1}{j}A\right)^{-1}\right\|\leqq M_1\frac{j}{j-\lambda}\leqq M_2 \tag{4.15}$$

so that the E_j are uniformly bounded in norm by a constant $M_2>0$.

For $x\in\mathscr{D}(A)\subseteq X$ we have

$$x-\left(I-\frac{1}{j}A\right)^{-1}x=x-E_jx=\frac{-1}{j}\left(I-\frac{1}{j}A\right)^{-1}Ax\to 0,\qquad j\to\infty.$$

Since $\mathscr{D}(A)$ is dense in X and the E_j are uniformly bounded, we are able to conclude that

$$\lim_{j\to\infty}E_jx=x,\qquad x\in X. \tag{4.16}$$

Since X is reflexive it is known [1] that A^* also generates a strongly continuous semigroup on X^* and that semigroup is precisely $S(t)^*$. The above calculations may now be applied to A^* instead of A to yield (4.15) and (4.16), but with E_j now replaced by E_j^*. We proceed now to show that $E_j^*\tilde{A}(t)^*$ converges uniformly to $\tilde{A}(t)^*$ for each fixed value of t. Indeed, suppose this were not the case. Then for some $\varepsilon>0$ we should have a sequence $\{\xi_j\}\subseteq X^*$ with $\|\xi_j\|_{X^*}=1$, and

$$\|(\tilde{A}(t)^*-E_j^*\tilde{A}(t)^*)\xi_j\|_{X^*}>\varepsilon, \qquad j=1,2,3,\cdots.$$

Since the compactness of $\tilde{A}(t)$ implies that of $\tilde{A}(t)^*$ there is a subsequence, we still call it $\{\xi_j\}$, and a point $\eta\in X^*$ such that

$$\lim_{j\to\infty}\|\eta-\tilde{A}(t)^*\xi_j\|_{X^*}=0.$$

Now

$$\|(\tilde{A}(t)^*-E_j^*\tilde{A}(t)^*)\xi_j\|_{X^*}\leqq\|\tilde{A}(t)^*\xi_j-\eta\|_{X^*}+\|\eta-E_j^*\eta\|_{X^*}+\|E_j^*\eta-E_j^*\tilde{A}(t)^*\xi_j\|_{X^*}$$

so we conclude that

$$\lim_{j\to\infty}\|(\tilde{A}(t)^*-E_j^*\tilde{A}(t)^*)\xi_j\|_{X^*}=0,$$

contradicting our earlier supposition. We conclude that

$$\lim_{j\to\infty}\|\tilde{A}(t)^*-E_j^*\tilde{A}(t)^*\|=0.$$

It now follows from a well-known theorem that

$$\lim_{j\to\infty}\|\tilde{A}(t)-\tilde{A}(t)E_j\|=0.$$

Let $\varepsilon>0$ be given and let j_ε be chosen large enough so that

$$\|\tilde{A}(t)S(t)x_0-\tilde{A}(t)E_{j_\varepsilon}S(t)x_0\|_X\leqq\frac{\varepsilon}{3}\|x_0\|_X \tag{4.17}$$

uniformly for $0\leqq t\leqq T$. Since E_{j_ε} commutes with $S(t)$ and has range in $\mathcal{D}(A)$, and since $E_{j_\varepsilon}A$ is bounded, we may differentiate:

$$\left\|\frac{d}{dt}\tilde{A}(t)E_{j_\varepsilon}S(t)x_0\right\|_X\leqq\|\tilde{A}'(t)E_{j_\varepsilon}S(t)x_0\|_X+\|\tilde{A}(t)AE_{j_\varepsilon}S(t)x_0\|_X. \tag{4.18}$$

For the second term on the right-hand side of (4.18) we have

$$\begin{aligned}\|\tilde{A}(t)AE_{j_\varepsilon}S(t)x_0\|_X&=\left\|\tilde{A}(t)A\left(I-\frac{1}{j_\varepsilon}A\right)^{-1}S(t)x_0\right\|_X\\&=\|\tilde{A}(t)(-j_\varepsilon I+j_\varepsilon E_{j_\varepsilon})S(t)x_0\|_X \quad \text{(cf. (4.8), (4.15), (4.16))}\\&\leqq Mj_\varepsilon(M_2+1)M_1\,e^{|\lambda|T}\|x_0\|_X\equiv M_3\|x_0\|_X\end{aligned} \tag{4.19}$$

uniformly for $0\leqq t\leqq T$, $x_0\in X$. For the first term on the right-hand side of (4.18) we have (cf. (4.8), (4.15) and hypothesis (ii))

$$\|\tilde{A}'(t)E_{j_\varepsilon}S(t)x_0\|_X\leqq M_0M_2M_1\,e^{|\lambda|T}\|x_0\|_X\equiv\hat{M}_3\|x_0\|_X \tag{4.20}$$

uniformly for $0 \leqq t \leqq T$, $x_0 \in X$. Combining (4.19) with (4.20) we have

$$\left\| \frac{d}{dt} \tilde{A}(t) E_{j_\varepsilon} S(t) x_0 \right\|_X \leqq M_4 \|x_0\|_X, \qquad 0 \leqq t \leqq T, \quad x_0 \in X, \tag{4.21}$$

where $M_4 = M_3 + \hat{M}_3$.

Hence for $t_1, t_2 \in [0, T]$ with $|t_1 - t_2| < \varepsilon/(3M_4)$ we have, using (4.17) and (4.21),

$$\begin{aligned} \|\tilde{A}(t_2)S(t_2)x_0 - \tilde{A}(t_1)S(t_1)x_0\|_X &\leqq \|\tilde{A}(t_2)S(t_2)x_0 - \tilde{A}(t_2)E_{j_\varepsilon}S(t_2)x_0\|_X \\ &\quad + \|\tilde{A}(t_2)E_{j_\varepsilon}S(t_2)x_0 - \tilde{A}(t_1)E_{j_\varepsilon}S(t_1)x_0\|_X \\ &\quad + \|\tilde{A}(t_1)E_{j_\varepsilon}S(t_1)x_0 - \tilde{A}(t_1)S(t)x_0\|_X \\ &\leqq \left(\frac{\varepsilon}{3} + M_4\left(\frac{\varepsilon}{3M_4}\right) + \frac{\varepsilon}{3}\right)\|x_0\|_X \leqq \varepsilon \|x_0\|_X \end{aligned}$$

and the equicontinuity of $\tilde{A}(t)S(t)x_0$ for bounded $\|x_0\|_X$ has been established.

The compactness of $\tilde{A}(t)$ implies that for $t \in [0, T]$, $B > 0$, the set $\{\tilde{A}(t)S(t)x_0 \,|\, \|x_0\|_X \leqq B\}$ is compact in X. Thus the familiar diagonal process of the Arzela–Ascoli argument may be applied to show that if $\|x_k\|_X \leqq B$, $k = 1, 2, 3, \cdots$, then $\{\tilde{A}(t)S(t)x_{k_l}\}$ will be Cauchy in $\mathscr{C}[0, T; X]$, and hence in $L^p[0, T; X]$, for some subsequence $\{x_{k_l}\} \subseteq \{x_k\}$.

Returning now to (4.11), the above Cauchy property for $\tilde{A}(t)S(t)x_{k_l}$ together with the uniform boundedness of $\|S(t-s)\|$ and the boundedness of H shows that $\{\hat{C}x_{k_l}\}$ is Cauchy in $[L^p 0, T; Y]$. We conclude therefore that $\hat{C}$ is a compact operator and, as we have shown in the paragraph following (4.11), that is enough to establish the theorem.

Remark. It is easy to see that all of the above results continue to hold for certain unbounded observing operators H, provided that

$$\int_0^t \|HS(t-s)\| \, ds \leqq \hat{M}, \qquad 0 \leqq t \leqq T,$$

where $\hat{M}$ is an appropriate positive constant. Such a situation arises, for example, when the operator A of (4.2) is an unbounded negative definite operator on a Hilbert space X and $H = (-A)^r$, $r \leqq 1$.

COROLLARY 4.2. *Consider the control system*

$$\frac{dy}{dt} = A^* y + H^* u, \qquad y \in X^*, \quad u \in Y^*, \tag{4.22}$$

and the perturbed system

$$\frac{d\tilde{y}}{dt} = (A^* + \tilde{A}(t)^*)\tilde{y} + H^* u, \tag{4.23}$$

where A, $\tilde{A}(t)$ and H refer to the operators in Theorem 4.1. *If* (4.22) *is exactly controllable and* (4.23) *is approximately controllable then* (4.23) *is also exactly controllable.*

Proof. Apply Theorem 2.1. Refer to § 1 for discussion of the duality theory as it relates control of (4.22) and (4.23) to observation in (4.2), (4.3) and (4.4), (4.5), respectively.

We remark that the result of Corollary 4.2 could also be proved directly in a manner rather similar to that of Theorem 4.1.

The above-exhibited stability of initial state observability and exact controllability fails completely to carry over into a corresponding result for F-observability and/or F^*-controllability. Results in [7] and [8] have already shown, in the context of the heat equation, that arbitrarily small changes in the observing operator H can destroy F-observability. The following example shows that arbitrarily small compact constant perturbations $\tilde{A}$ in (4.4) can also destroy this property.

Let $-A$ be a positive definite self-adjoint operator on the Hilbert space X with eigenvalues

$$0<\lambda_1<\lambda_2<\cdots<\lambda_k<\lambda_{k+1}<\cdots,$$

and let corresponding orthonormal eigenvectors be $\varphi_1, \varphi_2, \cdots, \varphi_k, \cdots$. (Consider for example $A=\partial^2/\partial\xi^2$ in $L^2[0, 1]$ with Dirichlet boundary conditions at $x=0$ and $x=1$). We consider the system

$$\frac{dx}{dt}=Ax, \qquad x(0)=x_0\in X, \tag{4.24}$$

and the scalar observation

$$w(t)=(h, x(t)), \tag{4.25}$$

where $(\cdot\,,\cdot)$ is the inner product in X. We suppose that

$$h=\sum_{k=1}^{\infty} h_k\varphi_k \tag{4.26}$$

with

$$|h_{k+1}|\leqq\rho|h_k|, \quad |h_k|>0, \quad k=1, 2, 3, \cdots, \quad \sum_{k=1}^{\infty}|\lambda_k h_k|^2<\infty, \tag{4.27}$$

and assume that

$$T_0=-\lim_{k\to\infty}\left(\frac{\log|h_k|}{\lambda_k}\right)\geqq 0 \tag{4.28}$$

exists. (This is true for $A=\partial^2/\partial\xi^2$, for example, when the h_k decay like $1/(\lambda_k)^r$ for some positive integer r. Then $T_0=0$.) It is shown in [7] that, under these circumstances we have

$$K(T)\|w(\cdot)\|_{L^2[0,T]}\geqq\|e^{AT}x_0\|_X \tag{4.29}$$

for any $T>T_0$ with appropriate $K(T)>0$ and that no such inequality is possible for $T<T_0$. As noted in § 1, this is F-observability with $F=e^{AT}$ for $T>T_0$. We will demonstrate the existence of a compact operator $\tilde{A}$ and a $T_1>T_0$ such that

for the perturbed system

$$\frac{d\tilde{x}}{dt} = (A + \tilde{A})\tilde{x} \tag{4.30}$$

no inequality like (4.29) can hold for $T < T_1$, showing that F-observability, via the observation (4.25), is actually destroyed by the perturbation $\tilde{A}$ in passing from (4.24) to (4.30) when $T_0 < T < T_1$.

Let

$$\tilde{h} = \sum_{k=1}^{\infty} \tilde{h}_k \varphi_k \tag{4.31}$$

where (cf. (4.26))

$$\begin{aligned} \tilde{h}_k &= h_{k+1}\chi_k, \\ \tilde{h}_{k+1} &= -h_k\chi_k, \\ \chi_k &= \frac{h_{k+1}(1 - e^{-\lambda_{k+1}})}{h_k \|h\|_X^2}, \end{aligned} \tag{4.32}$$

all for odd values of k. From (4.27) it is clear that the $\tilde{h}_k$ defined by (4.31) satisfy

$$\sum_{k=1}^{\infty} |\lambda_k \tilde{h}_k|^2 < \infty. \tag{4.33}$$

We define the finite rank (and hence compact) operator K by

$$Kx = (h, x)\tilde{h} + (\tilde{h}, x)h, \qquad x \in X.^1 \tag{4.34}$$

We assume without loss of generality that $(I + K)^{-1}$ exists. (This can be ensured by modifying a finite number of the $\tilde{h}_k$ to make $\|K\|$ sufficiently small, if necessary. Such a change does not affect (4.33).) Then

$$(I + K)^{-1} = I - K(I + K)^{-1}$$

together with (4.33), (4.34) shows that

$$\begin{aligned} A + \tilde{A} &= (I + K)A(I + K)^{-1} \\ &= A + KA - AK(I + K)^{-1} - KAK(I + K)^{-1} \end{aligned}$$

is well defined and

$$\tilde{A} = KA - AK(I + K)^{-1} - KAK(I + K)^{-1},$$

being of finite rank, is compact.

Now the eigenvalues of $A + \tilde{A}$ are still λ_k, $k = 1, 2, 3, \cdots$, and the eigenvectors are $(I + K)\varphi_k$, $k = 1, 2, 3, \cdots$. Using the formula of [8] for the system (4.30)

[1] The symmetry of K with respect to h and $\tilde{h}$ is needed to ensure that $A + \tilde{A}$ is still "regular", as is required for the application of the critical time formula in [8].

with $\tilde{A}$ as constructed above, we see that the system (4.30), (4.25) is not $\tilde{F}$-observable ($\tilde{F}$ now being $e^{(A+\tilde{A})T}$) for $T<T_1$, where

$$T_1 = -\liminf_{k\to\infty}\left(\frac{\log|(h,(I+K)\varphi_k)|}{\lambda_k}\right).$$

We have

$$(h,(I+K)\varphi_k) = h_k + (h, K\varphi_k) = h_k + h_k(h,\tilde{h}) + \tilde{h}_k\|h\|_X^2.$$

But (4.32) shows that $(h,\tilde{h})=0$ and

$$(h,(I+K)\varphi_k) = h_k\left(1+\frac{\tilde{h}_k}{h_k}\|h\|_X^2\right)$$

and therefore (cf. (4.28))

$$T_1 = -\lim_{k\to\infty}\left(\frac{\log|h_k|}{\lambda_k}\right) - \liminf_{k\to\infty}\left(\frac{\log|1+(\tilde{h}_k/h_k)\|h\|_X^2|}{\lambda_k}\right). \tag{4.35}$$

To show that $T_1 > T_0$ (thereby completing our example) it is only necessary to show that the second term in (4.35) is positive. For odd k we have (cf. (4.32))

$$\frac{-\log|1+(\tilde{h}_{k+1}/h_{k+1})\|h\|_X^2|}{\lambda_{k+1}}$$

$$= \frac{-\log|1-(h_k/h_{k+1})\|h\|_X^2\cdot(h_{k+1}/(h_k\|h\|_X^2))(1-e^{-\lambda_{k+1}})|}{\lambda_{k+1}} = \frac{-\log e^{-\lambda_{k+1}}}{\lambda_{k+1}} = 1.$$

Hence $T_1 \geqq T_0+1$ and for $T_0<T<T_0+1$ the perturbation $\tilde{A}$ destroys the property of final state observability, i.e. for such T (4.24), (4.25) is e^{AT} observable but (4.30), (4.25) is not $e^{(A+\tilde{A})T}$ observable.

The extreme fragility of F-observability and the associated null controllability in the presence of very weak perturbations provides some explanation as to why the extension of linear control and observation theories for parabolic equations [14], [15], [25], [26], [28], [29], [35], [34] to cover comparable nonlinear equations appears to be proceeding at a much slower rate than in the case of hyperbolic equations [12], [35], [34], [4], [5], [13] where observability and F-observability, reachability and null controllability are, due to the time reversibility of the process, equivalent, and the results of Theorem 4.1 apply equally well to initial and terminal state observability (and the results of Corollary 4.2 apply equally well to null controllability and reachability.)

5. Remarks on optimal reconstruction. In Definition 1.1, § 1, we have defined what we mean by F-constructibility. Given a system

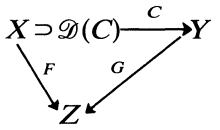
$$\begin{array}{ccc} X \supset \mathscr{D}(C) & \xrightarrow{\;C\;} & Y \\ & \searrow^{F} & \\ & & Z \end{array}$$

which we assume to be F-observable, the property of F-constructibility obtains if

there is a bounded operator

$$G\colon Y \to Z$$

such that $F = GC$, i.e., the following diagram is commutative:



One does not restrict the domain of definition of G to $\mathcal{R}(C)$ because one wishes to allow for the possibility that an observation $y = Cx$ may be "corrupted" by a noise or error term $\hat{y} \in Y$ with $\hat{y}$ not necessarily in $\mathcal{R}(C)$. The possibility of defining G on $\mathcal{R}(C)$ is implied by F-observability, as one sees quite readily. The possibility of extension of G to all of Y, thereby effecting a reconstruction of F, is in general a difficult question and is not fully explored here. It is clear that such an extension exists whenever there is a bounded projection $P\colon Y \to \overline{\mathcal{R}(C)}$. For wider discussion, see [31], [30].

Given a "corrupted" observation, $y + \hat{y}$, $y = Cx \in \mathcal{R}(C)$, $\hat{y} \in Y$, and given the existence of a reconstruction G, the reconstruction error can be defined as

$$\|Fx - G(Cx + \hat{y})\|_Z = \|G\hat{y}\|_Z \leqq \|G\|\|\hat{y}\|_Y.$$

The minimization of such reconstruction errors is consequently related to the norm of the reconstruction operator G. This prompts

DEFINITION 5.1. *The operator $\hat{G}$ provides an optimal reconstruction if*

$$\|\hat{G}\| = \min_{G \in \mathcal{G}} \|G\|,$$

where $\mathcal{G}$ is the class of all bounded operators $G\colon Y \to Z$ for which

$$(F - GC)x \equiv 0, \qquad x \in \mathcal{D}(C).$$

The existence of $\hat{G}$ can be demonstrated by a trivial variation of the argument given in [31], provided that Z is the dual of some Banach space (certainly true if Z is reflexive for example). In the present work we address ourselves to the question of uniqueness of $\hat{G}$. If $\hat{G}$ is unique, then in some sense it can be considered the natural reconstruction and one can, in applications, look for algorithms yielding its approximate realization. While this attractive situation sometimes does obtain, we shall see in the sequel that uniqueness is not to be expected in general.

Assuming F-observability, an operator $G \in \mathcal{G}$ is already determined on $\overline{\mathcal{R}(C)}$: there is an operator G_0 defined on $\overline{\mathcal{R}(C)}$ with

$$\sup_{\substack{y \in \overline{\mathcal{R}(C)} \\ \|y\|_Y = 1}} \|G_0 y\|_Z \leqq K$$

(where K is the constant in Definition 1.1) such that

$$Gy = G_0 y, \quad G \in \mathcal{G}, \quad y \in \overline{\mathcal{R}(C)}.$$

We shall see that the question of unique optimal extension of G_0, which is what optimal reconstruction amounts to, is closely related to whether or not G_0, as defined on $\overline{\mathcal{R}(C)}$, is a scalar multiple of an isometry.

LEMMA 5.2. *Assume that*

$$G_0: \overline{\mathcal{R}(C)} \to Z, \qquad \mathcal{R}(G_0) = Z,$$

is a scalar multiple of an isometry, i.e.

$$\|G_0 y\|_Z = \gamma \|y\|_Y, \qquad y \in \overline{\mathcal{R}(C)} \tag{5.1}$$

for some $\gamma > 0$. *If there is an optimal reconstruction operator* $\hat{G}$ *extending* G_0 *to* Y *with* $\|\hat{G}\| = K (\geqq \gamma)$, *then* $\hat{G}$ *is unique if for every* $\eta_0 \in \overline{\mathcal{R}(C)}^*$ *with*

$$\|\eta_0\|_{\overline{\mathcal{R}(C)}^*} = \gamma$$

there is only one extension of η_0, $\eta \in Y^*$, *with*

$$\|\eta\|_{Y^*} \leqq K.$$

Remark. There is, of course, at least one such extension η, as guaranteed by the Hahn–Banach theorem.

Proof. Clearly no generality is lost if we assume $\gamma = 1$. Suppose there were two extensions, $\hat{G}$ and $\tilde{G}$, of G_0 with $\|\hat{G}\| = \|\tilde{G}\| = K$. When G_0 is an isometry onto Z we have, for $\zeta \in Z^*$

$$\|G_0^* \zeta\|_{Y^*} = \sup_{\substack{y \neq 0 \\ y \in \overline{\mathcal{R}(C)}}} \frac{|\zeta(G_0 y)|}{\|y\|_Y} = \sup_{\substack{G_0 y \neq 0 \\ y \in \overline{\mathcal{R}(C)}}} \frac{|\zeta(G_0 y)|}{\|G_0 y\|_Z} = \sup_{\substack{z \neq 0 \\ z \in Z}} \frac{|\zeta z|}{\|z\|_Z} = \|\zeta\|$$

and G_0^* is also an isometry. Hence for $\zeta \in Z^*$ with $\|\zeta\|_{Z^*} = 1$, $\eta_0 = G_0^* \zeta$ is a linear functional on $\overline{\mathcal{R}(C)}^*$ with $\|\eta_0\|_{\overline{\mathcal{R}(C)}^*} = 1$. $\hat{G}^* \zeta$ and $\tilde{G}^* \zeta$ are then extensions of η_0 (i.e.,

$$(\hat{G}^* \zeta)_y = (\tilde{G}^* \zeta)_y = (G_0^* \zeta)_y, \qquad y \in \overline{\mathcal{R}(C)}^*)$$

with $\|\hat{G}^* \zeta\|_{Y^*}$ and $\|\tilde{G}^* \zeta\|_{Y^*}$ each $\leqq K$. The hypothesis of the lemma then yields

$$\hat{G}^* \zeta = \tilde{G}^* \zeta, \quad \zeta \in Z^*, \quad \|\zeta\| = 1.$$

Since this is true for all such ζ we conclude $\hat{G}^* = \tilde{G}^*$ and hence $\hat{G} = \tilde{G}_0$.

THEOREM 5.3. *Let* $G_0: \overline{\mathcal{R}(C)} \to Z$ *be a multiple of an isometry* (*cf.* (5.1)) *onto* Z. *If* Y *is reflexive and there is a projection* P *from* Y *onto the closed subspace* $\overline{\mathcal{R}(C)}$ *with*

$$\|P\| = 1,$$

then

$$\hat{G} = G_0 P$$

provides the unique optimal reconstruction operator for the observation operator C *with*

$$\|\hat{G}\| = \|G_0\| = \gamma.$$

Remark. When Y is a Hilbert space the existence of P is, of course, assured.

Proof of Theorem 5.3. Again we assume without loss of generality that $\gamma = 1$.

Let $\eta_0 \in \overline{\mathcal{R}(C)}^*$, $\|\eta_0\|_{\overline{\mathcal{R}(C)}^*} = 1$ and suppose that η_1 and η_2 are extensions to Y^* with

$$\|\eta_1\|_{Y^*} = \|\eta_2\|_{Y^*} = \|\eta_0\|_{\overline{\mathcal{R}(C)}^*} = 1.$$

Since Y is reflexive there is an element $y_0 \in Y$ with $\eta_0 y_0 = 1$. Let $\hat{y}$ be an arbitrary nonzero element of Y and let $\hat{Y}$ denote the two dimensional closed subspace of Y generated by y_0 and $\hat{y}$ with the topology induced by Y. Then $\hat{Y}$ is also reflexive and has a differentiable norm ([6]). Letting $\hat{\eta}_1$, $\hat{\eta}_2$ be the restrictions of η_1, η_2 to $\hat{Y}^*$, we see that $\hat{\eta}_1$ and $\hat{\eta}_2$ both have norm 1 and are solutions of

$$\min \|\hat{\eta}\|_{\hat{Y}^*},\ \hat{\eta} y_0 = 1.$$

Thus both $\hat{\eta}_1$ and $\hat{\eta}_2$ must be positive multiples of the differential of the norm function in $\hat{Y}$ (see [36]) and are therefore equal. Since this is true for each $\hat{y} \in Y$ we conclude $\eta_1 = \eta_2$ and η_0 has but one extension η with $\|\eta\|_{Y^*} = \|\eta_0\|_{\overline{\mathcal{R}(C)}^*}$.

If P is a projection onto $\overline{\mathcal{R}(C)}$ with $\|P\| = 1$, then

$$\hat{G} = G_0 P$$

is an extension of G_0 from $\overline{\mathcal{R}(C)}$ to Y with $\|\hat{G}\| = \|G_0\|$ and hence is an optimal reconstruction operator with norm

$$\|\hat{G}\| = K = 1.$$

Applying the above result on uniqueness of extensions of linear functionals together with Lemma 5.2 we conclude that G is the unique extension of G_0 with norm 1 and hence the unique optimal reconstruction operator.

As an example of application, let us consider the system whose evolution is described by the wave equation in R^1:

$$\frac{\partial^2 w}{\partial t^2} - \alpha^2 \frac{\partial^2 w}{\partial x^2} = 0, \quad w(0, t) = 0,$$

$$\frac{\partial w}{\partial x}(1, t) = 0, \qquad 0 \leqq x \leqq 1, \quad t \geqq 0, \tag{5.2}$$

and let the observing operator be

$$H\left(w(\cdot, t), \frac{\partial w}{\partial t}(\cdot, t)\right) = \frac{\partial w}{\partial t}(1, t).$$

The space X is the Hilbert space of initial displacements $w_0 \in H^1[0, 1]$, $w_0(0) = 0$, and initial velocities $v_0 \in L^2[0, 1]$ with the inner product

$$\langle (w_0, v_0), (\tilde{w}_0, \tilde{v}_0) \rangle_X = \int_0^1 \left(\alpha^2 \frac{\partial w_0}{\partial x} \frac{\partial \tilde{w}_0}{\partial x} + v_0 \tilde{v}_0 \right) dx$$

and norm

$$\|(w_0, v_0)\|_X = [\langle (w_0, v_0), (w_0, v_0) \rangle_X]^{1/2}.$$

The space Z coincides with X and F is the identity operator. For fixed T equal to a positive integer multiple of $2/\alpha$, i.e. $T = 2j/\alpha$, j a positive integer, the observation operator is

$$C\colon (w_0, v_0) \to \frac{\partial w}{\partial t}(1, \cdot) \in L^2[0, T] = Y, \tag{5.3}$$

where $w = w(x, t)$ is the unique generalized solution of (5.2) with initial data w_0, v_0. The initial states

$$w_{0,k}(x) = \frac{\sqrt{2}\sin((k+\frac{1}{2})\pi x)}{\alpha(k+\frac{1}{2})\pi}$$

$$v_{0,k}(x) = 0,$$

$$\hat{w}_{0,k}(x) = 0,$$

$$\hat{v}_{0,k}(x) = \sqrt{2}\sin((k+\tfrac{1}{2})\pi x)$$

form an orthonormal basis for X and correspond to solutions

$$w_k(x, t) = \frac{\cos(\alpha(k+\frac{1}{2})\pi t) w_{0,k}(x)}{\alpha(k+\frac{1}{2})\pi},$$

$$\hat{w}_k(x, t) = \frac{\sin(\alpha(k+\frac{1}{2})\pi t)\hat{v}_{0,k}(x)}{\alpha(k+\frac{1}{2})\pi}$$

for which the observations are

$$w_k(t) = -\sqrt{2}\sin(\alpha(k+\tfrac{1}{2})\pi t),$$

$$\hat{w}_k(t) = \sqrt{2}\cos(\alpha(k+\tfrac{1}{2})\pi t).$$

These form an orthogonal basis for $L^2[0, 2j/\alpha]$ with

$$\|w_k\|_{L^2[0,2j/\alpha]} = \|\hat{w}_k\|_{L^2[0,2j/\alpha]} = \frac{2j}{\alpha} \equiv \gamma.$$

The observation operator (5.3), which would at the outset be defined for initial states w_0, v_0 leading to continuously differentiable solutions, extends as a bounded operator $C\colon X \to L^2[0, 2j/\alpha]$ with C a multiple of an isometry, viz.:

$$\|w\|_{L^2[0,2j/\alpha]} = \|C(w_0, v_0)\|_{L^2[0,2j/\alpha]} = \frac{2j}{\alpha}\|(w_0, v_0)\|_X.$$

According to Theorem 5.3, then, there is only one optimal reconstruction operator, namely

$$G = C^{-1}P,$$

where P is the orthogonal projection from $L^2[0, 2j/\alpha]$ onto $\mathcal{R}(C) = \overline{\mathcal{R}(C)}$ = the span of the functions $\sin(\alpha(k+\frac{1}{2})\pi t)$, $\cos(\alpha(k+\frac{1}{2})\pi t)$, $k = 0, 1, 2, \cdots$, in $L^2[0, 2j/\alpha]$.

Despite this rather nice example, the typical situation is that optimal reconstruction operators are not unique. Even when we have a multiple of an isometry, as in Theorem 5.3, something like strict convexity is necessary. If we let $X = R^1$ with the absolute value as norm and let Y be R^2 with

$$\|(x_1, x_2)\|_{R^2} = |x_1| + |x_2|$$

(so that the unit ball in Y is not strictly convex) and let $C: R^1 \to R^2$ be given by

$$Cx = (x, 0),$$

then G_0, the inverse of C on its range, is defined by

$$G_0(x, 0) = x,$$

and clearly has norm 1. But there are many extensions of G_0 to transformations $G: R^2 \to R^1$ with $\|G\| = 1$. The transformations

$$G_\alpha(x_1, x_2) = x_1 + \alpha x_2, \qquad |\alpha| \leqq 1,$$

are easy examples.

When G_0 is not an isometry, which clearly must be considered the typical situation, all uniqueness vanishes, as we see from the following theorem applying to the Hilbert space case.

THEOREM 5.4. *Let us consider the system* (1.1) *with Y and Z assumed to be Hilbert spaces and suppose that $\overline{\mathcal{R}(C)} \neq Y$ and*

$$G_0: \overline{\mathcal{R}(C)} \subset Y \to Z$$

is not a multiple of an isometry. Then there are infinitely many extensions of G_0:

$$G: Y \to Z, \quad Gy = G_0 y, \quad y \in \overline{\mathcal{R}(C)},$$

for which $\|G\| = \|G_0\|$.

Remark. In view of Theorem 5.3 and the following remark, the interesting case in Theorem 5.4 is the case wherein $\overline{\mathcal{R}(G_0)} = Z$. This case is, of course, included since there is no assumption to the contrary.

Proof of Theorem 5.4. We may assume without loss of generality that $\|G_0\| = 1$ and that there is some element $y_0 \in \overline{\mathcal{R}(C)}$ such that $G_0 y_0 = z_0$ and

$$\|y_0\|_Y = 1, \qquad \|z_0\|_Z \equiv \mu_0^{1/2} < 1.$$

We will demonstrate that there are infinitely many extensions G of G_0,

$$G: Y \to Z, \quad Gy = G_0 y, \quad y \in \overline{\mathcal{R}(C)},$$

such that $\|G\| = 1$.

Since $\overline{\mathcal{R}(C)} \neq Y$ there is an element $y_1 \in \overline{\mathcal{R}(C)}^\perp$ with $\|y_1\|_Y = 1$. We let Y_1 be the closed subspace of Y spanned by $\mathcal{R}(C)$ and y_1 and let P be the orthogonal projection onto Y_1.

Let $G_0^*: Z \to \overline{\mathcal{R}(C)}$ be defined by

$$(G_0^* z, y)_Y = (z, G_0 y)_Z, \qquad y \in \overline{\mathcal{R}(C)}.$$

Then $G_0^*G_0$: $\overline{\mathcal{R}(C)} \to \overline{\mathcal{R}(C)}$ is a bounded self-adjoint operator and has the spectral representation

$$G_0^*G_0 = \int_{0-}^{1} \mu \, dE\mu,$$

where $E\mu$ is a spectral measure on $[0, 1]$. For each $y \in \overline{\mathcal{R}(C)}$ we then have

$$\|G_0 y\|_Z^2 = (G_0 y, G_0 y)_Z = (y, G_0^*G_0 y)_Y = \int_{0-}^{1} \mu \, d(y, E\mu y).$$

Since $\|G_0 y_0\|_Z^2 = \underline{\mu_0 \leq 1}$, the orthogonal projection $E\mu_0$ cannot be zero. We may then decompose $\mathcal{R}(C)$:

$$\overline{\mathcal{R}(C)} = E\mu_0\overline{\mathcal{R}(C)} \oplus (I - E\mu_0)\overline{\mathcal{R}(C)} \equiv R_0 \oplus R_0^{\perp}, \quad R_0 \equiv E\mu_0\overline{\mathcal{R}(C)}.$$

We note that

$$\|G_0 y\|_Z^2 = \int_{0-}^{\mu_0} \mu \, d(y, E\mu y) \leq \mu_0 \|y\|_Y^2, \qquad y \in R_0. \tag{5.4}$$

Each $y \in Y_1$ has the unique representation

$$y = r + \alpha y_1, \qquad r \in \overline{\mathcal{R}(C)}. \tag{5.5}$$

Let G_1 be defined on Y_1 by

$$G_1 y = G_1(r + \alpha y_1) = G_0 r + \alpha\varepsilon G_0 \hat{y}_0,$$

where $\hat{y}_0$ is an element of unit norm in R_0, and let G_0 then be extended to Y by setting

$$G = G_1 P,$$

where P is the orthogonal projection onto Y_1. Our theorem is consequently proved if we can show that

$$\|G_1\| \leq 1$$

for infinitely many values of ε.

For each $y \in Y_1$

$$\begin{aligned}
\|G_1 y\|_Z^2 &= \|G_0 r + \alpha\varepsilon G_0 \hat{y}_0\|_Z^2 \\
&= (r + \alpha\varepsilon \hat{y}_0, G_0^*G_0(r + \alpha\varepsilon \hat{y}_0))_Y \\
&= (r_0 + r_1 + \alpha\varepsilon \hat{y}_0, G_0^*G_0(r_0 + r_1 + \alpha\varepsilon \hat{y}_0))_Y \\
&= (r_1, G_0^*G_0 r_1)_Y + (r_0 + \alpha\varepsilon \hat{y}_0, G_0^*G_0(r_0 + \alpha\varepsilon \hat{y}_0))_Y,
\end{aligned}$$

where $r_0 = E\mu_0 r \in R_0$, $r_1 = (I - E\mu_0)r \in R_0^{\perp}$. The cross products of r_1 with $G_0^*G_0(r_0 + \alpha\varepsilon \hat{y}_0)$ vanish since R_0 is invariant under $G_0^*G_0$. We now have

$$\begin{aligned}
\|G_1 y\|_Z^2 &\leq \|G_0 r_1\|_Z^2 + \mu_0(\|r_0\|_Y^2 + \varepsilon^2 \|\alpha \hat{y}_0\|_Y^2) \\
&\quad + 2\varepsilon(r_0, G_0^*G_0(\alpha \hat{y}_0))_Y \\
&\leq \|r_1\|_Y^2 + \mu_0(\|r_0\|_Y^2 + \varepsilon^2 \|\alpha \hat{y}_0\|_Y^2) \\
&\quad + \varepsilon\mu_0(\|r_0\|_Y^2 + \|\alpha \hat{y}_0\|_Y^2),
\end{aligned} \tag{5.6}$$

repeatedly using the spectral representation for $G_0^* G_0$ and the fact that r_0, $\hat{y}_0$ lie in $E\mu_0$.

Let $\varepsilon > 0$ be small enough so that

$$\mu_0(1+\varepsilon) \leqq 1, \qquad \mu_0(\varepsilon^2+\varepsilon) \leqq 1. \tag{5.7}$$

Then (5.5) yields

$$\|G_1 y\|_Z^2 \leqq \|r_1\|_Y^2 + \|r_0\|_Y^2 + \|\alpha \hat{y}_0\|_Y^2 = \|r\|_Y^2 + \|\alpha y_1\|_Y^2 = \|y\|_Y^2, \tag{5.8}$$

the first equality following from the fact that y_1 and $\hat{y}_0$ each have unit norm. Since (5.7) is true for all $y \in Y_1$, $\|G_1\| \leqq 1$ for all ε satisfying (5.6) and, from our previous remarks, the theorem is proved.

In the case of boundary observation of the heat equation, as discussed in § 3, we see quite readily that as $k \to \infty$ the ratio of the norm in $L^2(\Gamma_1 \times [0, T])$ of the observation

$$\varphi_{k,j}(x) \exp(-\lambda_k t), \qquad x \in \Gamma_1, \tag{5.9}$$

on an eigenfunction solution $\varphi_{k,j}(x) \exp(-\lambda_k t)$ to the norm in $L^2(\Omega)$ of the final state

$$\varphi_{k,j}(x) \exp(-\lambda_k T) \tag{5.10}$$

tends to infinity. Thus in this case for every $\mu_0 > 0$ there is an element y_k in $\mathcal{R}(C)$ (namely (5.9) for k sufficiently large) such that $\|G_0 y_k\|_Z$ (i.e., the norm of (5.10)) is $< \mu_0$ and G_0 is not an isometry. Thus, while $G_0 P$, P being the orthogonal projection on $\mathcal{R}(C)$ (which is the closed span of the functions (5.8)) is an optimal reconstruction operator, it is not unique.

Acknowledgment. The authors wish to express their appreciation to Professor S. Rolewicz, Institute of Mathematics, Polish Academy of Sciences, and to Professor S. Kurcyusz, Institute of Automatic Control, Technical University of Warsaw, for extremely helpful conversations with the first author which aided in the formulation and clarification of certain concepts presented in this paper.

REFERENCES

[1] P. L. BUTZER AND H. BERENS, *Semigroups of Operators and Approximations*, Springer-Verlag, New York, 1967.

[2] H. T. BANKS, M. Q. JACOBS AND C. E. LANGENHOP, *Characterization of controlled states in $W_2^{(1)}$ of linear hereditary systems*, this Journal, 13 (1975), pp. 611–649.

[3] ———, *Function space controllability for linear functional differential equations*, Proc. Conf. on Differential Games and Control Theory, (Univ. of Rhode Island, Kingston, 1973); Marcel Dekker, New York, 1974.

[4] W. C. CHEWNING, *Controllability of the nonlinear wave equation in several variables*, this Journal, 14 (1976), pp. 19–25.

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

[6] M. M. DAY, *Normed Linear Spaces*, Springer-Verlag, Berlin, 1958.

[7] S. DOLECKI, *Observability for the one-dimensional heat equation*, Studia Math., 48 (1973), pp. 291–305.

[8] ———, *Observability for regular processes*, to appear.

[9] ———, *Duality of various notions of controllability and observability*, Proc. International Conference on Differential Equations, Univ. of Southern California, Los Angeles, 1974, Academic Press, New York, 1975.
[10] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part I: General Theory*, Interscience, New York, 1958.
[11] H. O. FATTORINI, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
[12] ———, *Controllability of higher order linear systems*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 301–311.
[13] ———, *Local controllability of a nonlinear wave equation*, Math. Systems Theory, 9 (1975), pp. 349–388.
[14] H. O. FATTORINI AND D. L. RUSSELL, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rational Mech. Anal., 43 (1971), pp. 272–292.
[15] ———, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. Appl. Math., 32 (1974), pp. 45–69.
[16] K. D. GRAHAM, *Separation of eigenvalues of the wave equation for the unit ball in R^n*, Studies in Appl. Math., 52 (1973), pp. 329–344.
[17] K. D. GRAHAM AND D. L. RUSSELL, *Boundary value control of the wave equation in a spherical region*, this Journal, 13 (1975), pp. 174–196.
[18] S. GOLDBERG, *Unbounded Linear Operators*, McGraw-Hill, New York, 1966.
[19] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[20] S. G. KREIN, *Linear Equations in Banach Space*, Nauka, Moscow, 1971 (in Russian).
[21] S. KURCYUSZ, Ph.D. Thesis, Institute of Automation, Warsaw Technical University.
[22] E. B. LEE AND L. W. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[23] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, Berlin, 1971.
[24] J. L. LIONS AND E. MAGENES, *Problèmes aux Limites Non-homogènes*, vols. 1, 2, Dunod, Paris, 1968.
[25] R. C. MACCAMY, V. J. MIZEL AND T. I. SEIDMAN, *Approximate boundary controllability of the heat equation*, J. Math. Anal. Appl., 23 (1968), pp. 699–703.
[26] ———, *Part II*, Ibid., 28 (1969), pp. 482–492.
[27] V. J. MIZEL AND T. I. SEIDMAN, *Observation and prediction for the heat equation*, Ibid., 28 (1969), pp. 303–312.
[28] ———, *Observation and prediction for the heat equation. II*, Ibid., 38 (1972), pp. 149–166.
[29] D. PRZEWORSKA-ROLEWICZ AND S. ROLEWICZ, *Equations in Linear Spaces*, Monografie Matematyczne, Vol. 56, Warsaw, 1968.
[30] S. ROLEWICZ, *On optimal observability of linear systems in infinite-dimensional states*, Studia Math., 44 (1972), pp. 411–416.
[31] ———, *On optimal observability of linear systems*, Berichte Ges. Math. Daten. mbH, Bonn, 77 (1973), pp. 139–141.
[32] D. L. RUSSELL, *Boundary value control of the higher dimensional wave equation*, this Journal, 9 (1971), pp. 29–42.
[33] ———, *Part II*, this Journal, 9 (1971), pp. 401–419.
[34] ———, *A unified boundary value controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., 52 (1973), pp. 189–211.
[35] ———, *Boundary value controllability of wave and heat processes in star-complemented regions*, Proc. Conference on Differential Games and Control Theory, (Kingston, R.I., 1973); Marcel Dekker, New York, 1974.
[36] ———, *The Kuhn–Tucker conditions in Banach space with an application to control theory*, J. Math. Anal. Appl., 15 (1966), pp. 200–212.
[37] T. I. SEIDMAN, *Problems of boundary control and observation for diffusion processes*, Math. Res. Rep. 73–10, Univ. of Maryland, Baltimore County, Baltimore, 1973.
[38] ———, *Observation and prediction for one dimensional diffusion equations*, Math. Res. Rep. 74–1, Univ. of Maryland, Baltimore County, Baltimore, 1974.
[39] ———, *A well-posed problem for the heat equation*, Bull. Amer. Math. Soc., 80 (1974), pp. 901–902.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 2, February 1977

NORMALIZED MARKOV DECISION CHAINS. II: OPTIMALITY OF NONSTATIONARY POLICIES*

URIEL G. ROTHBLUM†

Abstract. In this paper we consider finite state and action, discrete time parameters normalized Markov decision chains, i.e., Markov decision processes with transition matrices that are nonnegative with spectral radius not exceeding one (but not necessarily substochastic). We show that the periodical reward gained in period N is bounded by a polynom, uniformly over the set of all policies. The degree of this polynom can be obtained by considering only the set of stationary policies. Extending and improving results of Sladky (1974) for the stochastic case, we obtain necessary and sufficient conditions for n discount optimality of arbitrary (not necessarily stationary) policies.

1. Introduction. Normalized Markov decision chains were introduced in Rothblum (1975a). In this paper we give further results concerning these decision processes. In particular, we study optimality properties of policies that are not necessarily stationary. Known results for the stochastic case (e.g. Lippman (1968), Sladky (1974)) are improved and extended to the normalized case.

In § 2 we summarize a few notational conventions, and then our model and the n discount optimality criteria are introduced in § 3. In § 4 we show that the periodical reward gained in period N when a policy π is used, is bounded by a polynom in N, uniformly in π. In § 5 we develop necessary and sufficient conditions for n discount optimality of arbitrary (not necessarily stationary) policies. These conditions, which are stronger than those obtained by Sladky (1974) have a simple form when applied to stationary policies. We then show that the concepts of discount optimality coincide for all sufficiently large n.

2. Notational conventions. Let P be an $S \times S$ real matrix. We say that P is *nonnegative*, written $P \geqq 0$, if all its entries are nonnegative. We say P is *semipositive*, written $P > 0$, if $P \geqq 0$ and $P \neq 0$. We write $P \geqq (\text{resp.}, >)\ Q$ if $P - Q \geqq (\text{resp.}, >)\ 0$. Similar definitions apply to vectors. Let λ be a complex number and $Q \equiv P - \lambda I$. The *index* of λ for P, denoted $\nu_\lambda(P)$, is the smallest integer $n \geqq 0$ such that the null spaces of Q^n and Q^{n+1} coincide. *The* algebraic *eigenspace* of P at λ, $N_\lambda(P)$, is the null space of Q^ν where $\nu \equiv \nu_\lambda(P)$. The next lemma is well known (e.g., Rothblum–Veinott (1976)). It is stated for completeness.

LEMMA 2.1. *Let Q be an $S \times S$ complex matrix, ν be the index of zero for Q, $N \equiv N_0(Q)$ and R be the range of Q^ν. Then:*

(a) $C^S = R \oplus N$.

(b) *There is a unique projection E on N along R and so $Q^\nu E = EQ^\nu = 0$.*

* Received by the editors August 25, 1975, and in revised form April 2, 1976.

† School of Organization and Management, Yale University, New Haven, Connecticut 06520.

Parts of the research reported in this paper are based on the author's doctoral thesis submitted to the department of Operations Research at Stanford University. Research at Stanford University was supported by the National Science Foundation under Grant GK-18339 and by the Office of Naval Research under Contract N00014-67-A-0112-0050. Further research at the Courant Institute of Mathematical Sciences of New York University was supported by the National Science Foundation under Grant GP-37069.

(c) *E and Q commute.*

(d) *$E-Q$ is nonsingular.*

Proof. The proof of (a) is given in Halmos (1958, p. 113) and the proofs of (b) and (c) appear in Kato (1966, pp. 20, 21, 23). It remains only to prove (d). If $\nu = 0$, then $E = 0$ and $E-Q$ is nonsingular. Suppose $\nu > 0$ and $(E-Q)\mu = 0$. Thus, $0 = \sum_{k=0}^{\nu-1} Q^k E(E-Q)\mu = E\mu = Q\mu$. Thus $\mu \in N$, so $\mu = E\mu = 0$. Hence $E-Q$ is nonsingular, completing the proof of Lemma 2.1.

For a given square matrix P and complex number λ, the projection constructed in Lemma 2.1 for $Q \equiv P - \lambda I$ is called the *eigenprojection* of P at λ and is denoted $E_\lambda(P)$. The *deviation matrix* of P at λ is defined by $D_\lambda(P) \equiv (E-Q)^{-1}(I-E)$, where $E \equiv E_\lambda(P)$. Also let $\|P\| \equiv \max_{1 \leq i \leq S} \sum_{j=1}^{S} |P_{ij}|$ and let $\sigma(P)$ be the *spectral radius* of P.

Finally, for a finite set J, let $|J|$ be the number of elements in J and for a vector $x \in R^S$ let $\|x\| = \max_{1 \leq i \leq S} |x_i|$.

3. Description of the model. Consider a system that is observed at each of a sequence of points in time labeled $1, 2, \cdots$. At each of those points the system is found in one of S *states*, labeled $1, \cdots, S$. Each time the system is observed in state s, an *action* a is chosen from a finite set A_s of possible actions and a *reward* $r(s, a)$ is received. The *transition rate* of the system into state t at time $N+1$, given that it is found in state s at time N and that action a is taken at that time, and given the observed states and actions taken at times $1, \cdots, N-1$, is assumed to be a nonnegative function $p(t|s, a)$, depending only on t, s and a. These transition rates are not necessarily probabilities; i.e., we do *not* necessarily assume that $\sum_{t=1}^{S} p(t|s, a) = 1$. There are many interpretations of this generalized model, which for brevity we do not mention here.

Let $\Delta \equiv \prod_{s=1}^{S} A_s$ be the set of all *decision rules*, i.e., of all functions δ mapping each state s into an action $\delta^s \in A_s$. A *policy* is a sequence $\pi = (\delta_1, \delta_2, \cdots)$ of decision rules, often written for brevity $\pi = (\delta_N)$. Sometimes, to avoid double subscripts, we shall use $\delta(N)$ interchangeably with δ_N without mentioning that fact. The set of all policies will be denoted Δ^∞. We write δ^∞ for the *stationary policy* $(\delta, \delta, \cdots)$. If π is a policy, let π^N denote the first N components of π.

For each decision rule δ, let r_δ be the S element column vector of one period rewards earned by δ. Thus, the sth component of the r_δ is $r(s, \delta^s)$. Similarly, let P_δ be the $S \times S$ matrix of one step transition when δ is used. The stth component of P_δ is $p(t|s, \delta^s)$. If $\pi = (\delta_N)$ is a policy, let $P_\pi^N \equiv P_{\delta(1)} \cdots P_{\delta(N)}$ be the N step transition matrix resulting from the use of π. In particular, $P_\pi^0 \equiv I$, and if $\pi = \delta^\infty$, $P_\pi^N = (P_\delta)^N \equiv P_\delta^N$.

Let V_π^N be the S vector whose sth coordinate $(V_\pi^N)_s$ is the N period reward when using the policy $\pi = (\delta_N)$ and starting from state s. Evidently, for $\pi = (\delta_i)$,

$$V_\pi^N = \sum_{i=1}^{N} P_\pi^{i-1} r_{\delta(i)}, \qquad N = 1, 2, \cdots.$$

Suppose there is an interest rate $0 < \rho < \infty$. We suppress the dependence of the discount factor $\beta \equiv (1+\rho)^{-1}$ on ρ in the sequel for simplicity. We say that a policy π is *normalized* if for every $0 < \beta < 1$, $\sum_{N=0}^{\infty} \beta^N P_\pi^N$ converges. We say that a

decision rule δ is *normalized* if that is so of δ^∞. If $\pi = (\delta_N)$ is normalized, then the S vector $V_{\rho\pi}$ of expected total discounted returns starting from each state using policy $\pi = (\delta_N)$, which is given by

$$V_{\rho\pi} = \sum_{N=1}^{\infty} \beta^N P_\pi^{N-1} r_{\delta(N)},$$

converges absolutely for $0 < \beta < 1$. Policy π is *transient* if the above sequences converge for $\beta = 1$. For each policy π, let $V_\pi^0 = 0$. If $\pi = \delta^\infty$, we suppress the ∞ and work $V_\delta^N = V_\pi^N$ and $V_{\rho\delta} = V_{\rho\pi}$.

Using a characterization of Veinott (1969), the normalization condition was characterized in Rothblum (1975a) as follows:

LEMMA 3.1. *The following four statements are equivalent*:

(1) *Every (resp., some) stationary policy is normalized.*

(2) *Every (resp., some) policy is normalized.*

(3) $\sigma(P_\delta) \leqq 1$ *for every (resp., some) decision rule* δ.

(4) *For every* $N \geqq 1$, $\sigma(P_\pi^N) \leqq 1$ *for every (resp., some) policy* π.

An example in which every policy is normalized, but is not necessarily substochastic was illustrated in Rothblum (1975a).

In the remainder of this paper, we assume without further mention that we are in the *normalized case*, i.e., $\sigma_\delta \equiv \sigma(P_\delta) \leqq 1$ for all $\delta \in \Delta$. For each integer n, a policy τ is called *n discount optimal* if

$$\liminf_{\rho \downarrow 0} \rho^{-n}(V_{\rho\tau} - V_{\rho\pi}) \geqq 0 \quad \text{for all } \pi. \tag{3.1}$$

The limit inferior is, of course, componentwise. Similarly, we say that τ is ∞ *discount optimal*, or *sensitive discount optimal* if, for some $\rho^* > 0$,

$$V_{\rho\tau} - V_{\rho\pi} > 0 \quad \text{for all } \pi \text{ and } 0 < \rho < \rho^*. \tag{3.2}$$

For $-\infty < n \leqq \infty$ let Δ_n be the set of all δ in Δ for which δ^∞ is n discount optimal. Below we shall characterize these sets.

Following results from Miller–Veinott (1969) and Veinott (1969) (for the stochastic case), the Laurent expansion of the vector of expected return corresponding to a stationary policy (in the normalized case) was obtained in Rothblum (1975a, Thm. 3.1) as follows: For $\rho > 0$ sufficiently small

$$V_{\rho\delta} = \sum_{n=-\nu_\delta}^{\infty} \rho^n v_\delta^n, \tag{3.3}$$

where

$$v_\delta^n = \begin{cases} Q_\delta^{-n-1} E_\delta r_\delta & \text{if } n = -1, -2, \cdots, \\ (-1)^n D_\delta^{n+1} r_\delta & \text{if } n = 0, 1, \cdots, \end{cases}$$

$Q_\delta \equiv P_\delta - I$, E_δ and D_δ are the eigenprojection and deviation matrix of P_δ at one and ν_δ is the index of one for P_δ. Observe that $V_\delta^n = 0$ for integers $n < -\nu_\delta$.

The above expansion enables one to get a characterization of the Δ_n's. Before doing this we need a few additional definitions. Let C be a real matrix. We say that C is *lexicographically nonnegative*, written $C \geqq 0$, if the first nonvanishing element of each row of C is positive. We say that C is *lexicographically semipositive*, written $C > 0$, if $C \geqq 0$ and $C \neq 0$. Write $C \geqq$ (resp., $>$) B or $B \leqq$ (resp., $<$) C if $C - B \geqq$ (resp., $>$) 0. These definitions apply to infinite matrices as long as they are not vacuous. Let $\nu \equiv \max\{\nu_\delta \mid \delta \in \Delta\}$. For $\delta \in \Delta$ and $n = -\nu, -\nu+1, \cdots$, let $V_\delta^n \equiv (v_\delta^{-\nu}, \cdots, v_\delta^n)$ and $V_\delta^\infty \equiv (v_\delta^{-\nu}, v_\delta^{-\nu+1}, \cdots)$. Similar to results in Miller–Veinott (1969), Veinott (1969) and Rothblum (1975a), it follows from (3.3) that for all integers $n \geqq -\nu$,

$$\Delta_n = \{\delta \in \Delta \mid V_\delta^n \geqq V_\gamma^n \quad \text{for all } \gamma \in \Delta\} \tag{3.4}$$

and for integers $n < -\nu$, $\Delta_n = \Delta$.

For $\gamma, \delta \in \Delta$ and $n = -\nu, -\nu+1, \cdots$ let $C_{\gamma\delta}^n \equiv (c_{\gamma\delta}^{-\nu}, c_{\gamma\delta}^{-\nu+1}, \cdots, c_{\gamma\delta}^n)$ where $c_{\gamma\delta}^j \equiv r_\gamma^j + Q_\gamma v_\delta^j - v_\delta^{j-1}$ for $j = \cdots, -1, 0, 1, \cdots$, $r_\gamma^j = 0$ for $j \neq 0$ and $r_\gamma^0 = r_\gamma$. Also put $C_{\gamma\delta}^\infty \equiv (c_{\gamma\delta}^{-\nu}, c_{\gamma\delta}^{-\nu+1}, \cdots)$. Obviously $c_{\gamma\delta}^j = 0$ for all integers $j < -\nu$. Using a policy improvement algorithm, it was shown in Rothblum (1975a, § 4) that there exists a stationary ∞ discount optimal policy δ^∞ for which $C_{\gamma\delta}^\infty \leqq 0$ for every $\gamma \in \Delta$. It follows from (3.4) that $V_{(\cdot)}^n$ coincides for all decision rules in Δ_n, and therefore for every $\gamma \in \Delta$ and $-\nu \leqq n \leqq \infty$, $C_{\gamma(\cdot)}^n$ coincides on Δ_n. Thus

$$\Delta_n \subseteq \{\delta \in \Delta \mid C_{\gamma\delta}^n \leqq 0 \text{ for every } \gamma \in \Delta\}. \tag{3.5}$$

4. Uniform polynomial bound on the periodical reward. The purpose of this section is to give a uniform polynomial bound on the reward gained in the Nth period. We show that if $N^{-m}P_\delta^N$ is bounded for every $\delta \in \Delta$, then $N^{-m}P_\pi^N$ is uniformly bounded in $\pi \in \Delta^\infty$. We then characterize the least integer $m \geqq -1$ for which the above holds to be $\nu - 1$.

For the purpose of this section we introduce a few additional notations. For an $S \times S$ real matrix P and $I, J \subseteq \{1, \cdots, S\}$, denote by P_{IJ} the submatrix of P whose rows and columns correspond to I and J, and let $P_J \equiv P_{JJ}$. For $x \in R^S$ and $J \subseteq \{1, \cdots, S\}$, let x_J be the corresponding subvector of x.

We next summarize a few definitions from the theory of nonnegative matrices. Let P be an $S \times S$ nonnegative matrix. We say that states i and j *communicate* if there exist nonnegative integers n and m such that $(P^n)_{ij} > 0$ and $(P^m)_{ij} > 0$. This communication relation is an equivalence relation; hence one can partition the totality of states into equivalence classes. A class J of P is called *basic* if $\sigma(P_J) = \sigma(P)$. Spectral properties of nonnegative matrices that will be used in this section are discussed in Rothblum (1975b).

THEOREM 4.1. *Let m be a nonnegative integer. Then the following five conditions are equivalent*:

(1) $m \geqq \nu \equiv \max_{\delta \in \Delta} \nu_\delta$.

(2) $N^{-m+1}P_\delta^N$ *is uniformly bounded in* (N, δ), $(N \neq 0)$.

(3) $N^{-m+1}P_\pi^N$ *is uniformly bounded in* (N, π), $(N \neq 0)$.

(4) $N^{-m}P_\delta^N \to 0$ *as* $N \to \infty$ *for every* $\delta \in \Delta$.

(5) $N^{-m}P_\pi^N \to 0$ *as* $N \to \infty$ *for every* $\pi \in \Delta^\infty$.

Proof. The implications (3)$\Rightarrow$(2)$\Rightarrow$(4) and (3)$\Rightarrow$(5)$\Rightarrow$(4) are obvious. We next prove that (4)$\Rightarrow$(1). Let P be a square matrix where $\lim_{N\to\infty} N^{-m}P^{-N} = 0$.

We shall show that $\nu_1(P) \leqq m$. Assume that $(P-I)^{m+1}x = 0$ for some vector x. By the binomial formula,

$$P^N x = \sum_{i=0}^{N} \binom{N}{i} (P-I)^i x \quad \text{for } N = 0, 1, \cdots.$$

By premultiplying this equation by N^{-m}, letting $N \to \infty$ and observing that $\lim_{N\to\infty} N^{-m}\binom{N}{i} = 0$ for $i = 0, \cdots, m-1$, one can verify that

$$0 = \lim_{N\to\infty} N^{-m} P^N x = \lim_{N\to\infty} N^{-m}\binom{N}{m}(P-I)^m x = (m!)^{-1}(P-I)^m x.$$

We see that $(P-I)^m x = 0$ whenever $(P-I)^{m+1}x = 0$, proving that $\nu_1(P) \leqq m$ and therefore showing that (4) $\Rightarrow$ (1). It remains to show that (1) $\Rightarrow$ (3).

Assume first that $m = 0$. It follows from (1) that $\nu_\delta = 0$ for all $\delta \in \Delta$, i.e., one is not an eigenvalue of P_δ for any $\delta \in \Delta$. The Perron–Frobenius theorem (e.g., Varga (1962, p. 46)) implies that σ_δ is an eigenvalue of P_δ and so $\sigma_\delta < 1$ for all $\delta \in \Delta$. It follows from Veinott (1969, p. 1638) that there exists a diagonal matrix B having positive diagonal elements, and $\|BP_\delta B^{-1}\| < 1$ for all $\delta \in \Delta$. By the finiteness of Δ, $\alpha \equiv \max_{\delta\in\Delta} \|BP_\delta B^{-1}\| < 1$. Hence for every policy π, $\|NP_\pi^N\| < N\|B^{-1}\|\,\|B\|\alpha^N$, which implies (3) when $m = 0$.

Assume now that $m > 0$. Every $\delta \in \Delta$ has a characteristic number which equals the maximum number of positive coordinates in vectors which belong to $N_\delta \equiv N_1(P_\delta)$. Let γ be a decision rule which maximizes this number among all decision rules in Δ. By Theorem 3.1 of Rothblum (1975b) there exists a vector x in N_γ having the largest set of positive coordinates among all vectors in N_γ and $Q_\gamma^j x \geqq 0$ for all $j = 0, 1, \cdots$.

For the remainder of this proof let $K(v)$ denote the set of indices of the positive coordinates of the vector v. It follows from the definitions of γ and x that with $K \equiv K(x)$,

$$|K| \geqq |K(z)| \quad \text{for all } z \in \bigcup_{\delta\in\Delta} N_\delta. \tag{4.1}$$

By possibly reindexing the states, we may assume that $K = \{1, \cdots, |K|\}$.

Now set $r_\delta = x$ for every $\delta \in \Delta$. It follows from Rothblum (1975a, Thm. 4.1) that there exists a stationary ∞ discount optimal policy μ^∞ and $V_\delta^{-1} \leqq V_\mu^{-1}$ and $C_{\delta\mu}^{-1} \leqq 0$ for all $\delta \in \Delta$. By the finiteness of Δ, there exists a positive integer M such that for all $\delta \in \Delta$ and $k = 0, 1, \cdots$,

$$t^k \equiv \sum_{j=-\nu}^{-1} \binom{M}{-j} v_\mu^{j-k} \geqq \sum_{j=-\nu}^{-1} \binom{M}{-j} v_\delta^{j-k} \tag{4.2}$$

and

$$\sum_{j=-\nu}^{-1} \binom{M}{-j} c_{\delta\mu}^{j-k} \leqq 0. \tag{4.3}$$

Observe that $t^k = 0$ for $k \geqq \nu$. Since $x \in N_\gamma$, it follows that for $n = -1, -2, \cdots$,

$$v_\gamma^n = Q_\gamma^{-n-1} E_\gamma x = Q_\gamma^{-n-1} x \geqq 0. \tag{4.4}$$

For $n=-1,-2,\cdots$, $v_\mu^n = Q_\mu^{-n-1}E_\mu x$. The commutativity of E_μ and P_μ implies that $v_\mu^n \in N_\mu$. Hence, $t^k \in N_\mu$ for $k \geqq 0$ and by (4.2) and (4.4),

$$0 \leqq \sum_{j=-\nu}^{-1} \binom{M}{-j} v_\gamma^{j-k} \leqq t^k.$$

Setting $k=0$ in the above inequality and recalling the nonnegativity of the v_γ^n's, we find that $K(t^0) \supseteq K(v_\gamma^{-1}) = K(x) = K$. Moreover, since $t^0 \in N_\mu$, it follows from (4.1) that $K(t^0) = K$. In particular by (4.1) and the fact that $t^k \in N_\mu$,

$$K(t^k) \subseteq K(t^0) = K(x) \quad \text{for } k = 0, 1, \cdots. \tag{4.5}$$

It follows from (4.3) and the definition of the $c_{(\cdot)\mu}^n$ that for every $\delta \in \Delta$ and $k=0,1,\cdots$,

$$0 \geqq \sum_{j=-\nu}^{-1} \binom{M}{-j} c_{\delta\mu}^{j-k} = Q_\delta t^k - t^{k+1}. \tag{4.6}$$

We shall now show that for every positive integer N,

$$P_\pi^N t^0 \leqq \sum_{k=0}^{\nu-1} \binom{N}{k} t^k \quad \text{for every policy } \pi. \tag{4.7}$$

The proof is by induction on N. Obviously (4.6) implies (4.7) for the case $N=1$. Suppose now (4.7) holds for the positive integer $N-1$ and consider N. Let $\pi = (\delta_i)$ be a given policy. By the induction hypothesis, the nonnegativity of $P_{\delta(1)}$, (4.6) and the fact that $t^k = 0$ for all $k \geqq \nu$, it follows that

$$\begin{aligned} P_\pi^N t^0 = P_{\delta(1)} \cdots P_{\delta(N)} t^0 &\leqq P_{\delta(1)} \sum_{k=0}^{\nu-1} \binom{N-1}{k} t^k \\ &\leqq \sum_{k=0}^{\nu-1} \binom{N-1}{k} (t^k + t^{k+1}) = \sum_{k=0}^{\nu-1} \binom{N}{k} t^k. \end{aligned} \tag{4.8}$$

Now, from (4.7),

$$N^{-\nu+1} P_\pi^N t^0 \leqq \sum_{k=0}^{\nu-1} N^{-\nu+1} \binom{N}{k} t^k \to ((\nu-1)!)^{-1} t^{\nu-1} \quad \text{as } N \to \infty. \tag{4.9}$$

Also, $(t^0)_K \gg 0$ and $(t^0)_L = 0$ where $L = \{1, \cdots, S\} \setminus K$. Thus by (4.9), $(N^{-m+1} P_\pi^N)_K \leqq (N^{-\nu+1} P_\pi^N)_K$ is uniformly bounded.

It follows from (4.5), the nonnegativity of the t^k's and (4.7) that for every policy π and $N=0,1,\cdots$,

$$0 = \sum_{k=0}^{\nu-1} \binom{N}{k} (t^k)_L \geqq (P_\pi^N t^0)_L = (P_\pi^N)_{LK} (t^0)_K.$$

Since $(t^0)_K \gg 0$, $(P_\pi^N)_{LK} = 0$ and so $(P_\delta)_{LK} = 0$ for all $\delta \in \Delta$. Thus the model is decomposable in the sense that if $i \in L$ and $j \in K$, then there is no policy π and integer N such that $(P_\pi^N)_{ij} > 0$. This implies that for every $\delta \in \Delta$, L is a union of classes of P_δ.

We shall next show that $\sigma((P_\delta)_L) < 1$ for every $\delta \in \Delta$. The normality condition implies that $\sigma((P_\delta)_L) \leqq 1$. Assume now that for some $\delta \in \Delta$, $\sigma((P_\delta)_L) = 1$; then L

contains a basic class J of P_δ. Consider the decision rule θ, defined by

$$\theta^s = \begin{cases} \mu^s, & s \in K, \\ \delta^s, & s \in L. \end{cases}$$

Since $t^0 \in N_\mu$, $Q_\theta^\nu t^0 = Q_\mu^\nu t^0 = 0$, so $t^0 \in N_\theta$. Next, observe that J is a basic class of P_θ. Hence there exists a semipositive vector z in N_θ satisfying $z_J \gg 0$ by Rothblum (1975b, Thm. 3.1). Obviously $z + t^0 \in N_\theta$ and $K(z+t^0) \supset K(t^0) \cup J$, so by (4.5), (4.1) cannot hold, which is a contradiction. Thus $\sigma((P_\delta)_L) < 1$ for all $\delta \in \Delta$.

Since $(P_\delta)_{LK} = 0$ for every $\delta \in \Delta$, it follows that for every policy $\pi = (\delta_i)$, $(P_\pi^N)_L = (P_{\delta(1)})_L \cdots (P_{\delta(N)})_L$. Restricting attention to the states in L, and recalling that $\sigma((P_\delta)_L) < 1$ for every $\delta \in \Delta$, we see that it follows from Veinott (1969, p. 1639) that we are simply in the transient case on L, i.e., $\sum_{N=0}^\infty (P_\pi^N)_L$ converges for every policy π, and for some N, $\|(P_\pi^N)_L\| < 1$ for every π. Hence, $(P_\pi^N)_L$ is uniformly bounded in N and in π. Recalling that $m \geqq 1$, we see that the latter obviously implies the uniform boundedness of $N^{-m+1}(P_\pi^N)_L$.

We shall finally show that $N^{-m+1}(P_\pi^N)_{KL}$ is uniformly bounded. Let $\pi = (\delta_i)$. For the purpose of this proof define ${}^iP_\pi^N = P_{\delta(i+1)} \cdots P_{\delta(N)}$. Since $(P_\delta)_{LK} = 0$ for every $\delta \in \Delta$, it follows that

$$(P_\pi^N)_{KL} = \sum_{i=1}^N (P_\pi^{i-1})_K (P_{\delta(i)})_{KL} ({}^iP_\pi^N)_L. \tag{4.10}$$

Since $\sigma((P_\delta)_L) < 1$ for all $\delta \in \Delta$, there is a diagonal matrix B having positive diagonal elements such that $\alpha \equiv \max_{\delta \in \Delta} \|B(P_\delta)_L B^{-1}\| < 1$ (see Veinott (1969, p. 1638)). For every policy π and $N = 0, 1, \cdots$, $\|(P_\pi^N)_L\| \leqq \|B^{-1}\| \|B\| \alpha^N$. Let $k \equiv \|B^{-1}\| \|B\| \max_{\delta \in \Delta} \|P_\delta\|$. Then by (4.10),

$$\|N^{-m+1}(P_\pi^N)_{KL}\| \leqq k \sum_{i=1}^N N^{-m+1} \|(P_\pi^{i-1})_K\| \alpha^{N-i}.$$

Since we have already shown $N^{-\nu+1}(P_\pi^N)_K$ is uniformly bounded in π and N, it follows that the question of the uniform boundedness of $N^{-m+1}(P_\pi^N)_{KL}$ for $m \geqq \nu$ is reduced to the boundedness of the sequence $a_N \equiv \sum_{i=1}^N ((i-1)/N)^{\nu-1} \alpha^{N-i}$. Recalling that $m \geqq 1$, we see that

$$0 \leqq \sum_{i=1}^N \left(\frac{i-1}{N}\right)^{m-1} \alpha^{N-i} \leqq \sum_{i=1}^N \alpha^{N-i} < (1-\alpha)^{-1},$$

which shows the boundedness of a_N. This completes the proof of Theorem 4.1.

COROLLARY 4.2. *For any $\rho^* > 0$, $\rho^\nu V_{\rho\pi}$ is uniformly bounded in $0 < \rho < \rho^*$ and $\pi \in \Delta^\infty$.*

Proof. Let B be a uniform bound of $(N+1)^{-\nu+1} P_\pi^N$, (the existence of such a bound follows from Theorem 4.1) and let $r \equiv \max\{|r(s,a)| \mid 1 \leqq s \leqq S \text{ and } a \in A_s\}$. For any policy $\pi = (\gamma_i)$,

$$\|\rho^\nu V_{\rho\pi}\| = \left\| \rho^\nu \sum_{i=1}^\infty \beta^i P_\pi^{i-1} r_{\gamma(i)} \right\|$$

$$\leqq \rho^\nu \sum_{i=0}^\infty \beta^i i^{\nu-1} \cdot B \cdot r$$

The uniform boundedness of $\rho^{\nu}V_{\rho\pi}$ now follows directly from the boundedness of $\rho^{\nu}\sum_{i=0}^{\infty}\beta^{i}i^{\nu-1}$.

COROLLARY 4.3. *Every policy is n discount optimal for integers* $n<-\nu$.

5. Characterization of *n* discount optimal policies. The purpose of this section is to develop necessary and sufficient conditions for n discount optimality. These refine previous results obtained for the stochastic case by Sladky (1974). Our conditions coordinate Sladky's approach with ideas used by Lippman (1968).

We have shown (Corollary 4.3) that every policy is n discount optimal for integers $n<-\nu$. Thus, n discount optimality is uninteresting in this case. To this end we consider n discount optimality only for integers $n\geqq-\nu$.

We start by stating a necessary condition and a (different) sufficient condition for n discount optimality of stationary policies. These follow from results obtained in Rothblum (1975a) by using the methods of Miller–Veinott (1969) and Veinott (1969), (1975).

THEOREM 5.1. *Let* $n=-\nu, -\nu+1, \cdots$ *and* δ^{∞} *be a stationary n discount optimal policy. Then*

$$\{\gamma\in\Delta\,|\,C_{\gamma\delta}^{n+\nu}=0\}\subseteq\Delta_n\subseteq\{\gamma\in\Delta\,|\,C_{\gamma\delta}^{n}=0\}.$$

Proof. Let $\gamma\in\Delta_n$. It follows from (3.4) that $V_{(\cdot)}^{n}$ coincides for all decision rules in Δ_n and therefore so does $C_{\gamma(\cdot)}^{n}$. Observing the fact (see Rothblum (1975a, § 4)) that $C_{\gamma\gamma}^{\infty}=0$, we find that $C_{\gamma\delta}^{n}=C_{\gamma\gamma}^{n}=0$, completing the proof of the second inclusion. The first inclusion follows directly from the proof of Theorem 5.1 of Rothblum (1975a).

We next introduce a few additional definitions. Let $n=-\nu, -\nu+1, \cdots$ and let $\delta\in\Delta_n$. Define

$$\Delta_n^{*}=\{\gamma\in\Delta\,|\,C_{\gamma\delta}^{n}=0\}.$$

Observe that by (3.4) these definitions are independent of $\delta\in\Delta_n$. Theorem 5.1 says that for $n=-\nu, -\nu+1, \cdots$ $\Delta_n^{*}\supseteq\Delta_n\supseteq\Delta_{n+\nu}^{*}$.

It is clear that if the N period transition into a given state is zero, then the action taken in this state at period N is irrelevant. We next formulate this idea rigorously. For $N=0, 1, \cdots, n=-\nu, -\nu+1, \cdots$ and $\pi\in\Delta^{\infty}$, let

$$\mathscr{S}(N,\pi)\equiv\left\{t\,\middle|\,\sum_{s=1}^{S}(P_{\pi}^{N})_{st}>0\right\},$$

$$\Delta_n^{*}(N,\pi)=\{\gamma\in\Delta\,|\,(c_{\gamma\delta}^{j})_t=0 \text{ for } \delta\in\Delta_n,\ t\in\mathscr{S}(N,\pi), \text{ and } j=-\nu,\cdots,n\}$$

and

$$\Delta_{\infty}^{*}(N,\pi)=\{\gamma\in\Delta\,|\,(c_{\gamma\delta}^{j})_t=0 \text{ for } \delta\in\Delta_{\infty},\ t\in\mathscr{S}(N,\pi), \text{ and } j=-\nu,-\nu+1,\cdots\}.$$

We remark that the definition of the $\Delta_n^{*}(N,\pi)$'s do not depend on δ and that these sets are product spaces, i.e., it is of the form $\prod_{s=1}^{S}A_s^{*}$, where $A_s^{*}\subseteq A_s$ for $s=1,\cdots,S$.

We are now ready to give the first necessary and sufficient conditions for n discount optimality.

THEOREM 5.2. *Let $n=-\nu, -\nu+1, \cdots$ and let δ^∞ be a stationary n discount optimal policy. A policy $\pi=(\gamma_i)$ is n discount optimal if and only if*

$$\gamma_{N+1}\in\Delta_n^*(N,\pi) \quad \text{for } N=0,1,\cdots, \tag{5.1}$$

and

$$\begin{aligned}&\lim_{\rho\downarrow 0}\left(\rho^{-n}V_{\rho\pi}-\rho\sum_{N=0}^{\infty}\beta^N P_\pi^N v_\delta^n\right)=0 \quad \text{if } n=-\nu,\cdots,-1,\\ &\lim_{\rho\downarrow 0}\rho\sum_{N=0}^{\infty}\beta^N P_\pi^N v_\delta^n=0 \quad \text{if } n=0,1,\cdots.\end{aligned} \tag{5.2}$$

Proof. We first show that (5.1) is equivalent to

$$P_\pi^N c^j_{\gamma(N+1),\delta}=0 \text{ for } j=-\nu,\cdots,n \text{ and } N=0,1,\cdots. \tag{5.3}$$

Obviously (5.1)$\Rightarrow$(5.3); thus it suffices to show the reverse. Let $\pi=(\gamma_i)$ satisfy (5.3). Since $\delta\in\Delta_n$ it follows from (3.5) that for every fixed $N=0,1,\cdots$, $C^n_{\gamma(N+1),\delta}\leqq 0$. This implies that for some integer L,

$$\sum_{j=-\nu}^{n}\binom{M}{n-j}c^j_{\gamma(N+1),\delta}\leqq 0 \quad \text{for } M=L, L+1,\cdots. \tag{5.4}$$

Premultiplying this inequality by P_π^N and substituting (5.3) shows that

$$P_\pi^N\sum_{j=-\nu}^{n}\binom{M}{n-j}c^j_{\gamma(N+1),\delta}=\sum_{j=-\nu}^{n}\binom{M}{n-j}P_\pi^N c^j_{\gamma(N+1),\delta}=0.$$

The nonnegativity of P_π^N and (5.4) imply that $(\sum_{j=-\nu}^n\binom{M}{n-j}c^j_{\gamma(N+1),\delta})_t=0$ for all $t\in\mathcal{S}(N,\pi)$. Since M can be chosen arbitrarily large it immediately follows that $(c^j_{\gamma(N+1),\delta})_t=0$ for $j=-\nu,\cdots,n$ and $t\in\mathcal{S}(N,\pi)$, completing the proof of the equivalence of (5.1) and (5.3).

Next observe that since δ^∞ is n discount optimal, it follows that a policy π is n discount optimal if and only if

$$\lim_{\rho\downarrow 0}\rho^{-n}(V_{\rho\pi}-V_{\rho\delta})=0. \tag{5.5}$$

For $i=0,1,\cdots$ let $\tau(i)=(\pi^i,\delta^\infty)$. Then for $\rho>0$ sufficiently small

$$\begin{aligned}V_{\rho\pi}-V_{\rho\delta}&=\sum_{i=0}^{\infty}(V_{\rho,\tau(i+1)}-V_{\rho,\tau(i)})\\ &=\sum_{i=0}^{\infty}\beta^i P_\pi^i(\beta r_{\gamma(i+1)}+\beta P_{\gamma(i+1)}V_{\rho\delta}-V_{\rho\delta})\\ &=\sum_{i=0}^{\infty}\beta^{i+1}P_\pi^i\sum_{j=-\nu}^{\infty}\rho^j c^j_{\gamma(i+1),\delta}.\end{aligned} \tag{5.6}$$

We shall next approximate the above expression. Observing that there exists a positive constant K such that $\|c^j_{\gamma\delta}\| \leqq K^{j+1}$ for all $j = 0, 1, \cdots$, and all $\gamma \in \Delta$, we deduce from Corollary 4.2 that for $\rho < K^{-1}$,

$$\begin{aligned} &\left\| \sum_{i=0}^{\infty} \beta^i P^i_\pi \sum_{j=n+\nu+1}^{\infty} \rho^j c^j_{\gamma(i+1),\delta} \right\| \\ &\leqq \left\| \sum_{i=0}^{\infty} \beta^i P^i_\pi \right\| (\rho K)^{n+\nu+1} K (1-\rho K)^{-1} < \rho^{n+1} B, \end{aligned} \tag{5.7}$$

where B is a constant. Premultiplying (5.6) by ρ^{-n} and using the approximation given in (5.7) implies that (5.5) is equivalent to

$$\lim_{\rho \downarrow 0} \sum_{i=0}^{\infty} \beta^i \sum_{j=-\nu}^{n+\nu} \rho^{j-n} P^i_\pi c^j_{\gamma(i+1),\delta} = 0. \tag{5.8}$$

We are now ready to show the necessity of (5.3). Let π be n discount optimal, or equivalently, satisfy (5.8). It follows from (3.5) that for all $\gamma \in \Delta$, $C^n_{\gamma\delta} \leqq 0$. Thus, for ρ sufficiently small $\sum_{j=-\nu}^{n+\nu} \rho^j c^j_{\gamma\delta} \leqq 0$ for all $\gamma \in \Delta$. By combining this fact, the nonnegativity of the transition matrices and (5.8), we conclude that for $N = 0, 1, \cdots$.

$$\lim_{\rho \downarrow 0} \sum_{j=-\nu}^{n+\nu} \rho^{j-n} P^N_\pi c^j_{\gamma(N+1),\delta} = 0,$$

It is easily seen that (5.3) is a necessary condition for the above.

In order to show the conclusion of our theorem it suffices to show that a policy π which satisfies (5.3) is n discount optimal (or equivalently satisfies (5.8)) if and only if it satisfies (5.2). Thus it suffices to show that a given policy $\pi = (\gamma_i)$ satisfies

$$\lim_{\rho \downarrow 0} \sum_{j=n+1}^{n+\nu} \rho^{j-n} \sum_{i=0}^{\infty} \beta^i P^i_\pi c^i_{\gamma(i+1),\delta} = 0 \tag{5.9}$$

if and only if it satisfies (5.2). If $\nu = 0$, then by Corollary 4.2, $R_\rho \equiv \sum_{i=0}^{\infty} \beta^{i+1} P^i_\pi$ is bounded, and both (5.9) and (5.2) are always satisfied. If $\nu > 0$, let $V^j_{\rho\pi} \equiv 0$ if $j \neq 0$ and $V^0_{\rho\pi} \equiv V_{\rho\pi}$, i.e., $V^j_{\rho\pi} = \sum_{i=0}^{\infty} \beta^{i+1} P^i_\pi r^j_{\gamma(i+1)}$. Recalling the definitions of the c's, one gets that

$$\begin{aligned} &\sum_{j=n+1}^{n+\nu} \rho^{j-n} \sum_{i=0}^{\infty} \beta^{i+1} P^i_\pi c^j_{\gamma(i+1),\delta} \\ &= \sum_{j=n+1}^{n+\nu} \rho^{j-n} \sum_{i=0}^{\infty} \beta^{i+1} P^i_\pi (r^j_{\gamma(i+1)} + P_{\gamma(i+1)} v^j_\delta - v^j_\delta - v^{j-1}_\delta) \\ &= \sum_{j=n+1}^{n+\nu} \rho^{j-n} (V^j_{\rho\pi} + (1+\rho)(R_\rho - \beta I) v^j_\delta - R_\rho v^j_\delta - R_\rho v^{j-1}_\delta) \\ &= \sum_{j=n+1}^{n+\nu} \rho^{j-n} (V^j_{\rho\pi} + \rho R_\rho v^j_\delta - R_\rho v^{j-1}_\delta - v^j_\delta) \\ &= \rho^{-n} V^{\max\{(n+1,0\}}_{\rho\pi} + \rho^{\nu+1} R_\rho v^{n+\nu}_\delta - \rho R_\rho v^n_\delta - \sum_{j=n+1}^{n+\nu} \rho^{j-n} v^j_\delta. \end{aligned}$$

The equivalence of (5.2) and (5.9) now follows immediately from the fact that $\rho^{\nu}R_{\rho}$ is bounded in ρ, thus completing the proof of Theorem 5.2.

An immediate corollary of Theorem 5.2 gives a necessary and sufficient condition for ∞ discount optimality.

COROLLARY 5.3. *Let δ^{∞} be a stationary ∞ discount optimal policy. A policy $\pi=(\gamma_i)$ is ∞ discount optimal if and only if*

$$\gamma_{N+1}\in\Delta_{\infty}^{*}(N,\pi) \quad \text{for } N=0,1,\cdots.$$

Proof. The necessity follows immediately from Theorem 5.2 and the fact that an ∞ discount optimal policy is n discount optimal for all integers $n=-\nu,-\nu+1,\cdots$. To prove the sufficiency observe that (5.1) implies (5.3). Thus by (5.6), one gets that $V_{\rho\pi}=V_{\rho\delta}$; completing the proof of Corollary 5.3.

We shall next compare our results with those obtained by Sladky (1974, Thm. 2.2). Sladky considered only the stochastic case, in which it is known that $\nu=1$. His necessary and sufficient conditions for n discount optimality are (5.3) and (5.9). Observe that (5.3) (in the stochastic case) means that for every $N=0,1,\cdots$, the expected value of the comparison functions $c^{j}_{\gamma(N+1),\delta}$, $j=-\nu,\cdots,n$, is zero, whereas (5.1) says that $c^{j}_{\gamma(N+1),\delta}=0$ with probability one, $j=-\nu,\cdots,n$. Also observe that (5.9) involves v^{j}_{δ} for integers $j>n$, which are irrelevant for n discount optimality (see (3.4)). We remark that Lippman (1968) gave a necessary and sufficient condition for 0 discount optimality in the transient case (i.e., $\nu=0$) which is precisely (5.1) (Lippman considered the case where all matrices are stochastic and there is a fixed interest rate; however, his methods apply to the general transient case.)

Applying Theorem 5.2 to stationary policies gives a simpler form of the necessary and sufficient conditions for n discount optimality. Namely

THEOREM 5.4. *Let $n=-\nu,-\nu+1,\cdots$ and let δ^{∞} be a stationary n discount optimal policy. A stationary policy γ^{∞} is n discount optimal if and only if $C^{n}_{\gamma\delta}=0$ and in addition*

$$\begin{aligned} E_{\gamma}Q_{\gamma}^{j-n+1}r_{\gamma}-E_{\gamma}Q_{\gamma}^{j}v_{\delta}^{n}=0 \quad & \text{for } j=0,\cdots,\nu-1 \\ & \text{if } n=-1,-2,\cdots; \\ E_{\gamma}Q_{\gamma}^{j}v_{\delta}^{n}=0 \quad & \text{for } j=0,\cdots,\nu-1 \\ & \text{if } n=0,1,\cdots. \end{aligned} \tag{5.10}$$

Proof. First observe that $\mathscr{S}(0,\gamma^{\infty})=\{1,\cdots,S\}$; thus, (5.1) is equivalent to the requirement $C^{n}_{\gamma\delta}=0$. Next observe that by (3.3),

$$\rho\sum_{N=0}^{\infty}\beta^{i}P_{\gamma}^{i}=\rho\sum_{j=1}^{\nu}\rho^{-j}E_{\gamma}Q_{\gamma}^{j-1}-\sum_{j=0}^{\infty}\rho^{j+1}(-D_{\gamma})^{j+1}.$$

The equivalence of (5.2) and (5.10) now follows immediately.

In the proof of Theorem 5.2 we showed that (5.1) together with (5.9) form necessary and sufficient conditions for n discount optimality. This enables one to obtain a simple sufficient condition for n discount optimality. Applying this condition to stationary policies we get a slight extension of Rothblum (1975a, Thm. 5.1).

COROLLARY 5.5. *Let $n = -\nu, -\nu+1, \cdots$ and let δ^∞ be a stationary n discount optimal policy. A policy $\pi = (\gamma_i)$ is n discount optimal if for every $N = 0, 1, \cdots, \gamma_{N+1} \in \Delta^*_{n+\nu}(N, \pi)$. A stationary policy γ^∞ is n discount optimal if $C^{n+\nu}_{\gamma\delta} = 0$.*

We shall next show that one has to consider n discount optimality only for a finite number of integers n.

THEOREM 5.6. *Let δ^∞ be a stationary ∞ discount optimal policy and let M be the rank of E_δ. A policy $\pi = (\gamma_i)$ is ∞ discount optimal if and only if it is $S-M$ discount optimal.*

Proof. By Corollary 5.6 and the equivalence of (5.1) and (5.3) it suffices to show that if for a given $N = 0, 1, \cdots$,

$$P^N_\pi c^j_{\gamma(N+1),\delta} = 0 \tag{5.11}$$

for $j = -\nu, \cdots, S-M$, then this holds for $j = S-M+1, S-M+2, \cdots$. By the explicit expression of the c's and the v's, (5.11), for $j > 0$, is equivalent to $P^N_\pi(I + Q_{\gamma(N+1)}D_\delta)(-D_\delta)^j r_\delta = 0$, i.e.,

$$(-D_\delta)^j r_\delta \in \text{Null } P^N_\pi(I + Q_{\gamma(N+1)}D_\delta). \tag{5.12}$$

By recalling that $D_\delta = (E_\delta - Q_\delta)^{-1}(I - E_\delta)$ it follows that rank $D_\delta = S - M$. It now follows (e.g., Veinott (1975, Lem. 3, p. 32) that if (5.12) holds for $j = 1, \cdots, S-M$, then it holds for $j = S-M+1, S-M+2, \cdots$. Thus, the proof of Theorem 5.6 is completed.

Acknowledgments. I would like to take this opportunity to express my thanks and gratitude to Professor Arthur F. Veinott Jr., my dissertation advisor, for his guidance and advice, for his insight and perspective, during the preparation of my Ph.D. dissertation.

REFERENCES

P. R. HALMOS (1958), *Finite Dimensional Vector Spaces*, Van Nostrand, Princeton, N.J.

T. KATO (1966), *Perturbation Theory for Linear Operators*, Springer-Verlag, New York.

S. A. LIPPMAN (1968), *On the set of optimal policies in discrete dynamic programming*, J. Math. Anal. Appl., 24, pp. 440–445.

B. L. MILLER AND A. F. VEINOTT, JR. (1969), *Discrete dynamic programming with small interest rate*, Ann. Math. Statist., 40, pp. 366–370.

U. G. ROTHBLUM (1975a), *Normalized Markov decision chains. I: Sensitive discount optimality*, Operations Res., 23, pp. 785–795.

——— (1975b), *Algebraic eigenspaces of nonnegative matrices*, Linear Algebra Appl., 12, pp. 281–292.

U. G. ROTHBLUM AND A. F. VEINOTT, JR. (1976), *Average-overtaking cumulative optimality for polynomially bounded Markov decision chains*, to appear.

K. SLADKY (1974), *On the set of optimal controls for Markov chains with rewards*, Kybernetika, 10, pp. 350–367.

R. S. VARGA (1962), *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J.

A. F. VEINOTT, JR. (1969), *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist., 40, pp. 1635–1660.

——— (1975), *Markov decision chains*, Studies of Mathematics, vol. 10, Studies of Optimization, G. B. Dantzing and B. C. Eaves, eds., Mathematical Association of America, Washington, D.C.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 2, February 1977

DIFFERENTIAL GAMES WITH NO INFORMATION*

D. J. WILSON†

Abstract. This paper considers a broad class of differential games of prescribed duration in which the players obtain no information about the state variables. It is shown that such games always have a value when the players can choose their controls by means of a mixed strategy.

It is further shown that a player can always implement a mixed strategy by drawing a random number from the uniform distribution on the unit interval and then using a control which is determined by the number so chosen.

1. Introduction. The problem of finding *mixed strategy solutions* of differential games with partial information has received scant attention in the literature. It is only recently that a few authors have addressed themselves to this problem [1], [17], [25]–[26]. Although the term "mixed strategy" has been used by Russian authors [16], [22] to denote a form of relaxed control, introduced to differential games by Smol'yakov [22], the type of control to which they refer is not a mixed strategy in the usual sense of game theory. Such relaxed controls cannot be expected to provide a saddle point for a differential game with partial information, except in special cases (e.g., see [7]).

In this paper, we shall consider two person differential games of prescribed duration in which the players receive *no information* about the state variables during the game, except for their initial values, which are known to both players from the start. In such games, the pure strategies of the players will be the so-called *open-loop controls.* Although special cases of such games have been shown to possess pure strategy saddle points [9]–[10], [19]–[21], [25], solutions of the general case must be sought among the mixed strategies [26].

In the present work, a mixed strategy is defined to be a Borel probability measure on the space of pure strategies, and the concept of a *mixed relaxed control* (a probability measure on the space of relaxed controls) is introduced. Warga [24] has shown that the games under consideration always have a saddle point among the mixed relaxed controls. Here, it is shown that any mixed relaxed control can be approximated, in a quite strong sense, by a mixed strategy. Together with Warga's theorem, this implies that the *mixed extension* (that is, the game in which the players may choose an arbitrary mixed strategy) of the original game always has a *value.*

The description of a mixed strategy as a probability measure on a function space is inconvenient, both for implementation during the play and for obtaining necessary or sufficient conditions satisfied by optimal mixed strategies. However, in § 6 it is shown that any mixed strategy can be implemented by using a pure strategy which is completely determined by the value of a number chosen at random from the unit interval. When this method is used to implement a mixed strategy, the expressions for the (random) trajectory and expected payoff are

* Received by the editors June 26, 1975, and in revised form April 29, 1976.
† Department of Mathematics, University of Melbourne, Parkville, Victoria 3052, Australia.

greatly simplified. Necessary conditions satisfied by optimal mixed strategies can then be easily obtained from these simplified expressions.

2. Pure and mixed strategies. Consider a *two player, zero sum, differential game*, the *state* of which lies in *Euclidean n-space* $\mathbb{R}^n$. The state moves along a *trajectory* x determined by the differential equation

$$x'(t) = f(x(t), u(t), v(t)), \qquad x(0) = x_0. \tag{2.1}$$

At each instant t, one *player P* chooses the *control* $u(t)$ from a compact subset U of $\mathbb{R}^p$ and a *second player E* chooses the *control* $v(t)$ from a compact subset V of $\mathbb{R}^q$. When choosing their controls, the players have no information about the present or past values of the state, except for its *initial value* x_0, which is known to both of them from the start. At some *fixed, positive time* T the game finishes and E receives a *payoff* J given by

$$J(u, v) = h(x(T)) + \int_0^T g(x(t), u(t), v(t))\, dt, \tag{2.2}$$

which he strives to make as large as possible and which P tries to keep as small as possible. The functions $f: \mathbb{R}^n \times U \times V \to \mathbb{R}^n$, $g: \mathbb{R}^n \times U \times V \to \mathbb{R}$ and $h: \mathbb{R}^n \to \mathbb{R}$ are assumed to be continuous and f is also assumed to satisfy the following:

Conditions 2.1. (i) For each compact subset K of $\mathbb{R}^n$ there is a constant k such that $\|f(x, u, v) - f(y, u, v)\| < k\|x - y\|$ for all x and y in K, u in U and v in V.

(ii) There is a constant c such that $|x \cdot f(x, u, v)| \leqq c\|x\|^2$ for all x in $\mathbb{R}^n$, u in U and v in V.

These conditions guarantee that the trajectory and payoff of the game are well-defined by equations (2.1) and (2.2) whenever the control functions u and v are measurable [13, p. 7].

Since the players of the game are completely *blind*, their strategies will simply be *open-loop control functions*—that is, functions of time only, with values in the control sets U or V. A Borel measurable function $u: [0, T] \to U$ or $v: [0, T] \to V$ will therefore be called a *pure strategy* for P, or a *pure strategy* for E, respectively. Let $\mathcal{U}$ and $\mathcal{V}$ be the sets of pure strategies of P and E respectively. The sets $\mathcal{U}$ and $\mathcal{V}$ will be regarded as subsets of the spaces $L_2^p[0, T]$ and $L_2^q[0, T]$ of square integrable $\mathbb{R}^p$- and $\mathbb{R}^q$-valued functions respectively. A two person, zero sum game of the form described above will be called a *differential game of prescribed duration with no information*. For the remainder of this paper, we shall regard the functions f, g and h, the terminal time T and the initial condition x_0 as *fixed*. The differential game thus determined by equations (2.1) and (2.2) will be referred to simply as "the game".

As in the case of finite games, the players' lack of information makes it unlikely that the payoff J will have a saddle point among the pure strategies. Therefore, in analogy with the finite case, we define mixed strategies to be probability measures on the spaces of pure strategies.

Notation 2.2. For any topological space X, let $\mathcal{B}(X)$ be the set of *Borel probability measures* on X, and $\mathcal{F}(X)$ the set of those members of $\mathcal{B}(X)$ which have finite support. Further, let $\mathcal{B}(X)$ be equipped with the *topology of weak convergence* [4].

An element of $\mathcal{B}(\mathcal{U})$ or $\mathcal{B}(\mathcal{V})$ will be called a *mixed strategy* for P, or a *mixed strategy* for E, respectively. In § 3 it will be shown that the payoff is a bounded, Borel measurable function on the Cartesian product of the pure strategy sets. It follows that for any pair of mixed strategies, μ for P and ν for E, the *expected payoff* $\int_{\mathcal{U}} \int_{\mathcal{V}} J(u, v)\nu(dv)\mu(du)$, which we set equal to $J(\mu, \nu)$, exists and is independent of the order of integration.

The game with strategy sets $\mathcal{B}(\mathcal{U})$ and $\mathcal{B}(\mathcal{V})$ and payoff J to the second player will be called the *mixed extension* of the original game. If

$$\sup_{\nu \in \mathcal{B}(\mathcal{V})} \inf_{\mu \in \mathcal{B}(\mathcal{U})} J(\mu, \nu) = \inf_{\mu \in \mathcal{B}(\mathcal{U})} \sup_{\nu \in \mathcal{B}(\mathcal{V})} J(\mu, \nu) = W, \tag{2.3}$$

then (the mixed extension of) the game will be said to have the *value* W.

Example 2.3. In the one-dimensional differential game with state equation $x'(t) = 8v(t)^2 - 4(u(t) - v(t))^2 - 6v(t)$, $x(0) = 0$, control constraints $u(t) \in [0, 1]$, $v(t) \in [0, 1]$, and payoff

$$J(u, v) = \int_0^1 x(t)\, dt,$$

neither player has an optimal pure strategy. (This is true even if the players are given complete information; cf. the example of Berkovitz [3].) If the players obtain no information while playing the game, an optimal (mixed) strategy for E is to choose each of the controls $v \equiv 0$ and $v \equiv 1$ with probability $\frac{1}{2}$, and an optimal strategy for P is to choose the controls $u \equiv 0$ and $u \equiv 1$ with probabilities $\frac{3}{4}$ and $\frac{1}{4}$ respectively. The value (of the mixed extension) of the game is $-\frac{1}{2}$.

3. Relaxed and mixed relaxed controls. In order to use currently available minimax theorems to prove that a game has a value, it is necessary to provide the strategy spaces with topologies in which they are at least precompact and the payoff function semicontinuous. The author has previously tried to do this for differential games with no information [25] by adapting a device due to Wald [23]; namely, by using the payoff function and trajectory of the game to define a metric on the pure strategy spaces so that they become precompact. However, the original proof of precompactness given in [25] contains an error, and now a much more tidy approach has become available through the introduction of (open-loop) relaxed controls to differential games by Warga [24, Chap. IX], and Elliott, Kalton and Markus [7]. A *relaxed control* for P (or E resp.) is a function $\rho: [0, T] \to \mathcal{B}(U)$ (or $\sigma: [0, T] \to \mathcal{B}(V)$ resp.) which is Borel measurable. For each pair of relaxed controls, ρ for P and σ for E, a *relaxed trajectory* $\hat{x}$ is defined by the differential equation

$$\hat{x}'(t) = \int_U \int_V f(\hat{x}(t), u, v)\sigma_t(dv)\rho_t(du), \qquad \hat{x}(0) = x_0, \tag{3.1}$$

and a *relaxed payoff* $\hat{J}$ as the integral

$$\hat{J}(\rho, \sigma) = h(\hat{x}(T)) + \int_0^T \int_U \int_V g(\hat{x}(t), u, v)\sigma_t(dv)\rho_t(du)\, dt. \tag{3.2}$$

Note that the conditions satisfied by the functions f, g and h are sufficient to guarantee that the relaxed trajectory and payoff are well-defined. Also, for each pure strategy u of P (*or* v of E), an associated relaxed control ρ^u (or σ^v) can be defined by $\rho_t^u(A) = 1$ if $u(t) \in A$ and $\rho_t^u(A) = 0$ if $u(t) \notin A$ (or $\sigma_t^v(B) = 1$ if $v(t) \in B$ and $\sigma_t^v(B) = 0$ if $v(t) \notin B$).

The relaxed trajectory and payoff generated by the associated relaxed controls of a pair of pure strategies, u for P and v for E, are then identical to the trajectory and payoff determined by equations (2.1) and (2.2). Thus, we can identify a pure strategy with its associated relaxed control and regard the spaces of pure strategies as subsets of the spaces of relaxed controls. The payoff J can then be regarded as the restriction of the relaxed payoff $\hat{J}$ to $\mathcal{U} \times \mathcal{V}$. Let $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ be the sets of relaxed controls for P and E respectively. The game with strategy sets $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$, and payoff $\hat{J}$ to the second player, will be called the *relaxation* of the original game. As f, g, h, T and x_0 are being kept fixed throughout, the differential game determined by (3.1) and (3.2) will be referred to simply as "the relaxed game".

There is a superficial resemblance between relaxed controls and mixed strategies which is a little misleading. This, apparently, has led to the use of the term "mixed strategy" to denote a type of relaxed control [16], [22], and also to the claim [7] that the introduction of relaxed controls to differential games is analogous to Borel and von Neumann's introduction of mixed strategies to matrix games.

The relaxation of the game in Example 2.3, for instance, has a saddle point with the same value, $-\frac{1}{2}$, as the mixed strategy solution given above. The optimal relaxed control for E is the *constant* probability measure on $[0, 1]$ which assigns an equal weight of $\frac{1}{2}$ to each of the points 0 and 1; P's optimal, relaxed control is the constant measure which assigns weight $\frac{3}{4}$ to the point 0 and $\frac{1}{4}$ to the point 1. However, the optimal relaxed trajectory, given by $\hat{x}(t) \equiv -t$, is a *deterministic* function, whereas the optimal mixed strategies, as described above, give rise to a *random* trajectory, which may have any one of the functions $x(t) \equiv 0$, $x(t) \equiv -4t$, $x(t) \equiv -2t$ or $x(t) \equiv 2t$ as its realization (the probabilities of these realizations being $\frac{3}{8}$, $\frac{1}{8}$, $\frac{3}{8}$ and $\frac{1}{8}$ respectively).

Also, there is a straightforward method for implementing a mixed strategy during the play of a game—namely, that of choosing at random from among the pure strategies with probabilities determined by the given mixed strategy (at least, this may be done if the mixed strategy is atomic); but a corresponding physical interpretation of relaxed controls is lacking. In optimal control problems a relaxed control can, to all intents and purposes, be implemented by using an approximating ordinary control (that is, a pure strategy, in our context). However, in a differential game this will not always work, because the approximation cannot always be made to hold uniformly over all the opponent's strategies.

In some differential game models of military and economic conflicts [5], [15], events which are really of finite duration occur instantaneously in the model. In such game models a relaxed control may be interpreted as an instantaneously mixed strategy (somewhat similar to a behavior strategy in a discrete-time game) which approximates some mixed strategy that could be used over intervals of real time (see also a related discussion by Warga [24, pp. 457–459]). Here, I do not

wish to assume that the resolution of sequences of events is quite so coarse. It is therefore stressed that, in this paper, no such interpretation of relaxed controls is contemplated. A relaxed control, as used herein, is simply a mathematical device for completing (in the *topological* sense) the pure strategy spaces, and is in no way thought of as capable of being used by a player. Thus, a solution of the game in terms of relaxed controls will not be of much help to a player unless he can find a strategy (either pure or mixed) which will give him nearly the same payoff, whatever the pure strategy of his opponent. In § 4 it will be shown that such an approximation is always possible.

In order to attain the desired ends, it is necessary to topologize the spaces of relaxed controls so that they are compact, so that the spaces of pure strategies are dense, and so that the payoff function is at least continuous in each relaxed control separately. The topology used by Warga [24, p. 272], and Elliott, Kalton and Markus [7] admirably fulfills all these requirements. A neighborhood of a member ρ of $\hat{\mathcal{U}}$ is taken to be any set containing a finite intersection of sets of the form $\{\rho' \in \hat{\mathcal{U}}; |\int_0^T [\int_U \phi(u,t)\rho_t'(du) - \int_U \phi(u,t)\rho_t(du)]\, dt| < \varepsilon\}$, where $\varepsilon > 0$ and the function $\phi: U \times [0, T] \to \mathbb{R}$ satisfies the following:

Conditions 3.1.

(i) for each u in U, $\phi(u, \cdot)$ is measurable;

(ii) for each t in $[0, T]$, $\phi(\cdot, t)$ is continuous, and

(iii) $\int_0^T \sup_{u \in U} |\phi(u,t)|\, dt < \infty$.

Let $\hat{\mathcal{U}}$ be equipped with the topology generated by these neighborhoods and let $\hat{\mathcal{V}}$ be equipped with a similarly defined topology. The spaces $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ and the payoff $\hat{J}$ have the following useful properties.

LEMMA 3.2. (i) *$\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ are compact and metrizable*, [24, p. 272].

(ii) *$\mathcal{U}$ and $\mathcal{V}$ are dense in $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ respectively*, [24, p. 287].

(iii) *The function $\hat{J}$ is bounded on $\hat{\mathcal{U}} \times \hat{\mathcal{V}}$. Also, for any ρ in $\hat{\mathcal{U}}$ and σ in $\hat{\mathcal{V}}$, $\hat{J}(\rho, \cdot)$ is continuous on $\hat{\mathcal{V}}$ and $\hat{J}(\cdot, \sigma)$ is continuous on $\hat{\mathcal{U}}$*, [24, pp. 349, 477].

The boundedness of $\hat{J}$ follows from the continuity of f, g and h and from the fact that, over any finite time interval, all the trajectories of (3.1) lie within a common compact set [24, p. 349]. The continuity properties of $\hat{J}$ are asserted by Warga [24, p. 477] to hold under the condition that f, g and h have bounded, continuous partial derivatives. However, they are retained under much more general conditions. The argument at the bottom of p. 325 in [24] shows that the relaxed trajectory defined by (3.1) is continuous in each relaxed control. The last conclusion of Lemma 3.2 then follows from IV.2.9 of [24, p. 278].

DEFINITION 3.3. An element of $\mathcal{B}(\hat{\mathcal{U}})$ (or of $\mathcal{B}(\hat{\mathcal{V}})$ resp.) will be called a *mixed relaxed control* for P (or for E resp.)

Remarks 3.4. The sets $\mathcal{U}$ and $\mathcal{V}$ have now been provided with two apparently different topologies. Besides the L_2 topologies which they acquire as subsets of $L_2^p[0, T]$ and $L_2^q[0, T]$ respectively, they are also equipped with the relativized topologies of $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ which they acquire through being imbedded in these sets. It is tempting to identify mixed strategies with those mixed relaxed controls which have supports contained in $\mathcal{U}$ or $\mathcal{V}$. However, to do this it is necessary to show that the Borel subsets of $\mathcal{U}$ and $\mathcal{V}$ generated by the L_2 topologies are the same as those generated by the relativized topologies of $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$. Indeed this can be done.

A referee has pointed out[1] to me that the topology of $\hat{\mathcal{U}}$, when restricted to $\mathcal{U}$, is identical with the $L_s^p[0, T]$ topology of $\mathcal{U}$, for any s in $[1, \infty)$. As $\mathcal{U}$ is also a Borel subset of $\hat{\mathcal{U}}$, it follows that the members of $\mathcal{B}(\mathcal{U})$ are precisely the restrictions, to the Borel subsets of $\mathcal{U}$, of those mixed relaxed controls which have supports contained in $\mathcal{U}$. Since the only mixed strategies which explicitly appear in the main results of § 4 all have finite support, these observations are not used. However I have included their proofs in an Appendix, for the sake of completeness.

The following lemma justifies the subsequent definition of expected relaxed payoff. It is a slightly stronger version of a result of Michael and Rennie's [28].

LEMMA 3.5. *$\hat{J}$ is Borel measurable on $\hat{\mathcal{U}} \times \hat{\mathcal{V}}$.*

Proof. Let $\hat{\mathcal{U}}$ be equipped with a metric (see Lemma 3.2). Since $\hat{\mathcal{U}}$ is compact, then for every positive integer r there is a partition $\{S_1^r, S_2^r, \cdots, S_{j_r}^r\}$ of $\hat{\mathcal{U}}$ into Borel subsets of diameter less than $1/r$. Whenever i and r are positive integers such that $1 \leqq i \leqq j_r$, choose any member α_i^r of S_i^r and put $\phi_r(\rho, \sigma) = \sum_{i=1}^{j_r} \chi_{S_i^r}(\rho)\hat{J}(\alpha_i^r, \sigma)$ for every (ρ, σ) in $\hat{\mathcal{U}} \times \hat{\mathcal{V}}$. Then $\{\phi_r\}$ is a sequence of functions, each bounded and Borel measurable on $\hat{\mathcal{U}} \times \hat{\mathcal{V}}$, which converges (pointwise) to $\hat{J}$. Since the limit of such a sequence must be Borel measurable, the lemma is proved.

In view of the above remarks, Lemma 3.5 also justifies the assertion made in § 2 that J is Borel measurable on $\mathcal{U} \times \mathcal{V}$.

For each pair of mixed relaxed controls, $\hat{\mu}$ for P and $\hat{\nu}$ for E, an *expected relaxed payoff* $\hat{J}(\hat{\mu}, \hat{\nu})$ is defined by $\hat{J}(\hat{\mu}, \hat{\nu}) = \int_{\hat{\mathcal{U}}} \int_{\hat{\mathcal{V}}} \hat{J}(\rho, \sigma)\hat{\nu}(d\sigma)\hat{\mu}(d\rho)$. The existence of this integral, along with that of the integral defining expected payoff, follows from Lemma 3.5.

The game with strategy sets $\mathcal{B}(\hat{\mathcal{U}})$ and $\mathcal{B}(\hat{\mathcal{V}})$, and payoff $\hat{J}$ to the second player, will be referred to throughout as "the mixed extension of the relaxed game" or, more briefly, as "the mixed relaxed game". If

$$\sup_{\hat{\nu} \in \mathcal{B}(\hat{\mathcal{V}})} \inf_{\hat{\mu} \in \mathcal{B}(\hat{\mathcal{U}})} \hat{J}(\hat{\mu}, \hat{\nu}) = \inf_{\hat{\mu} \in \mathcal{B}(\hat{\mathcal{U}})} \sup_{\hat{\nu} \in \mathcal{B}(\hat{\mathcal{V}})} \hat{J}(\hat{\mu}, \hat{\nu}) = \hat{W},$$

then $\hat{W}$ will be referred to as the "relaxed value" of the game.

4. The existence of a value. In this section it is shown that a differential game of prescribed duration with no information always has an arbitrarily approximate solution among the mixed strategies; more precisely its *mixed extension* is shown to always have a value. The proof of this relies on the fact, noted previously by Warga [24, p. 477], that the corresponding mixed relaxed game always has a saddle point.

In order that the results of this section may be stated in their most general form, let $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ now denote *arbitrary* compact metric spaces and $\mathcal{U}$ and $\mathcal{V}$ dense Borel subsets of them. Let $\hat{J}$ be a bounded, real-valued function on $\hat{\mathcal{U}} \times \hat{\mathcal{V}}$ such that $\hat{J}(\rho, \cdot)$ is continuous on $\hat{\mathcal{V}}$ and $\hat{J}(\cdot, \sigma)$ is continuous on $\hat{\mathcal{U}}$ for any ρ in $\hat{\mathcal{U}}$ and σ in $\hat{\mathcal{V}}$. Note that the proof of Lemma 3.5 does not depend on any special properties of $\hat{J}$ as defined in § 3, but is also valid when $\hat{J}$ has just the properties

[1] As his source of the observation, he cited an oral communication of Artstein [27] who has kindly permitted me to include the proof which appears in the Appendix.

stated above. We are thus able to extend $\hat{J}$ to $\mathcal{B}(\hat{\mathcal{U}})\times\mathcal{B}(\hat{\mathcal{V}})$ by letting $\hat{J}(\hat{\mu}, \hat{\nu})$ denote the expected value of $\hat{J}$ with respect to the product of the measures $\hat{\mu}$ of $\mathcal{B}(\hat{\mathcal{U}})$ and $\hat{\nu}$ of $\mathcal{B}(\hat{\mathcal{V}})$.

Suppose the players P and E now play the game with strategy sets $\mathcal{U}$ for P and $\mathcal{V}$ for E, and payoff function $\hat{J}$ to E. Their *mixed strategies* will again be the elements of $\mathcal{B}(\mathcal{U})$ and $\mathcal{B}(\mathcal{V})$. Lemma 3.2 shows that this game is a generalization of the differential game defined by (2.1) and (2.2), for which all the following results thus hold. The first of these is a variation on von Neumann's minimax theorem. A proof of it has already been given by Warga [24, p. 273], but the following one is much shorter.

THEOREM 4.1. *The function $\hat{J}$ has a saddle point in $\mathcal{B}(\hat{\mathcal{U}})\times\mathcal{B}(\hat{\mathcal{V}})$; that is, there exist measures $\hat{\mu}^*$ in $\mathcal{B}(\hat{\mathcal{U}})$ and $\hat{\nu}^*$ in $\mathcal{B}(\hat{\mathcal{V}})$ such that*

$$\hat{J}(\hat{\mu}^*, \hat{\nu}) \leqq \hat{J}(\hat{\mu}, \hat{\nu}^*) \tag{4.1}$$

for every $\hat{\mu}$ in $\mathcal{B}(\hat{\mathcal{U}})$ and $\hat{\nu}$ in $\mathcal{B}(\hat{\mathcal{V}})$.

Proof. In the sense of Fan [8], the function $\hat{J}$ is convex on $\mathcal{B}(\hat{\mathcal{U}})$ and concave on $\mathcal{B}(\hat{\mathcal{V}})$. Also, since $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ are compact, so are $\mathcal{B}(\hat{\mathcal{U}})$ and $\mathcal{B}(\hat{\mathcal{V}})$ [4, pp. 35–37]. Let ρ and $\hat{\nu}$ belong to $\hat{\mathcal{U}}$ and $\mathcal{B}(\hat{\mathcal{V}})$ respectively, and suppose that $\{\rho_r\}$ is a sequence of members of $\hat{\mathcal{U}}$, which converges to ρ. It follows, from the Lebesgue dominated convergence theorem, that

$$\lim_{r\to\infty}\int_{\hat{\mathcal{V}}} \hat{J}(\rho_r, \sigma)\hat{\nu}(d\sigma) = \int_{\hat{\mathcal{V}}} \hat{J}(\rho, \sigma)\hat{\nu}(d\sigma).$$

Thus, since $\hat{\mathcal{U}}$ is metrizable, then the function $\psi\colon \hat{\mathcal{U}}\to\mathbb{R}$ defined by $\psi(\rho)\equiv \int_{\hat{\mathcal{V}}} \hat{J}(\rho, \sigma)\hat{\nu}(d\sigma)$ is continuous on $\hat{\mathcal{U}}$. Therefore [4, p. 7], $\int_{\hat{\mathcal{V}}} \psi(\rho)\hat{\mu}(d\rho)$, regarded as a function of $\hat{\mu}$, is continuous on $\mathcal{B}(\hat{\mathcal{U}})$; that is, $\hat{J}(\cdot\,, \hat{\nu})$ is continuous on $\mathcal{B}(\hat{\mathcal{U}})$. Similarly, for each $\hat{\mu}$ in $\mathcal{B}(\hat{\mathcal{U}})$, it can be shown that $\hat{J}(\hat{\mu}, \cdot\,)$ is continuous on $\mathcal{B}(\hat{\mathcal{V}})$. The result now follows from Fan's minimax theorem [8].

Theorem 4.1 implies there exists a value of the game with strategy sets $\mathcal{B}(\hat{\mathcal{U}})$ and $\mathcal{B}(\hat{\mathcal{V}})$, and payoff $\hat{J}$ to the second player. As this game is a generalization of the relaxed game of § 3, I shall denote its value also by $\hat{W}$.

When applied to the differential game of §§ 2 and 3, Theorem 4.1 shows that it always has a relaxed value. As noted earlier, however, there is no known way of implementing a relaxed control in the play of a differential game. Even a mixed strategy might not be practicable unless it were atomic, since otherwise it would require randomization from an *uncountable* set of pure strategies. Thus arises the important question of whether or not a player can guarantee himself a payoff arbitrarily close to the relaxed value by making use only of atomic mixed strategies. The following results show that the answer is yes. In fact they show that he can do it even if he is restricted to the use of mixed strategies with *finite support.* Such strategies can be implemented by randomization from a *finite* set of pure strategies.

Note that a game has a value if and only if it has an *ε-saddle point* for every positive number ε. An ε-saddle point is a pair of mixed strategies, μ_ε for P and ν_ε for E, such that

$$\hat{J}(\mu_\varepsilon, \nu) - \varepsilon/2 \leqq \hat{J}(\mu, \nu_\varepsilon) + \varepsilon/2 \tag{4.2}$$

for every ν in $\mathcal{B}(\mathcal{V})$ and μ in $\mathcal{B}(\mathcal{U})$.

The main results of this section are the following:

THEOREM 4.2. *For every positive number ε, $\hat{\mu}$ in $\mathcal{B}(\hat{\mathcal{U}})$ and $\hat{\nu}$ in $\mathcal{B}(\hat{\mathcal{V}})$, there exists μ_ε in $\mathcal{F}(\mathcal{U})$ and ν_ε in $\mathcal{F}(\mathcal{V})$ such that*

$$\left| \int_{\mathcal{U}} \hat{J}(u, \sigma)\mu_\varepsilon(du) - \int_{\hat{\mathcal{U}}} \hat{J}(\rho, \sigma)\hat{\mu}(d\rho) \right| < \varepsilon \tag{4.3}$$

for every σ in $\hat{\mathcal{V}}$, and

$$\left| \int_{\mathcal{V}} \hat{J}(\rho, v)\nu_\varepsilon(dv) - \int_{\hat{\mathcal{V}}} \hat{J}(\rho, \sigma)\hat{\nu}(d\sigma) \right| < \varepsilon \tag{4.4}$$

for every ρ in $\hat{\mathcal{U}}$.

THEOREM 4.3. *For every positive number ε, there exists μ_ε in $\mathcal{F}(\mathcal{U})$ and ν_ε in $\mathcal{F}(\mathcal{V})$ such that inequality* (4.2) *is satisfied. Consequently the game always has a value.*

Theorem 4.3 is an almost immediate consequence of Theorems 4.1 and 4.2. The proof of Theorem 4.2 proceeds via a couple of auxiliary lemmas.

Let $C(\hat{\mathcal{V}})$ be the Banach space of continuous, real-valued functions on $\hat{\mathcal{V}}$ equipped with the supremum norm (defined by $\|\phi\| = \sup_{\sigma \in \hat{\mathcal{V}}} \phi(\sigma)$). For each subset X of $C(\hat{\mathcal{V}})$ let $\operatorname{co}(X)$, $\operatorname{cl} X$ and $\operatorname{wcl} X$ be the convex hull, closure and weak closure of X, respectively.

LEMMA 4.4. *Let Y be a dense subset of a compact metric space Z, and $I: Z \to C(\hat{\mathcal{V}})$ a bounded function such that $\{I(\cdot)\}(\sigma)$ is continuous on Z for each σ in $\hat{\mathcal{V}}$. Then $I(Z) \subset \operatorname{wcl} I(Y)$.*

Proof. Let ϕ belong to $I(Z)$ and put $\phi = I(z)$, where $z \in Z$. Then there exists a sequence $\{y_r\}$, of members of Y, which converges to z. It follows that $\lim_{r\to\infty} \{I(y_r)\}(\sigma) = \{I(z)\}(\sigma) = \phi(\sigma)$. Therefore [6, pp. 265–266], ϕ is the weak limit of the sequence $\{I(y_r)\}$ and so belongs to $\operatorname{wcl} I(Y)$. This completes the proof of the lemma.

Now consider the conclusion of Lemma 4.4 for the function $I: \hat{\mathcal{U}} \cup \mathcal{B}(\hat{\mathcal{U}}) \to C(\hat{\mathcal{V}})$ defined by $I(\rho) = \hat{J}(\rho, \cdot)$ when $\rho \in \hat{\mathcal{U}}$, and $I(\hat{\mu}) = \int_{\hat{\mathcal{U}}} \hat{J}(\rho, \cdot)\hat{\mu}(d\rho)$ when $\hat{\mu} \in \mathcal{B}(\hat{\mathcal{U}})$. Lemma 3.2 implies that $I(\rho) \in C(\hat{\mathcal{V}})$ whenever $\rho \in \hat{\mathcal{U}}$; in conjunction with the Lebesgue dominated convergence theorem it also implies that $I(\hat{\mu}) \in C(\hat{\mathcal{V}})$ whenever $\hat{\mu} \in \mathcal{B}(\hat{\mathcal{U}})$. Clearly,

$$I(\mathcal{F}(\hat{\mathcal{U}})) = \operatorname{co}(I(\hat{\mathcal{U}})) \quad \text{and} \quad I(\mathcal{F}(\mathcal{U})) = \operatorname{co}(I(\mathcal{U})). \tag{4.5}$$

Also, from Lemma 3.2 it follows that the conditions of Lemma 4.4 are satisfied when $Z = \hat{\mathcal{U}}$ and $Y = \mathcal{U}$. Therefore,

$$I(\hat{\mathcal{U}}) \subset \operatorname{wcl} I(\mathcal{U}). \tag{4.6}$$

Since $\mathcal{F}(\hat{\mathcal{U}})$ is dense in $\mathcal{B}(\hat{\mathcal{U}})$ [4, p. 237], and $\mathcal{B}(\hat{\mathcal{U}})$ is compact and metrizable [4, pp. 35–37, 236–238], then the conditions of Lemma 4.4 are satisfied when $Z = \mathcal{B}(\hat{\mathcal{U}})$ and $Y = \mathcal{F}(\hat{\mathcal{U}})$. (This follows from Lemma 3.2 and the definition of the topology of $\mathcal{B}(\hat{\mathcal{U}})$ [4, p. 7]). Thus, Lemma 4.4 also gives

$$I(\mathcal{B}(\hat{\mathcal{U}})) \subset \operatorname{wcl} I(\mathcal{F}(\hat{\mathcal{U}})). \tag{4.7}$$

LEMMA 4.5. $I(\mathcal{B}(\hat{\mathcal{U}})) \subset \operatorname{cl} I(\mathcal{F}(\mathcal{U}))$.

Proof. Since the space $C(\hat{\mathcal{V}})$ is locally convex, then [6, p. 422] cl co $(I(\mathcal{U}) =$ wcl co $(I(\mathcal{U}))$. Thus $I(\mathcal{B}(\hat{\mathcal{U}})) \subset$ wcl $I(\mathcal{F}(\hat{\mathcal{U}}))$ (from (4.7)) = wcl co $(I(\hat{\mathcal{U}}))$ (from 4.5) $\subset$ wcl co (wcl $I(\mathcal{U})$) (from (4.6)) = wcl co $(I(\mathcal{U}))$ = cl co $(I(\mathcal{U}))$ (from above) = cl $I(\mathcal{F}(\mathcal{U}))$ (from (4.5)), which proves the lemma.

Proof of Theorem 4.2. Only the proof of (4.3) is given; the proof of (4.4) is essentially the same, but it requires the recasting of Lemma 4.5 into a suitable form (viz. $\mathcal{U}$ has to be replaced by $\mathcal{V}$).

Let $\hat{\mu}$ belong to $\mathcal{B}(\hat{\mathcal{U}})$ and ε be a positive number. Since $I(\hat{\mu}) \in I(\mathcal{B}(\hat{\mathcal{U}}))$, it follows from Lemma 4.5 that there is a function ϕ in $I(\mathcal{F}(\mathcal{U}))$ such that $\|\phi - I(\hat{\mu})\| < \varepsilon$.

If μ_ε is now chosen from $\mathcal{F}(\mathcal{U})$ so that $I(\mu_\varepsilon) = \phi$, then μ_ε satisfies (4.3). This completes the proof.

Proof of Theorem 4.3. Let $\hat{\mu}^*$ and $\hat{\nu}^*$ be members of $\mathcal{B}(\hat{\mathcal{U}})$ and $\mathcal{B}(\hat{\mathcal{V}})$, respectively, which satisfy inequality (4.1) for all $\hat{\mu}$ and $\hat{\nu}$. Let ε be any positive number. Choose μ_ε in $\mathcal{F}(\mathcal{U})$ and ν_ε in $\mathcal{F}(\mathcal{V})$ so that (4.3) and (4.4) are satisfied when $\hat{\mu}$ is replaced by $\hat{\mu}^*$, $\hat{\nu}$ by $\hat{\nu}^*$ and ε by $\varepsilon/2$. From (4.3) and (4.1) we get $\int_{\mathcal{U}} \hat{J}(u, v)\mu_\varepsilon(du) \leqq \int_{\hat{\mathcal{U}}} \hat{J}(\rho, v)\hat{\mu}^*(d\rho) + (\varepsilon/2) \leqq \hat{W} + (\varepsilon/2)$ for every v in $\mathcal{V}$. Similarly (4.4) and (4.1) give $\hat{W} - (\varepsilon/2) \leqq \int_{\mathcal{V}} \hat{J}(u, v)\nu_\varepsilon(dv)$ for every u in $\mathcal{U}$. Combining these last inequalities, we have

$$\int_{\mathcal{U}} \hat{J}(u, v)\mu_\varepsilon(du) - \frac{\varepsilon}{2} \leqq \hat{W} \leqq \int_{\mathcal{V}} \hat{J}(u, v)\nu_\varepsilon(dv) + \frac{\varepsilon}{2} \tag{4.8}$$

for all u in $\mathcal{U}$ and v in $\mathcal{V}$. If now, μ and ν are any mixed strategies for P and E respectively, then (4.2) is obtained by integrating the right-hand inequality of (4.8) with respect to μ, and the left-hand inequality with respect to ν. This completes the proof of 4.3.

Remark 4.6. The above results can be applied to more general forms of payoff function than those considered here. This paper has followed Isaacs [14] in considering payoff functions with an integral and a terminal component. A common generalization is to replace the terminal component with a functional on the space of trajectories. Provided this functional is continuous with respect to the supremum norm, all our results will still apply.

5. Representation of mixed strategies. According to our definition, a mixed strategy is a measure on a function space (the space of pure strategies). This is inconvenient both for implementation during the play of a game and for deriving the necessary or sufficient conditions satisfied by optimal strategies. In 1954, Fleming [12] showed that certain games over a function space have mixed strategy solutions of a simple kind. In the games which he considered, the pure strategies were functions mapping a compact metric space into a compact subset of an Euclidean space and satisfying constraints imposed by functional equations and inequalities. He showed that these games have a mixed strategy solution which can be implemented by the players' independently drawing numbers α and β at random from the unit interval and then playing pure strategies u_α and v_β determined by the numbers so chosen. Fichefet [11] has applied these results to differential games with payoff functions that can be cast into the same form as those considered by Fleming. Unfortunately, most differential games do not have

payoff functions of this type, so the results are not applicable to the vast majority of them.

The following theorem implies that if the strategy spaces of a game are Borel subsets of complete, separable metric spaces, then *all* mixed strategies (that is, Borel probability measures) can be implemented in the simple way described above. Since the strategy spaces of differential games are usually Borel subsets of separable Banach spaces, this result is immediately applicable. A proof of this theorem was first given in [25] for *compact* metric spaces; the following proof is much shorter.

THEOREM 5.1. *If Z is a complete separable metric space and Π a Borel probability measure on Z, then there exists a Borel measurable function $\phi\colon (0,1)\to Z$ which has Π as its distribution with respect to Lebesgue measure on $(0,1)$. That is, $\mathrm{m}(\{\alpha\in(0,1);\ \phi(\alpha)\in A\})=\Pi(A)$ for any Borel subset A of Z.*

Proof. We first prove the theorem in the case when Z is countable. Let $\{z_r\}_{r=1}^{\infty}$ be an enumeration of Z, and $p_r=\Pi(\{z_r\})$ for each r. Then define ϕ by setting $\phi(\alpha)=z_r$ whenever $\sum_{j=1}^{r-1}p_j\leqq\alpha<\sum_{j=1}^{r}p_j$. Clearly ϕ has distribution Π. Next the theorem is proved for the case $Z=\mathbb{R}$. Let F be the distribution function of Π and define ϕ by setting $\phi(\alpha)=\sup\{z\in\mathbb{R};\ F(z)\leqq\alpha\}$ for α in $(0,1)$. Since F is monotonically increasing, so is ϕ; therefore, ϕ is Borel measurable. Now suppose that $\alpha\in\phi^{-1}((-\infty,z])$, where $z\in\mathbb{R}$. It follows that $\phi(\alpha)\leqq z$, and therefore that $F(\phi(\alpha))\leqq F(z)$, since F is increasing. But from the definition of ϕ and the right continuity of F, we may conclude that $F(\phi(\alpha))\geqq\alpha$. Thus we have $0<\alpha\leqq F(\phi(\alpha))\leqq F(z)$ and so $\alpha\in(0,F(z)]$. We have now shown that $\phi^{-1}((-\infty,z])\subset(0,F(z)]$ for any z in $\mathbb{R}$. Conversely, suppose that $\alpha\in(0,F(z))$, where $z\in\mathbb{R}$. Then for any real number y such that $F(y)\leqq\alpha$, we must have $F(y)<F(z)$, and therefore $y<z$ (since F is increasing). It thus follows, from the definition of ϕ, that $\phi(\alpha)\leqq z$. That is, $\alpha\in\phi^{-1}((-\infty,z])$. We have now shown that $(0,F(z))\subset\phi^{-1}((-\infty,z])\subset(0,F(z)]$. Consequently $\mathrm{m}(\phi^{-1}((-\infty,z]))=F(z)=\Pi((-\infty,z])$ for any real number z. Since the measures $\mathrm{m}\phi^{-1}$ and Π agree on half-lines, then they must agree on all Borel sets. That is, Π is the distribution of ϕ. To prove the theorem in the general case, we now make use of the following isomorphism theorem [18, pp. 7, 12, 14]:

If Z is uncountable then there exists a one-to-one Borel measurable function ζ mapping $\mathbb{R}$ onto Z such that ζ^{-1} is Borel measurable.

Put $\Pi_1(A)=\Pi(\zeta(A))$ for all Borel subsets A of $\mathbb{R}$. It follows that Π_1 is a Borel probability measure. Therefore, by the last part of the proof, there exists a Borel measurable function $\phi_1\colon(0,1)\to\mathbb{R}$ with distribution Π_1. Thus, $\zeta\circ\phi_1\colon(0,1)\to Z$ is a Borel measurable function with distribution Π. This completes the proof.

From Theorem 5.1 it follows that a mixed strategy μ, for P say, can be represented by a function $u\colon(0,1)\to\mathcal{U}$ which has distribution μ. Alternatively, we may regard u as a function from $(0,1)\times[0,T]$ into U. Now, if P draws a random number α from the uniform distribution on $(0,1)$, then the pure strategy $u(\alpha,\cdot)$ will be distributed over $\mathcal{U}$ according to the distribution μ. He can therefore implement the mixed strategy μ by playing a pure strategy $u(\alpha,\cdot)$ chosen in this way. Similarly, if ν is a mixed strategy for E, there is a function $v\colon(0,1)\times[0,T]\to V$ which he can use to implement ν. To play ν, he first chooses a random number β from the uniform distribution on $(0,1)$ and then plays the pure strategy $v(\beta,\cdot)$. The (random) trajectory generated by strategies chosen in this way will satisfy the

differential equation

$$\dot{x}(\alpha, \beta; t) = f(x(\alpha, \beta; t), u(\alpha, t), v(\beta, t)), \qquad x(\alpha, \beta; 0) = x_0,$$

for every α and β in $(0, 1)$. The expected payoff will be given by

$$J(\mu, \nu) = \int_0^1 \int_0^1 \{h(x(\alpha, \beta; T)) + \int_0^T g(x(\alpha, \beta; t), u(\alpha, t), v(\beta, t))\, dt\}\, d\alpha\, d\beta.$$

With the trajectory and payoff cast in this form, it is now a simple matter to write down a set of necessary conditions satisfied by any functions $u^*: (0, 1) \times [0, T] \to U$ and $v^*: (0, 1) \times [0, T] \to V$ which represent an optimal pair μ^*, ν^* of mixed strategies. Let $x^*(\alpha, \beta; \cdot)$ be the trajectory generated from x_0 by the controls $u^*(\alpha, \cdot)$ and $v^*(\beta, \cdot)$. Suppose that f, g and h have continuous partial derivatives with respect to the state variables and let $\lambda: (0, 1) \times (0, 1) \times [0, T] \to \mathbb{R}^n$ be the costate trajectory defined by the system of differential equations,

$$\dot{\lambda}_j(\alpha, \beta; t) = -\sum_{i=1}^{n} \lambda_i(\alpha, \beta; t) \frac{\partial f_i}{\partial x_j}(x^*(\alpha, \beta; t), u^*(\alpha, t), v^*(\beta, t))$$

$$-\frac{\partial g}{\partial x_j}(x^*(\alpha, \beta; t), u^*(\alpha, t), v^*(\beta, t)),$$

$$\lambda_j(\alpha, \beta; T) = \frac{\partial h}{\partial x_j}(x^*(\alpha, \beta; T)).$$

If we define the Hamiltonian $\mathcal{H}: (0, 1) \times (0, 1) \times U \times V \times [0, T] \to \mathbb{R}$ by $\mathcal{H}(\alpha, \beta, u, v, t) = \lambda(\alpha, \beta; t) \cdot f(x^*(\alpha, \beta; t), u, v) + g(x^*(\alpha, \beta; t), u, v)$, then it is necessary that $\int_0^1 \int_0^1 \mathcal{H}(\alpha, \beta, u^*(\alpha, t), v^*(\beta, t), t)\, d\alpha\, d\beta$ be essentially constant on $[0, T]$ and that u^*, v^* satisfy the following minimax principle:

(i) $$\int_0^1 \mathcal{H}(\alpha, \beta, u^*(\alpha, t), v^*(\beta, t), t)\, d\beta = \inf_{u \in U} \int_0^1 \mathcal{H}(\alpha, \beta, u, v^*(\beta, t), t)\, d\beta$$

for a.e. $(\alpha, t) \in (0, 1) \times [0, T]$,

(ii) $$\int_0^1 \mathcal{H}(\alpha, \beta, u^*(\alpha, t), v^*(\beta, t), t)\, d\alpha = \sup_{v \in V} \int_0^1 \mathcal{H}(\alpha, \beta, u^*(\alpha, t), v, t)\, d\alpha$$

for a.e. $(\beta, t) \in (0, 1) \times [0, T]$.

The proof of these assertions is tedious [25] and will appear elsewhere.

It is possible to show that the functions representing mixed strategies, in the way described above, are Lebesgue measurable on $(0, 1) \times [0, T]$. For the purposes of this paper, no regularity of these functions was required, and so the proof of measurability has been omitted. It also is tedious, though straightforward, and may be found in [25].

Theorem 5.1 may also be used to justify a similar representation of relaxed controls. If $\rho \in \hat{\mathcal{U}}$, then by Theorem 5.1, for each t in $[0, T]$, the measure ρ_t of $\mathcal{B}(U)$ must be the distribution of some Borel measurable function $u_t: (0, 1) \to U$. Similarly, if $\sigma \in \hat{\mathcal{V}}$ and $t \in [0, T]$, there exists $v_t: (0, 1) \to V$ with distribution σ_t.

Again, it is possible to show that the functions $\hat{u}: (0, 1)\times[0, T]\to U$ and $\hat{v}: (0, 1)\times[0, T]\to V$ defined by $\hat{u}(\omega, t) = u_t(\omega)$ and $\hat{v}(\zeta, t) = v_t(\zeta)$ are measurable. The relaxed trajectory $\hat{x}$ corresponding to ρ and σ now satisfies the differential equation

$$\hat{x}'(t) = \int_0^1 \int_0^1 f(\hat{x}(t), \hat{u}(\omega, t), \hat{v}(\zeta, t))\, d\omega\, d\zeta, \qquad \hat{x}(0) = x_0,$$

and the relaxed payoff is given by

$$\hat{J}(\rho, \sigma) = h(\hat{x}(T)) + \int_0^T \int_0^1 \int_0^1 g(\hat{x}(t), \hat{u}(\omega, t), \hat{v}(\zeta, t))\, d\omega\, d\zeta\, dt.$$

Finally it is clear that a further application of Theorem 5.1 will show that functions $\hat{u}: (0, 1)\times(0, 1)\times[0, T]\to U$ and $\hat{v}: (0, 1)\times(0, 1)\times[0, T]\to V$ can be used to represent mixed relaxed controls $\hat{\mu}$ of $\mathcal{B}(\hat{\mathcal{U}})$ and $\hat{\nu}$ of $\mathcal{B}(\hat{\mathcal{V}})$, respectively. The (random) relaxed trajectory satisfies the differential equation

$$\frac{d\hat{x}(\alpha, \beta; t)}{dt} = \int_0^1 \int_0^1 f(\hat{x}(\alpha, \beta; t), \hat{u}(\alpha, \omega, t), \hat{v}(\beta, \zeta, t))\, d\omega\, d\zeta,$$

$$\hat{x}(\alpha, \beta; 0) = x_0,$$

and the expected relaxed payoff is given by

$$\hat{J}(\hat{\mu}, \hat{\nu}) = \int_0^1 \int_0^1 \{h(\hat{x}(\alpha, \beta; T)) + \int_0^T \int_0^1 \int_0^1 g(\hat{x}(\alpha, \beta; t), \hat{u}(\alpha, \omega, t), \hat{v}(\beta, \zeta, t))\, d\omega\, d\zeta\}\, d\alpha\, d\beta.$$

Appendix A. The following theorem justifies the remarks made in Remark 3.4.

THEOREM A.1. (i) (Artstein [27]). *The topologies induced on $\mathcal{U}$ and $\mathcal{V}$ by $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$, respectively, are identical with their L_s topologies whenever $s \in [1, \infty)$.*

(ii) *$\mathcal{U}$ and $\mathcal{V}$ are Borel subsets of $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ respectively.*

Proof. Proofs will be given for $\mathcal{U}$ only. Those for $\mathcal{V}$ are identical.

(i) Since all the topologies are metric, it is sufficient to show that convergence in $\hat{\mathcal{U}}$ is the same as convergence in L_s. First suppose that a sequence $\{u_r\}$ of $\mathcal{U}$ has limit $\bar{u}$ in $\mathcal{U}$ with respect to the topology of $\hat{\mathcal{U}}$. Then

$$\lim_{r\to\infty} \int_0^T \phi(t, u_r(t))\, dt = \int_0^T \phi(t, \bar{u}(t))\, dt \tag{A.1}$$

whenever ϕ satisfies Conditions 3.1. Putting $\phi(t, u) = \|u - \bar{u}(t)\|^s$ in (A.1), we see that $\{u_r\}$ converges to $\bar{u}$ in L_s.

Conversely, suppose that $\{u_r\}$ converges to $\bar{u}$ in L_s. It follows that $\{u_r\}$ converges to $\bar{u}$ in measure. Let ϕ satisfy 3.1 and define functions $\psi_r: [0, T]\to\mathbb{R}$ by $\psi_r(t) = \sup\{|\phi(t, u) - \phi(t, u')|;\ u, u' \in U \text{ and } \|u - u'\| \leqq 1/r\}$. Because $\phi(t, \cdot)$ is continuous for each t, the sequence $\{\psi_r\}$ converges to zero in measure. If ε is any

positive number, then

$$\left| \int_0^T \phi(t, u_r(t))\, dt - \int_0^T \phi(t, \bar{u}(t))\, dt \right|$$

(A.2)

$$\leqq 2 \int_{A_l'} \sup_{u \in U} |\phi(t, u)|\, dt + 2 \int_{B_l} \sup_{u \in U} |\phi(t, u)|\, dt + \int_C |\phi(t, u_r(t)) - \phi(t, \bar{u}(t))|\, dt,$$

where $A_l' = \{t; \|u_r(t) - \bar{u}(t)\| > 1/l\}$, $B_l = \{t; \psi_l(t) \geqq \varepsilon/3T\}$ and $C = \{t; |\phi(t, u_r(t)) - \phi(t, \bar{u}(t))| \leqq \varepsilon/3T\}$. Since $\{\psi_r\}$ converges to zero in measure, we can find an integer l such that the second term on the right side of (A.2) is less than $\varepsilon/3$. The first term will then be less than $\varepsilon/3$ whenever r is sufficiently large, because $\{u_r\}$ converges to $\bar{u}$ in measure. Thus the left side of (A.2) is less than ε whenever r is sufficiently large and therefore (A.1) holds. As ϕ was arbitrary this shows that $\{u_r\}$ converges to $\bar{u}$ in the topology of $\hat{\mathcal{U}}$.

(ii) Since we identify functions of $\hat{\mathcal{U}}$ which are equal a.e., an element ρ of $\hat{\mathcal{U}}$ will belong to $\mathcal{U}$ if and only if, for almost all t in $[0, T]$, the measure ρ_t is concentrated at a single atom (which may depend on t, however). This will be true if and only if the variance of the random vector u with respect to the measure $\rho_t(du)$ is zero for almost all t in $[0, T]$. Thus, $\rho \in \mathcal{U}$ if and only if $\int_0^T \int_U \|u' - \int_U u \rho_t(du)\|^2 \rho_t(du')\, dt = 0$. That is, $\rho \in \mathcal{U}$ if and only if ρ satisfies the equation

$$0 = \int_0^T \int_U \|u\|^2 \rho_t(du)\, dt - \sum_{i=1}^{p} \int_0^T \left(\int_U u_i \rho_t(du) \right)^2 dt. \tag{A.3}$$

The functions $\psi_i : \hat{\mathcal{U}} \times \hat{\mathcal{U}} \to \mathbb{R}$ defined by $\psi_i(\rho, \sigma) = \int_0^T \int_U u_i \rho_t(du) \int_U u_i \sigma_t(du)\, dt$ for $i = 1, 2, \cdots, p$, are separately continuous with respect to each of the variables ρ, σ and are therefore Borel measurable by Lemma 3.5. It follows that the function $\psi_i(\rho, \rho)$ is a Borel measurable function of the variable ρ on $\hat{\mathcal{U}}$ for each $i = 1, 2, \cdots, p$. Therefore, the expression on the right-hand side of (A.3) is a Borel measurable function of the variable ρ on $\hat{\mathcal{U}}$ (the first term is in fact continuous). Thus $\mathcal{U}$ is the zero set of a Borel measurable function and is consequently Borel measurable.

Acknowledgment. I would like to thank the reviewer for many helpful comments which have enabled me to simplify the original version of this paper. I am especially grateful for his drawing my attention to the results which are proved in the Appendix. I am also grateful to Z. Artstein who has allowed me to include the previously unpublished proof of Theorem A.1(i).

REFERENCES

[1] T. BASAR AND M. MINTZ, *A Multistage pursuit-evasion game that admits a Gaussian random process as a maximin control policy*, Stochastics, 1 (1973), pp. 25–69.

[2] A. BENSOUSSAN, *Saddle points of convex concave functionals*, Differential Games and Related Topics, H. W. Kuhn and G. P. Szegö, eds., North-Holland, Amsterdam, 1971, pp. 177–199.

[3] L. D. BERKOVITZ, *A differential game with no pure strategy solution*, Annals of Mathematics Studies No. 52, Princeton University Press, Princeton, N.J., 1964, pp. 175–194.

[4] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[5] J. CASE, *Applications of the theory of differential games to economic problems*, Differential Games and Related Topics, H. W. Kuhn and G. P. Szegö, eds., North-Holland, Amsterdam, 1971, pp. 345–371.
[6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators I*, Interscience, New York, 1968.
[7] R. J. ELLIOTT, N. J. KALTON AND L. MARKUS, *Saddle points for linear differential games*, this Journal, 11 (1973), pp. 100–112.
[8] K. FAN, *Minimax theorems*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 42–47.
[9] J. FICHEFET, *Sur l'existence de points de selle dans les jeux différentiels linéaires de durée fixée*, Cahiers Centre Études Recherche Opér., 10 (1968), pp. 5–19.
[10] ———, *Sur l'existence de points de selle dans les jeux différentiels de durée fixée II*, Ibid., 10 (1968), pp. 171–199.
[11] ———, *Un concept de stratégies mixtes pour les jeux différentiels de durée fixée*, Ibid., 11 (1969), pp. 3–25.
[12] W. H. FLEMING, *On a class of games over function space and related variational problems*, Ann. of Math., 60 (1954), pp. 578–594.
[13] A. FRIEDMAN, *Differential Games*, Interscience, New York, 1971.
[14] R. ISAACS, *Differential Games*, John Wiley, New York, 1967.
[15] Y. KAWARA, *An allocation problem of support fire in combat as a differential game*, Operations Res., 21 (1973), pp. 942–951.
[16] N. N. KRASOVSKIY, *On a differential game of rendezvous*, Soviet Math. Dokl., 11 (1970), pp. 921–924.
[17] J. MEDANIC AND M. ANDJELIC, *On a class of differential games without saddle point conditions*, J. Optimization Theory Appl., 8 (1971), pp. 413–430.
[18] K. R. PARTHASARATHY, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.
[19] Z. V. REKASIUS, *On open-loop and closed-loop solutions of linear differential games*, Proc. 1st Internat. Conference on the Theory and Applications of Differential Games, Amherst, Mass., 1969, pp. VII-20–VII-22.
[20] R. C. SCALZO, *N-person linear-quadratic differential games with constraints*, this Journal, 12 (1974), pp. 419–425.
[21] W. E. SCHMITENDORF, *Existence of optimal open-loop strategies for a class of differential games*, J. Optimization Theory Appl., 5 (1970), pp. 363–375.
[22] E. R. SMOL'YAKOV, *Differential games in mixed strategies*, Soviet Math. Dokl., 11 (1970), pp. 330–334.
[23] A. WALD, *Foundation of a general theory of sequential decision functions*, Econometrica, 15 (1947), pp. 279–313.
[24] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[25] D. J. WILSON, *Differential games with no information*, Ph.D. thesis, University of Adelaide, Adelaide, Australia, 1971.
[26] ———, *Mixed strategy solutions for quadratic games*, J. Optimization Theory Appl., 13 (1974), pp. 319–333.
[27] Z. ARTSTEIN, Oral communication.
[28] J. H. MICHAEL AND B. C. RENNIE, *Measurability of Functions of Two Variables*, J. Austral. Math. Soc., 1 (1959), pp. 21–26.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 2, February 1977

A RING OF DELAY OPERATORS WITH APPLICATIONS TO DELAY-DIFFERENTIAL SYSTEMS*

N. S. WILLIAMS† AND V. ZAKIAN‡

Abstract. A ring of delay operators is used to obtain a representation of the solution of systems of linear delay-differential equations. With the aid of this representation an algebraic rank-test is obtained for the R^n-controllability of the systems.

1. Introduction. Let $C([a, b], R^n)$ and $L([a, b], R^n)$ respectively denote the space of all continuous functions and the space of Lebesgue integrable functions which map $[a, b]$ into R^n.

Consider the delay-differential system

$$\frac{dx}{dt} = \sum_{i=0}^{q} A_i x(t-\alpha_i) + \sum_{i=0}^{r} B_i v(t-\beta_i), \tag{1.1a}$$

$$x(t) = \phi(t) \quad \text{for all } t \in [-\alpha_q, 0] \tag{1.1b}$$

where $\phi \in C([-\alpha_q, 0], R^n)$, $v \in L([-\beta_r, t], R^l)$ and the A_i, B_i are real matrices of appropriate dimensions. The delays α_i, β_i are nonnegative real numbers ordered so that $0 \leqq \alpha_0 < \alpha_1 \cdots < \alpha_q$ and $0 \leqq \beta_0 < \beta_1 \cdots < \beta_r$.

Following Zakian and Williams [1], a ring of delay operators is developed in § 2. Use is made of this ring in § 3 to obtain a representation of the solution $x(t)$ of (1.1). With the aid of this representation an algebraic rank test is derived in §§ 4 and 5 for the R^n-controllability of the system (1.1). This test includes the results of Sebakhy and Bayoumi [2] for the case $q = 0$ and the results of Kirrillova and Čurakova [3] for the case $q = 1$ and $r = 0$.

2. A ring of delay operators. Let Ω^n denote a linear space of functions x with domain R and range in R^n, and let T be a linear mapping from Ω^n into Ω^m. The image of x is denoted by Tx, and this is clearly a function with domain R and range in R^m. For any t in the domain of Tx the corresponding image is denoted by $T(x, t)$; for example, if I is the identity transformation in Ω^n then $I(x, t) = x(t)$ for all $x \in \Omega^n$ and all $t \in R$.

Consider the linear transformation $\hat{\alpha}$, called a *delayor*, defined by

$$\hat{\alpha}(x, t) = x(t-\alpha), \qquad x \in \Omega, \quad \alpha \geqq 0, \quad t \in R \tag{2.1}$$

where α is called the *delay*. A *delay operator* $\hat{a}$ is any expression of the form

$$\hat{a} = \sum_{i=0}^{m} a_i \hat{\alpha}_i, \qquad a_i \in R \tag{2.2}$$

* Received by the editors May 25, 1972, and in revised form May 27, 1976.

† Control Systems Centre, University of Manchester Institute of Science and Technology, Manchester, England. Now at The Post Office, Telecommunications Headquarters, London EC1 1AR, England.

‡ Control Systems Centre, University of Manchester Institute of Science and Technology, Manchester M60 1QD, England.

where the $\hat{\alpha}_i$ are delayors. Clearly

$$\hat{a}(x,t)=\sum_{i=0}^{m} a_i x(t-\alpha_i), \qquad x\in\Omega. \tag{2.3}$$

A nonzero delay operator $\sum_{i=0}^{m} a_i\hat{\alpha}_i$ is said to be in *reduced* form if and only if

$$0\leqq\alpha_0<\alpha_1\cdots<\alpha_m \quad \textit{and} \quad a_i\neq 0. \tag{2.4}$$

The reduced form of the zero operator is the zero operator.

It is easy to verify that any delay operator has a unique reduced form.

Two delay operators $\hat{a}$ and $\hat{b}$ are equal if and only if $\hat{a}x=\hat{b}x$ for all $x\in\Omega$. Let $\hat{R}$ denote the set of all delay operators.

Addition and multiplication of two elements $\hat{a},\hat{b}\in\hat{R}$ are defined respectively by

$$(\hat{a}+\hat{b})x=\hat{a}x+\hat{b}x, \qquad x\in\Omega, \tag{2.5}$$

$$(\hat{a}\hat{b})x=\hat{a}(\hat{b}x), \qquad x\in\Omega. \tag{2.6}$$

It follows that if $\hat{a}=\sum_{i=0}^{m} a_i\hat{\alpha}_i$ and $\hat{b}=\sum_{i=0}^{n} b_i\hat{\beta}_i$, then

$$\hat{a}+\hat{b}=\sum_{i=0}^{m} a_i\hat{\alpha}_i+\sum_{i=0}^{n} b_i\hat{\beta}_i, \tag{2.7}$$

$$\hat{a}\hat{b}=\sum_{i=0}^{m}\sum_{j=0}^{n} a_i b_j\widehat{(\alpha_i+\beta_j)}. \tag{2.8}$$

For example, let $\hat{a}=1+\hat{1}$ and $\hat{b}=1-\hat{1}$. Multiplication gives $\hat{a}\hat{b}=-\hat{2}+0+1$, which in reduced form is $1-\hat{2}$.

THEOREM 2.1 (Zakian and Williams [1]). *The set $\hat{R}$, together with the operations of addition and multiplication of elements of $\hat{R}$, form an integral domain.*

Proof. $\hat{R}$ contains R and takes as its unity and zero elements the real numbers 1 and 0. An integral domain is a commutative ring without divisors of zero (see, for example, Archbold [4]). It is easy to verify that $\hat{R}$ satisfies all the properties of a commutative ring. To prove that $\hat{R}$ contains no divisors of zero let $\hat{a}$, $\hat{b}$ be nonzero elements of $\hat{R}$ expressed in reduced form with respective coefficients a_i, b_j. Write the product $\hat{a}\hat{b}$ in reduced form with coefficients c_k. Clearly $c_0=a_0b_0$. Since $a_0\neq 0$, $b_0\neq 0$ it follows that $c_0\neq 0$ and $\hat{a}\hat{b}\neq 0$.

3. A representation of solution of system. Let $\hat{A}$ and $\hat{B}$ denote matrices over $\hat{R}$, expressed in reduced form by $\sum_{i=0}^{q} A_i\hat{\alpha}_i$ and $\sum_{i=0}^{r} B_i\hat{\beta}_i$ where A_i and B_i are the matrices of (1.1). The system (1.1) can now be written in the form

$$\frac{dx}{dt}=\hat{A}(x,t)+\hat{B}(v,t), \tag{3.1a}$$

$$x(t)=\phi(t) \quad \text{for all } t\in[-\alpha_q,0]. \tag{3.1b}$$

Let $\psi(t)=v(t)$ for all $t\in[-\beta_r,0]$ and let $u(t)=v(t)$ for $t>0$. When convenient write $x(t)=x(t;\phi,\psi,u)$ in order to show explicitly the dependence of $x(t)$ on ϕ, ψ and u.

For $t>0$ the solution $x(t)$ of (3.1) is unique (see, for example, Hale [5, pp. 81–85]) and takes the form

$$x(t)=x(t;\phi,0,0)+\int_0^t N(t-\lambda)\hat{B}(v,\lambda)\,d\lambda \tag{3.2}$$

where the $n\times n$ matrix $N(t)$ is uniquely defined by the system

$$\frac{dN}{dt}=\hat{A}(N,t) \quad \text{for } t>0, \tag{3.3a}$$

$$N(t)=I \quad \text{for } t=0, \tag{3.3b}$$

$$N(t)=0 \quad \text{for } t<0. \tag{3.3c}$$

It can be verified from (3.1) and (3.2) that for $t>0$

$$x(t)=x(t;\phi,\psi,0)+\int_0^t M(t-\lambda)u(\lambda)\,d\lambda \tag{3.4}$$

where

$$M(t)=\sum_{i=0}^{r} N(t-\beta_i)B_i, \tag{3.5}$$

that is,

$$M=\sum_{i=0}^{r}\hat{\beta}_i N B_i. \tag{3.6}$$

Now define the system

$$\frac{d\omega}{dt}=\hat{C}(\omega,t) \quad \text{for } t>0, \tag{3.7a}$$

$$\omega(t)=\operatorname{col}(1,0,\cdots,0) \quad \text{for } t=0, \tag{3.7b}$$

$$\omega(t)=0 \quad \text{for } t<0 \tag{3.7c}$$

where $\hat{C}$ is an $n\times n$ matrix over $\hat{R}$ defined by

$$\hat{C}=\left[\begin{array}{c|c} O & -\hat{a}_0 \\ \hline I_{n-1} & \begin{array}{c} -\hat{a}_1 \\ \vdots \\ -\hat{a}_{n-1} \end{array} \end{array}\right] \tag{3.8}$$

and the $\hat{a}_i$ satisfy

$$\det(sI-\hat{A})=s^n+s^{n-1}\hat{a}_{n-1}+\cdots+s\hat{a}_1+\hat{a}_0. \tag{3.9}$$

The main result of this section is the following.

THEOREM 3.1. *Let* $\omega=\operatorname{col}(\omega_0,\omega_1,\cdots,\omega_{n-1})$; *then*

$$N=\sum_{j=0}^{n-1}\hat{A}^j\omega_j \tag{3.10}$$

where N and $\hat{A}$ are as defined in (3.3) *and ω is defined in* (3.7).

Proof. Let p denote the differential operator. Equation (3.7a) implies that

$$p\omega_j = \omega_{j-1} - \hat{a}_j\omega_{n-1} \quad \text{for } j = 1, 2, \cdots, n-1. \tag{3.11}$$

The derivative $p\omega$ is defined by (3.7) at all points in R but not at $t = 0$. Allow that at $t = 0$ the derivative be given the value 0. Therefore $p\hat{A}^j\omega_j$ is defined everywhere in R and moreover

$$p\hat{A}^j\omega_j = \hat{A}^j p\omega_j. \tag{3.12}$$

On combining (3.11) and (3.12) we find that

$$p\sum_{j=1}^{n-1} \hat{A}^j\omega_j = \sum_{j=1}^{n-1} \hat{A}^j(\omega_{j-1} - \hat{a}_j\omega_{n-1}). \tag{3.13}$$

By virtue of (3.9), the Cayley–Hamilton theorem (see, for example, Jacobson [6]) gives:

$$\hat{A}^n = -\sum_{j=0}^{n-1} \hat{a}_j\hat{A}^j, \qquad \hat{A}^0 = I \tag{3.14}$$

and hence (3.13) becomes:

$$p\sum_{j=1}^{n-1} \hat{A}^j\omega_j = \sum_{j=1}^{n} \hat{A}^j\omega_{j-1} + \hat{a}_0\omega_{n-1}I \tag{3.15}$$

$$= \hat{A}K + \hat{a}_0\omega_{n-1}I \tag{3.16}$$

where $K = \sum_{j=0}^{n-1} \hat{A}^j\omega_j$. Therefore

$$pK = \hat{A}K + (p\omega_0 + \hat{a}_0\omega_{n-1})I \tag{3.17}$$

and because of (3.7) one gets

$$\frac{dK}{dt} = \hat{A}(K, t), \qquad t > 0, \tag{3.18a}$$

$$K(t) = I, \qquad t = 0, \tag{3.18b}$$

$$K(t) = 0, \qquad t < 0. \tag{3.18c}$$

But this is the system (3.3) which has a unique solution, and hence $K = N$ and the theorem is proved.

Note that (3.6) and (3.10) give

$$M = \sum_{i=0}^{r}\sum_{j=0}^{n-1} \hat{\beta}_i\hat{A}^j\omega_j B_i = \sum_{j=0}^{n-1} \hat{A}^j\hat{B}\omega_j. \tag{3.19}$$

4. R^n-controllability. The framework developed in the previous sections is used here to derive a test of R^n-controllability of the system (1.1). The question that such a test must answer is whether there is a function, called the *control* and defined as the restriction of v to some interval of the form $(0, \theta]$, such that for a given $x^1 \in R^n$ the condition $x(\theta) = x^1$ holds. It should be noted that the test says nothing about $x(t)$ for $t > \theta$. More precise definitions now follow.

For a given initial pair (ϕ, ψ) and a given point $x^1 \in R^n$ the notation $(\phi, \psi) \to (x^1, \theta)$ means that there is a control $u \in L((0, \theta], R^l)$ such that

$$x(\theta; \phi, \psi, u) = x^1.$$

The system (1.1) is said to be R^n-*controllable* if and only if for every pair (ϕ, ψ) and every x^1 there is a $\theta > 0$ such that the condition $(\phi, \psi) \to (x^1, \theta)$ is satisfied.

The following notation will be required. For any $n \times m$ matrix $\hat{X} = \sum_{k=0}^{\nu} X_k \hat{\gamma}_k$ over $\hat{R}$ let $[\underline{\hat{X}}]$ denote the $n \times m(\nu+1)$ matrix

$$[X_0 \vdots X_1 \vdots \cdots \vdots X_\nu]$$

where the X_k are the coefficients of the reduced form of $\hat{X}$. Let $C(\hat{A}, \hat{B})$ denote the $n \times nl$ matrix over $\hat{R}$

$$[\hat{B} \vdots \hat{A}\hat{B} \vdots \cdots \vdots \hat{A}^{n-1}\hat{B}]$$

where $\hat{A}$ and $\hat{B}$ are as defined in (3.1) and let

$$W(\theta) = \int_0^\theta M(\theta - \lambda) M^T(\theta - \lambda)\, d\lambda \tag{4.1}$$

where M is defined in (3.6) and T indicates transposition.

The following two theorems are the main results of this section.

THEOREM 4.1. (a) $(\phi, \psi) \to (x^1, \theta)$ *if and only if*

$$x^1 - x(\theta; \phi, \psi, 0) \in \text{Range } W(\theta).$$

(b) *For* $\theta > \{(n-1)\alpha_q + \beta_r\}$, $(\phi, \psi) \to (x^1, \theta)$ *if and only if*

$$x^1 - x(\theta; \phi, \psi, 0) \in \text{Range } [\underline{C(\hat{A}, \hat{B})}].$$

THEOREM 4.2. *The system* (1.1) *is* R^n*-controllable if and only if one of the following equivalent conditions is satisfied*:

$$\begin{aligned} &\text{(a)} && \text{rank } W(\theta) = n \quad \textit{for some } \theta > 0, \\ &\text{(b)} && \text{rank } [\underline{C(\hat{A}, \hat{B})}] = n. \end{aligned} \tag{4.2}$$

The proofs of these results are in § 5. Part (a) of Theorem 4.2 was given by Chyung [7]. Notice, however, that a distinctive feature of condition (b) of Theorem 4.2 is that it can be readily evaluated from the matrices $\hat{A}$ and $\hat{B}$ using the operations defined in § 2. A number of interesting sufficient conditions for R^n-controllability are derived from Theorem 4.2(b) in a straightforward manner. For example, since the reduced form of the matrix $\hat{A}^i\hat{B}$ contains the terms

$$A_0^i B_0 \widehat{(j\alpha_o + \beta_0)} \quad \text{and} \quad A_q^i B_r \widehat{(j\alpha_q + \beta_r)}$$

it follows that condition (b) of Theorem 4.2 is satisfied if any of the real matrices $C(A_0, B_0)$, $C(A_q, B_r)$ or $[C(A_0, B_0) \vdots C(A_q, B_r)]$ have rank n.

Consider the following special case of (3.1):

$$\frac{dx}{dt} = A_0 x(t) + \hat{B}(v, t) \tag{4.3}$$

where the reduced form of $\hat{B}$ is $\sum_{i=0}^{r} B_i\hat{\beta}_i$. The reduced form of $C(A_0, \hat{B})$ is $\sum_{i=0}^{r} C(A_0, B_i)\hat{\beta}_i$ and consequently condition (4.2) becomes

$$\text{rank}\,[C(A_0, B_0)\,|\,C(A_0, B_1)\,|\,\cdots\,|\,C(A_0, B_r)] = n. \tag{4.4}$$

This result was recently obtained by Sebakhy and Bayoumi [2].

Another special case of (3.1) is

$$\frac{dx}{dt} = A_0x(t) + A_1x(t-\alpha) + B_0v(t). \tag{4.5}$$

Let Q_i^j be the real $n \times l$ matrices defined by

$$Q_1^1 = B_0, \qquad Q_i^j = 0 \quad \text{for } i = 0 \text{ and } i > j$$

and

$$Q_i^{j+1} = A_0Q_i^j + A_1Q_{i-1}^j.$$

It is easy to show by induction that the reduced form of $(A_0 + A_1\hat{\alpha})^jB_0$ is given by

$$\sum_{i=0}^{j} Q_{i+1}^{j+1}(\hat{\alpha})^i.$$

Consequently for the system (4.5) condition (4.2) becomes

$$\text{rank}\,(Q_1^1Q_1^2\cdots Q_1^n|Q_2^2Q_2^3\cdots Q_2^n|\cdots|Q_n^n) = n. \tag{4.6}$$

This was given in Kirrillova and Čurakova [3], see also Weiss [8].

5. Proofs of Theorems 4.1 and 4.2. Before proving Theorems 4.1 and 4.2, some relations between the range spaces of the matrices $W(\theta)$ and $[\underline{C(\hat{A}, \hat{B})}]$ are established.

LEMMA 5.1. *For any* $\theta > 0$,

$$\text{Range } W(\theta) \subseteq \text{Range } [\underline{C(\hat{A}, \hat{B})}].$$

Proof. Let $X = \text{Range } [\underline{C(\hat{A}, \hat{B})}]$. If $\lambda \notin X$, there are two orthogonal vectors $\lambda_1, \lambda_2 \in R^n$ such that $\lambda = \lambda_1 + \lambda_2$, $\lambda_2 \neq 0$, $\lambda_1 \in X$ and λ_2 is orthogonal to X. Hence

$$\lambda_2^T\hat{A}^j\hat{B} = 0 \quad \text{for } j = 0, 1, 2, \cdots, n-1.$$

Now by (3.19)

$$\lambda_2^TM = \sum_{j=0}^{n-1} \lambda_2^T\hat{A}^j\hat{B}\omega_j = 0$$

and hence $\lambda_2^TW(\theta) = 0$ for any $\theta > 0$. Consequently λ_2 is orthogonal to Range $W(\theta)$ and $\lambda \notin$ Range $W(\theta)$.

LEMMA 5.2. *Let* $\hat{X}$ *and* $\hat{B}$ *be* $n \times n$ *and* $n \times l$ *matrices over* $\hat{R}$ *with the respective reduced forms*

$$\hat{B} = \sum_{i=0}^{r} B_i\hat{\beta}_1, \qquad \hat{X} = \sum_{j=0}^{\nu} X_j\hat{\gamma}_j.$$

If, for $\lambda \in R^n$,

$$\lambda^T \sum_{i=0}^{r} \hat{X}(N, t-\beta_i)B_i = 0 \tag{5.1}$$

for $t \in [0, \theta]$ *and* $\theta > (\beta_r + \gamma_\nu)$, *then*

$$\lambda^T \hat{X}\hat{B} = 0.$$

Proof. Note that $\hat{X}\hat{B} = \sum_{i=0}^{r} \sum_{j=0}^{\nu} X_j B_i \widehat{(\beta_i + \gamma_j)}$. Let the reduced form of $\hat{X}\hat{B}$ be

$$\sum_{l=0}^{\mu} Y_l \hat{\delta}_l, \tag{5.2}$$

that is, $Y_l \neq 0$ and $0 \leqq \delta_0 < \delta_1 < \cdots < \delta_\mu$. Consider the set of couples $c(l) = \{(i, j) | (\gamma_j + \beta_i) = \delta_l\}$. Then Y_l is given by

$$Y_l = \sum_{c(l)} X_j B_i. \tag{5.3}$$

Consequently (5.1) can be written

$$\lambda^T \left\{ \sum_{i=0}^{r} \sum_{j=0}^{\nu} X_j N(t-\beta_i-\gamma_j)B_i \right\} = \lambda^T \sum_{l=0}^{\mu} \left\{ \sum_{c(l)} X_j N(t-\delta_l)B_i \right\} = 0.$$

Hence as $t \to \delta_k^+$

$$\lambda^T \sum_{l=0}^{k} \left\{ \sum_{c(l)} X_j N(\delta_k^+ - \delta_l)B_i \right\} = 0 \tag{5.4}$$

and as $t \to \delta_k^-$

$$\lambda^T \sum_{l=0}^{k} \left\{ \sum_{c(l)} X_j N(\delta_k^- - \delta_l)B_i \right\} = 0. \tag{5.5}$$

Since $N(0^-) = 0$, $N(0^+) = I_n$ and $N(t)$ is continuous for $t > 0$, on subtracting (5.5) from (5.4) we find that

$$\lambda^T \left\{ \sum_{c(k)} (X_j N(0^+)B_i - X_j N(0^-)B_i) \right\} = 0,$$

which with (5.3) implies that $\lambda^T Y_k = 0$. This result holds for $k = 0, 1, \cdots, \mu$ and hence from (5.2) it follows that $\lambda^T \hat{X}\hat{B} = 0$.

LEMMA 5.3. *For any* $\theta > (n-1)\alpha_q + \beta_r$

$$\text{Range}\,[\underline{C(\hat{A}, \hat{B})}] \subseteq \text{Range}\, W(\theta).$$

Proof. Let $\lambda \in R^n$, $\lambda \neq 0$ and $\lambda \notin$ Range $W(\theta)$. Then since $W(\theta)$ is a real symmetric matrix, $\lambda = \lambda_1 + \lambda_2$ where $\lambda_1 \in$ Range $W(\theta)$ and $\lambda_2 \in$ Null $W(\theta)$. Now

$$\int_0^\theta \|\lambda_2^T M(\theta - \sigma)\|^2 \, d\sigma = \lambda_2^T W(\theta) \lambda_2 = 0 \tag{5.6}$$

where $\|\cdot\|$ denotes the Euclidean norm.

Since the integrand in (5.6) is nonnegative and M is a piecewise continuous function, it follows that

$$\lambda_2^T M(t) = 0 \tag{5.7}$$

for $t \in [0, \theta]$. Hence by (3.5)

$$\lambda_2^T \sum_{i=0}^{r} N(t-\beta_i) B_i = 0 \tag{5.8}$$

for $t \in [0, \theta]$. Differentiating (5.8) j times gives

$$\lambda_2^T \sum_{i=0}^{r} \hat{A}^j(N, t-\beta_i) B_i = 0$$

for $t \in [0, \theta]$ and hence, provided $\theta > ((n-1)\alpha_q + \beta_r)$, Lemma 5.2 gives

$$\lambda_2^T \hat{A}^j \hat{B} = 0$$

for $j = 0; 1, 2, \cdots, n-1$. It follows that $\lambda_2^T C(\hat{A}, \hat{B}) = 0$ and that $\lambda \notin \text{Range}\,[\underline{C(\hat{A}, \hat{B})}]$.

Proof of Theorem 4.1. (a) To prove sufficiency, let $\eta \in R^n$ be given by

$$x^1 - x(\theta; \phi, \psi, 0) = W(\theta)\eta;$$

then the control

$$u(\sigma) = \left(\sum_{i=0}^{r} N(\theta - \sigma - \beta_i) B_i \right)^T \eta, \tag{5.9}$$

which is in $L((0, \theta], R^l)$, will transfer (ϕ, ψ) to (x^1, θ), as can be seen by substituting (5.9) into (3.4).

To prove necessity let $\lambda = x^1 - x(\theta; \phi, \psi, 0)$. The hypothesis is $(\phi, \psi) \to (x^1, \theta)$ and implies that $(0, 0) \to (\lambda, \theta)$. Since $W(\theta)$ is a real symmetric matrix there is a $\lambda_1 \in \text{Range}\ W(\theta)$ and a $\lambda_2 \in \text{Null}\ W(\theta)$ such that $\lambda = \lambda_1 + \lambda_2$. Hence following the proof of Lemma 5.3, $\lambda_2^T M(t) = 0$ for $t \in [0, \theta]$. Since $\lambda_1 \in \text{Range}\ W(\theta)$ then, by the sufficiency part of the theorem, $(0, 0) \to (\lambda_1, \theta)$. By virtue of linearity, $(0, 0) \to (\lambda, \theta)$ and $(0, 0) \to (\lambda_1, \theta)$ imply $(0, 0) \to (\lambda_2, \theta)$. Therefore there is a $u \in L([0, \theta], R^l)$ such that

$$\lambda_2 = \int_0^{\theta} M(\theta - \lambda) u(\lambda)\, d\lambda. \tag{5.10}$$

Since $\lambda_2^T M(t) = 0$ for $t \in [0, \theta]$ it follows that $\lambda_2^T \lambda_2 = 0$. This implies that $\lambda_2 = 0$ and hence $\lambda \in \text{Range}\ W(\theta)$.

(b) From Lemmas 5.1 and 5.3 it is clear that for $\theta > ((n-1)\alpha_q + \beta_r)$

$$\text{Range}\ W(\theta) \equiv \text{Range}\,[\underline{C(\hat{A}, \hat{B})}]. \tag{5.11}$$

Proof of Theorem 4.2. (a) To prove sufficiency let rank $W(\theta) = n$ for some $\theta > 0$. Then Range $W(\theta) = R^n$ and from Theorem 4.1(a), $(\phi, \psi) \to (x^1, \theta)$ for all choices of ϕ, ψ and x^1. Consequently the system (1.1) is R^n-controllable.

To prove necessity, suppose to the contrary that rank $W(\theta) < n$ for all $\theta > 0$. Since by (5.11) $W(\theta)$ has the same range space for all $\theta > ((n-1)\alpha_q + \beta_r)$, it follows that there is $\lambda \in R^n$, such that $\lambda \in \text{Null } W(\theta)$ for all sufficiently large θ. Hence $\lambda^T M(t) = 0$ for all t and $\lambda \in \text{Null } W(\theta)$ for all $\theta > 0$. Consequently $(0, 0)$ cannot be transferred to (λ, θ) for any $\theta > 0$ and the system (1.1) is not R^n-controllable.

(b) The proof of this part of the theorem follows from part (a) and Lemmas 5.1 and 5.3 in a straightforward manner.

6. Conclusions. The ring of delay operators has proved useful in the study of delay-differential systems. It provides a compact notation which exposes algebraic features of the systems. Also it serves as a heuristic device for suggesting possible results analogous to those of ordinary differential systems. The main results of §§ 3 and 4 were in fact suggested in this way, although it should be stressed that the concept of R^n-controllability of delay-differential systems is only partly analogous to that of ordinary-differential systems, since it is not concerned with what happens to $x(t)$ at $t > \theta$. Nevertheless R^n-controllability is a useful concept in certain contexts and the test given in § 4 is readily applicable and more general than hitherto known tests.

It appears that the recent results of Manitius and Olbrot [9] and Zmood [10] can be derived by the techniques of this paper.

REFERENCES

[1] V. ZAKIAN AND N. S. WILLIAMS (1972), *Algebraic treatment of delay differential systems*, Control Systems Centre Rep. 183, Univ. of Manchester Inst. of Sci. and Tech., Manchester, England.

[2] O. SEBAKHY AND M. M. BAYOUMI (1973), *Controllability of linear time-varying systems with delay in control*, Internat. J. Control, 17, pp. 127–135.

[3] F. M. KIRRILLOVA AND S. V. ČURAKOVA (1967), *The problem of controllability of linear systems with after affect*, Differencial'nye Uravnenija, 3, pp. 436–445.

[4] J. W. ARCHBOLD (1964), *Algebra*, Isaac Pitman, London.

[5] J. HALE (1971), *Functional Differential Equations*, Springer-Verlag, New York.

[6] N. JACOBSON (1951, 1953), *Lectures in Abstract Algebra*, vols. I, II, Van Nostrand, Princeton, NJ.

[7] D. H. CHYUNG (1971), *Controllability of linear time-varying systems with delays*, IEEE Trans. Automatic Control, AC-16, pp. 493–494.

[8] L. WEISS (1970), *An algebraic criterion for controllability of linear systems with time delay*, Ibid., AC-15, pp. 443–444.

[9] A. MANITIUS AND A. W. OLBROT (1972), *Controllability conditions for linear systems with delayed state and control*, Arch. Automat. i. Telemech., 17, pp. 119–131.

[10] R. B. ZMOOD (1974), *The Euclidean space controllability of control systems with delay*, this Journal, 12, pp. 609–623.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 2, February 1977

THE HIGH ORDER MAXIMAL PRINCIPLE AND ITS APPLICATION TO SINGULAR EXTREMALS*

ARTHUR J. KRENER†

Abstract. The high order maximal principle (HMP) which was announced in [11] is a generalization of the familiar Pontryagin maximal principle. By using the higher derivatives of a large class of control variations, one is able to construct new necessary conditions for optimal control problems with or without terminal constraints. In particular, we show how the HMP can be used to prove the generalized Legendre–Clebsch condition of Kelley, Kopp, Moyer and Goh. The principle advantage of this derivation is that, unlike previous ones, it remains valid even when there are terminal constraints.

1. Introduction. Although we are interested in high order necessary conditions for optimal control problems, let us first consider the following nonlinear programming problem. Minimize the smooth function $y_0(x)$ subject to the smooth constraints $y_i(x)=0$ for $i=1,\cdots,m$ and $x\in\mathcal{A}\subseteq\mathbb{R}^n$. The set $\mathcal{A}$ is not explicitly described, instead, given $x^e\in\mathcal{A}$ we assume there are ways of generating smooth curves $s\mapsto x(s)\in\mathcal{A}$ for $s\in[0,\varepsilon)$ such that $x(0)=x^e$. To develop first order necessary conditions for this problem we adjoin the constraints y_i to y_0 via Lagrange multipliers $\nu_0,\nu_1,\cdots,\nu_m$, where ν_0 is normalized to be nonpositive. If x^e is a minimum, then every curve $x(s)$ as above generates a necessary condition

$$\frac{d}{ds}\sum_{i=0}^{m}\nu_i y_i(x(0))=\sum_{i=0}^{m}\nu_i\frac{\partial}{\partial x}y_i(x^e)\frac{d}{ds}x(0)\leqq 0. \tag{1.1}$$

The use of the Lagrange multipliers requires some assumption of local convexity on the set $\{x: y_i(x)=0,\ i=1,\cdots,m\}\cap\mathcal{A}$ around x^e. Since $\mathcal{A}$ is not explicitly given, this cannot be verified. Instead we assume the following: the gradients of the functions $y_0,\cdots,y_m$ are linearly independent at x^e and whenever $x^1(s)$ and $x^2(s)$ are used to develop necessary conditions via (1.1), for any $0\leqq\hbar\leqq 1$ there exists a curve $x^3(s)\in\mathcal{A}$ such that $x^3(0)=x^e$ and

$$\frac{d}{ds}x^3(0)=\mu\frac{d}{ds}x^1(0)+(1-\mu)\frac{d}{ds}x^2(0). \tag{1.2}$$

As we shall see later, this form of convexity suffices to justify the multipliers. Of course if $m=0$, no convexity assumption or multipliers are needed and ν_0 can be set to be -1.

The goal of any collection of necessary conditions is to isolate a hopefully unique candidate for the minimum. Additional conditions may be required to narrow the field of possibilities and to distinguish between potential maxima and minima. If a collection of necessary conditions of the form (1.1) does not

* Received by the editors October 14, 1975, and in revised form February 10, 1976.

† Department of Mathematics, University of California, Davis, California 95616. This research was supported in part by the U.S. Office of Naval Research under the Joint Services Electronics Program Contract N00014-75-C-0648, while the author was a research fellow at the Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts, 1974–75, and also by the National Science Foundation under Grant MPS75-05248.

completely accomplish this task, one can look for additional curves $x(s)$ or obtain higher order conditions by differentiating (1.1) further.

If there are no y_i constraints ($m = 0$), then it is clear that the first nonzero derivative of $\nu_0 y_0(x(s))$ must be negative for x^e to be a minimum. In general this involves higher order partial derivatives of $y_0(x)$. For example, if (1.1) is assumed to be zero, then the second derivative test is

$$\nu_0 \frac{d^2}{ds^2} y_0(x(0)) = \nu_0 \left(\frac{d}{ds} x(0)\right)^T \frac{\partial^2}{\partial x^2} y_0(x^e) \left(\frac{d}{ds} x(0)\right) + \nu_0 \frac{\partial}{\partial x} y_0(x^e) \frac{d^2}{ds^2} x(0) \leqq 0. \tag{1.3}$$

Suppose $x^e \in \text{interior } \mathscr{A} \subseteq \mathbb{R}^n$. Then (1.1) implies that $(\partial/\partial x) y_0(x^e) = 0$ and so (1.3) reduces to the familiar condition where the Hessian $\nu_0(\partial^2/\partial x^2) y_0(x^e)$ is negative semidefinite at a minimum. On the other hand, if for some $x(s)$, $(d/ds)x(0) = 0$, then (1.1) is trivially satisfied and (1.3) yields a condition which involves only the gradient of y_0. The same condition can be obtained from (1.1) by reparametrizing $x(s)$ as $x(s^{1/2})$.

If there are terminal constraints, then second order conditions similar to (1.3) can be developed with some difficulty, since the use of the Lagrange multiplier must be justified. For higher derivatives, this justification is so difficult as to make the resulting necessary conditions of little practical value. The difficulties arise because, in general, these conditions involve second and higher order partial derivatives of $y_0, y_1, \cdots, y_m$. As was seen above, there is an exception to that; if the first $h-1$ derivatives of $x(s)$ are zero at $s = 0$, then

$$\frac{d^h}{ds^h} \sum_{i=0}^{m} \nu_i y_i(x(0)) = \sum_{i=0}^{m} \nu_i \frac{\partial}{\partial x} y_i(x^e) \frac{d^h}{ds^h} x(0) \leqq 0 \tag{1.4}$$

involves only the first partial derivatives of $y_0, \cdots, y_m$. In this case, justifying the Lagrange multiplier requires only a convexity assumption for higher derivatives similar to (1.2). It is this type of necessary condition which we consider in this paper.

Now we turn to optimal control problems which generate nonlinear programming problems of the type we have been considering. Suppose we wish to minimize $y_0(x(t^e))$ subject to $\dot{x} = f(x(t), u(t))$, $x(t^0) = x^0$, $y_i(x(t^e)) = 0$, $i = 1, \cdots, m$, and $u(t) \in \Omega$ for $t \in [t^0, t^e]$. Let $\mathscr{A}$ denote the set of points accessible from x^0 using admissible controls. Suppose a control $u(t)$ and trajectory $x(t)$ defined on $[t^0, t^e]$ is a candidate for an optimal solution. We can generate curves lying in $\mathscr{A}$ by considering the locus of endpoints $x(t^e; s)$ of a family of trajectories $x(t; s)$ generated by controls $u(t; s)$ which are variations of $x(t)$ and $u(t)$ depending on the parameter s. The controls $u(t; s)$ are obtained by replacing $u(t)$ by some other control $v(t)$ for $t \in [t^1 - s, t^1]$ where $t^1 \in (t^0, t^e)$. The reference control and trajectory are obtained when $s = 0$. In this way, using (1.1), one develops the usual linear necessary conditions, i.e., the Pontryagin maximum principle (PMP), which is most conveniently expressed in a Hamiltonian format.

It frequently happens in nonlinear control problems that the set of first derivatives of the curves obtained by the above procedure does not fully represent

all the degrees of freedom within the set $\mathcal{A}$ of accessible points around the reference endpoint $x(t^e)$. Such controls and trajectories are called *singular* (in the sense of the PMP as opposed to the classical definition in the calculus of variations). For this reason the PMP can prove to be inadequate in determining either a unique candidate or distinguishing between minimizing and maximizing trajectories. (See [2].)

The high order maximum principle [HMP) is an attempt to overcome these difficulties. More complicated control variations are used which have the property that lower order derivatives of $x(t^e; s)$ are zero and the first nonzero derivatives lie in directions within $\mathcal{A}$ which were not available as first derivatives. Since the lower derivatives are zero and a convexity assumption for higher derivatives similar to (1.2) is satisfied, equation (1.4) can be applied to obtain new necessary conditions which can also be expressed in terms of the Hamiltonian.

The organization of the rest of the paper is as follows. The statement of the HMP is found in § 2 and the proof in § 3. Then the HMP is used to develop linear and quadratic necessary conditions for singular extremals. Scalar controls are treated in §§ 4 and 5, and in § 6, vector controls are treated. (These conditions are called linear and quadratic not because they are linear or quadratic with respect to the parameter s mentioned above, but rather because they are linear or quadratic with respect to the L^1 norm of the control variation. We elaborate on this later.)

The linear conditions are those implied by the PMP. The quadratic conditions reduce to the generalized Legendre–Clebsch (GLC) of Kelley, Kopp and Moyer [8] (scalar control) and Goh [4] (vector controls) when the problem in question is normal or there are no terminal constraints. Using the HMP we can extend the GLC to problems which do not satisfy these assumptions.

We wish to emphasize that these are not the only applications of the HMP, rather, the HMP is a very powerful tool for constructing necessary conditions, the simplest of which are the ones mentioned above. We hope that by studying this paper the reader will be able to construct new necessary conditions in an ad hoc fashion which are appropriate to the problem of interest.

2. The high order maximal principle. Consider a system whose dynamics are given by

$$\dot{x} = f(x, u) \tag{2.1}$$

subject to $x(t^0) = x^0$ and $u(t) \in \Omega$, where $x = (x_0, x_1, \cdots, x_n)$ with $x_0 = t$, $u = (u_1, \cdots, u_l)$, f a C^∞-function of x and u, Ω some subset of $\mathbb{R}^l$. The state variables x are local coordinates on an $(n+1)$-dimensional C^∞-manifold M. However we proceed as if they are globally defined and leave to the reader the task of "patching things together", i.e., supplying the intrinsic meaning for all of the objects described in a coordinate-dependent fashion.

The problem is to find a piecewise C^∞-control $u(t) \in \Omega$ for $t \in [t^0, t^e]$ which generates a trajectory $x(t)$ satisfying the boundary conditions

$$x(t^0) = x^0 \quad \text{and} \quad y_i(x(t^e)) = 0, \qquad i = 1, \cdots, m, \tag{2.2}$$

which minimizes

$$y_0(x(t^e)). \tag{2.3}$$

The functions $y_0, y_1, \cdots, y_m$ are assumed to be C^∞ and linearly independent everywhere of interest. Since time is a state variable, (2.1) could be time-dependent and the functions $y_0, \cdots, y_m$ could also depend on time. Control problems where the integral of a Lagrangian are to be minimized can easily be converted to the above format by the addition of another state variable.

The assumption of infinite differentiability is not required, it is only invoked to avoid counting the degree of differentiability needed in a particular argument. Piecewise differentiability means left and right limits always exist and there are only a finite number of jumps in any compact interval. Throughout the paper we assume that the controls being considered are C^∞ at the times in question. At other times, similar results can be deduced by restricting to left or right limits and by continuity. Since the details are tedious, we choose the convenient expedient of leaving them to the reader.

Corresponding to each *admissible control*, $u^i(t) \in \Omega$, is an *admissible vector field*

$$f^i(x) = f(x, u^i(x_0))$$

which generates an *admissible flow* $\gamma^i(s)x$ defined as the family of integral curves of the differential equation

$$\frac{d}{ds}\gamma^i(s)x = f^i(\gamma^i(s)x)$$

satisfying the initial conditions

$$\gamma^i(0)x = x.$$

Suppose the reference trajectory $x(t) = \gamma^0(t-t^0)x^0$ is generated by the control $u^0(t)$ for $t \in [t^0, t^e]$. Then a standard proof of the PMP is to replace the reference control by another control $u^1(t)$ for $t \in [t^1 - s, t^1]$ where $t^1 \in (t^0, t^e)$. The result is a family of trajectories $x(t; s)$ indexed by small $s \geqq 0$ whose locus of endpoints is given by

$$\begin{aligned} x(t^e; s) &= \gamma^0(t^e - t^1)\gamma^1(s)\gamma^0(t^1 - t^0 - s)x^0 \\ &= \gamma^0(t^e - t^1)\gamma^1(s)\gamma^0(-s)x^1, \end{aligned}$$

where $x^1 = x(t^1)$. If we define $\alpha(s)x = \gamma^1(s)\gamma^0(-s)x$, then this can be written as

$$\gamma^0(t^e - t^1)\alpha(s)x^1.$$

For this reason, we call the map $\alpha(s)x$ a *control variation* to u^0 before x.

Alternately, $u^0(t)$ could be replaced by $u^1(t)$ on the interval $[t^1, t^1 + s]$ resulting in a locus of endpoints

$$\begin{aligned} x(t^e; s) &= \gamma^0(t^e - t^1 - s)\gamma^1(s)x^1 \\ &= \gamma^0(t^e - t^1)\gamma^0(-s)\gamma^1(s)x^1. \end{aligned}$$

This time we have a control variation $\alpha(s)x = \gamma^0(-s)\gamma^1(s)x$ to u^0 after x. Various combinations of the above are possible, for example, $\alpha(s)x = \gamma^0(-s/2)\gamma^1(s)\gamma^0(-s/2)x$, a control variation to u^0 at x. As we shall see in § 3, if

$u^0(t)$ and $u^1(t)$ are smooth at t^1, then all of the above yield the same necessary conditions.

The important point about these variations $\alpha(s)x$ is that when they are inserted into a trajectory generated by the control $u^0(t)$, the result is a family of admissible trajectories indexed by small $s \geqq 0$ whose locus of endpoints is a smooth function of s. With this in mind we define a *control variation* $\alpha(s)x$ to $u^0(t)$ at x as being of the following form:

$$\alpha(s)x = \gamma^0(q_2(s))\gamma^k(p_k(s)) \cdots \gamma^1(p_1(s))\gamma^0(q_1(s))x, \tag{2.4}$$

where $\gamma^0, \gamma^1, \cdots, \gamma^k$ are the flows of admissible controls $u^0(t), u^1(t), \cdots, u^k(t)$ and $q_i(s)$ and $p_i(s)$ are polynomials in s satisfying $q_i(0) = p_i(0) = 0$ and $p_i(s) \geqq 0$ for small $s \geqq 0$. This is similar to the bundle variation of Gabasov and Kirillova [2]. The reader should note that

$$x(t^e; s) = \gamma^0(t^e - t^1)\alpha(s)\gamma^0(t^1 - t^0)x^0$$

is the locus of endpoints of a family of admissible trajectory for small $s \geqq 0$, and hence a curve in $\mathscr{A}$. Moreover, $x(t^e; s)$ is a smooth function of s and $x(t^e; 0)$ is the endpoint of the reference trajectory. Notice that if $q_1(s) + q_2(s) + \sum p_i(s) \neq 0$, then the control variation changes the terminal time, t^e. In particular, the variations $\alpha(s)x = \gamma^0(\pm s)x$ lengthen or shorten the reference trajectory.

A control variation $\alpha(s)x$ is said to be of *order h* at $x^1 = x(t^1)$ if there exists an $\varepsilon > 0$ such that

$$\frac{d^j}{ds^j}\alpha(0)x(t) = 0 \tag{2.5}$$

for $j = 1, \cdots, h-1$ and $|t - t^1| < \varepsilon$. In particular, for $h = 1$, there are no lower derivatives for which (2.5) must hold and so every variation is of order at least one. A control variation of order h is a fortiori of order $1, \cdots, h-1$. Because the earlier derivatives are zero, it is the h derivative of $x(t^e; s)$ which supplies the necessary condition via (1.4). As we show in the next section, it is necessary to require (2.5) to hold in a time interval around t^1 so that the convexity assumption for higher derivatives holds and the use of multipliers can be justified.

The high order maximal principle (HMP). Let $u^0(t)$ be an admissible control generating the trajectory $x(t) = \gamma^0(t - t^0)x^0$ for $t \in [t^0, t^e]$. If $u^0(t)$ minimizes $y_0(x(t^e))$ subject to the boundary condition $y_i(x(t^e)) = 0$ for $i = 1, \cdots, m$, then there exists a nontrivial adjoint variable $\lambda(t) = (\lambda_0(t), \cdots, \lambda_n(t))$ defined for $t \in [t^0, t^e]$ and satisfying

$$\lambda(t) = -\lambda(t)\frac{\partial}{\partial x} f(x(t), u^0(t)), \tag{2.6}$$

$$\lambda(t^e) = \sum \nu_i \frac{\partial}{\partial x} y_i(x(t^e)), \quad \text{where } \nu_0 \leqq 0, \tag{2.7}$$

$$\lambda(t)f(x(t), u) \leqq \lambda(t)f(x(t), u^0(t)) = 0 \quad \forall\, u \in \Omega, \tag{2.8}$$

and for every control variation $\alpha(s)x$ of order h at $x(t)$,

$$\lambda(t)\frac{d^h}{ds^h}\alpha(0)x(t) \leqq 0. \tag{2.9}$$

Conditions (2.6), (2.7) and (2.8) are the familiar PMP and (2.1), (2.6) and (2.8) can be conveniently expressed in terms of the Hamiltonian, $H(\lambda, x, u) = \lambda f(x, u)$, as Hamilton's differential equation

$$\dot{x} = \frac{\partial}{\partial \lambda} H(\lambda(t), x(t), u^0(t)), \tag{2.10}$$

$$\lambda = -\frac{\partial}{\partial x} H(\lambda(t), x(t), u^0(t)) \tag{2.11}$$

and the Pontryagin–Weierstrass condition,

$$0 = H(\lambda(t), x(t), u^0(t)) = \max_{u \in \Omega} H(\lambda(t), x(t), u). \tag{2.12}$$

(Equations (2.8) and (2.12) are zero because $x_0 = t$.) A $u^0(t)$ and $x(t)$ for which there exists a $\lambda(t)$ satisfying (2.6) and (2.8) are called an *extremal control and extremal trajectory*. If $u^0(t) \in$ interior Ω, then (2.12) implies

$$\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) = 0 \tag{2.13}$$

and the Legendre–Clebsch condition

$$\frac{\partial^2}{\partial u^2} H(\lambda(t), x(t), u^0(t)) \leqq 0. \tag{2.14}$$

The new condition (2.9) is a generalization of the Pontryagin–Weierstrass condition to control variations of higher order. It in turn leads to a generalization of the Legendre–Clebsch condition for extremal trajectories which are singular in the classical sense (i.e., (2.14) not of full rank). This is demonstrated in §§ 5 and 6.

We have fixed the initial point $x(t^0) = x^0$ but there is a straightforward extension of the HMP to problems where the initial point is only partially constrained. In this case (2.6)–(2.9) still hold and, in addition, $\lambda(t^0)$ must satisfy a transversality condition similar to (2.7) as in the PMP.

When applying the HMP it is highly desirable to choose a minimal realization of the problem under consideration where $y(x) = (y_0(x), \cdots, y_m(x))$ is considered as the output map. (For the theory of minimal realizations of nonlinear systems, see Sussmann [16].) The reason for this is that the less state dimensions there are, the easier it is to find higher order control variations satisfying (2.5), and so the more necessary conditions result. An example of just this point is given in § 5.

3. Proof of the HMP. We start by noting that the order of a control variation can easily be shifted upward.

LEMMA 3.1. *Suppose $\alpha^1(s)x$ is a control variation of order h at x^1. Then for any integer k there exists a control variation $\alpha^2(s)x$ of order $h \cdot k$ at x^1 whose $h \cdot k$ derivative is a positive multiple of the h derivative of $\alpha^1(s)x$ and hence yields the same necessary condition when used in* (2.9) *of the HMP.*

Proof. Define $\alpha^2(s)x = \alpha^1(s^k)x$. It is straightforward to verify that $\alpha^2(s)x$ is a control variation as in (2.4) and of order $h \cdot k$ with the $h \cdot k$ derivative as described.

To compute higher derivatives of control variations, it is convenient to let them operate on smooth functions as a partial differential operator.

LEMMA 3.2. *A necessary and sufficient condition for $\alpha(s)x$ to be a control variation of order h at x^1 is that for some $\varepsilon > 0$ and for every C^∞ real-valued function $\varphi(x)$ defined in some neighborhood of x^1,*

$$\frac{d^j}{ds^j}\varphi(\alpha(0)x(t)) = 0$$

for $j = 1, \cdots, h-1$ and $|t - t^1| < \varepsilon$. Moreover if $\alpha(s)x$ is of order h at x^1,

$$\frac{d^h}{ds^h}\varphi(\alpha(0)x^1) = \frac{\partial}{\partial x}\varphi(x^1)\frac{d^h}{ds^h}\alpha(0)x^1.$$

Proof. The proof is straightforward.

We next show that if all the controls involved are C^∞ at t^1, then it does not matter whether a variation is made before, after or at $x^1 = x(t^1)$.

LEMMA 3.3. *Let $\alpha^1(s)x$ be a control variation to $u^0(t)$ of order h at x^1 and $q(s)$ be a polynomial in s such that $q(0) = 0$. Define a new control variation $\alpha^2(s)x = \gamma^0(q(s))\alpha^1(s)\gamma^0(-q(s))x$. Then α^2 is also of order h at x^1 and furthermore yields the same necessary condition in* (2.9) *for*

$$\frac{d^h}{ds^h}\alpha^2(0)x^1 = \frac{d^h}{ds^h}\alpha^1(0)x^1.$$

Proof. Consider $\alpha^2(s)x$ as a function of four variables,

$$\alpha^2(s_1, s_2, s_3)x = \gamma^0(q(s_3))\alpha^1(s_2)\gamma^0(-q(s_1))x,$$

where $s_1 = s_2 = s_3 = s$. Then for any C^∞-function, by the chain rule,

$$\frac{d^l}{ds^l}\varphi(\alpha^2(0)x) = \sum \binom{l}{i, j, k} \frac{\partial^i}{\partial s_1^i}\frac{\partial^j}{\partial s_2^j}\frac{\partial^k}{\partial s_3^k}\varphi(\alpha^2(0, 0, 0)x), \tag{3.1}$$

where the sum is over all $i, j, k \geqq 0$, $i + j + k = l$. For any k define a C^∞-function,

$$\psi(x) = \frac{\partial^k}{\partial s_3^k}\varphi(\gamma^0(q(0))x).$$

For small s_1 there exists a $t(s_1)$ near t^1 such that

$$\gamma^0(-q(s_1))x^1 = x(t(s_1)).$$

Since $\alpha^1(s)x$ is of order h at x^1, for $1 \leqq j \leqq h-1$ and small s_1,

$$\frac{\partial^j}{\partial s_2^j}\frac{\partial^k}{\partial s_3^k}\varphi(\alpha^2(s_1, 0, 0)) = \frac{\partial^j}{\partial s_2^j}\psi(\alpha^1(0)\gamma^0(-q(s_1))x^1 = \frac{\partial^j}{\partial s_2^j}\psi(\alpha^1(0)x(t(s_1))) = 0.$$

So for $1 \leqq l \leqq h-1$, equation (3.1) becomes

$$\frac{d^l}{ds^l}\varphi(\alpha^2(0)x^1) = \sum_{i=0}^{l}\binom{l}{i}\frac{\partial^i}{\partial s_1^i}\frac{\partial^{l-i}}{\partial s_3^{l-i}}\varphi(\alpha(0,0,0)x^1)$$

$$= \frac{d^l}{ds^l}\varphi(\gamma^0(q(0))\gamma^0(-q(0))x^1)$$

$$= \frac{d^l}{ds^l}\varphi(x^1) = 0,$$

since $\gamma^0(q(s))\gamma^0(-q(s))x^1 = \gamma^0(q(s)-q(s))x^1 = \gamma^0(0)x^1 = x^1$. The same arguments can be repeated at each $x(t)$, $|t-t^1| < \varepsilon$ to show $\alpha^2(s)x$ satisfies (2.5) in an interval around t^1.

Similarly evaluating (3.1) for $l = h$, we have

$$\frac{d^h}{ds^h}\varphi(\alpha^2(0)x^1) = \frac{d^h}{ds^h}\varphi(\alpha^1(0)x^1) + \frac{d^h}{ds^h}\varphi(\gamma^0(q(0))\gamma^0(-q(0))x^1)$$

$$= \frac{d^h}{ds^h}\varphi(\alpha^1(0)x^1). \qquad \text{Q.E.D.}$$

If the control $u^0(t)$ is not continuous at t^1, then the trajectory has a corner at x^1 and the effect of control variations on either side are different. By comparing these differences one can deduce various corner conditions for optimality, but this is a topic we shall not pursue any further. We refer the interested reader to Kelley, Kopp and Moyer [8], Gabasov and Kirillova [2], McDanell and Powers [12] and Maurer [13].

The next two lemmas are crucial to the HMP because they show that for higher order control variations satisfying (2.5), one can "add" them and, in particular, form convex combinations as required by the use of Lagrange multipliers. First we deal with control variations made at the same point of the trajectory.

LEMMA 3.4. *Suppose $\alpha^1(s)x, \cdots, \alpha^r(s)x$ are control variations to $u^0(t)$ at $x^1 = x(t^1)$ of order $h_1, \cdots, h_r$ respectively. Let $c = (c_1, \cdots, c_r)$ be a vector of nonnegative real numbers. Then there exists a family of control variations $\alpha(s; c)x$ of order h (= the least common multiple $\{h_i\}$) such that*

$$\frac{d^h}{ds^h}\alpha(0; c)x^1 = \sum_{i=1}^{r} c_i \frac{d^{h_i}}{ds^{h_i}}\alpha^i(0)x^1.$$

Moreover $\alpha(s, c)x$ is continuous in c for small $s \geqq 0$.

Proof. For notational simplicity, assume $r = 2$; the general case follows by a similar argument. Using Lemma 3.1, we can assume that $h = h_1 = h_2$, and using Lemma 3.3, that both variations are made before x^1, for example,

$$\alpha^1(s)x = \gamma^k(p_k(s)) \cdots \gamma^1(p_1(s))\gamma^0(q(s))x.$$

Define a family of new variations

$$\alpha(s; c)x = \gamma^k(p_k(c_1^{1/h}s)) \cdots \gamma^1(p_1(c_1^{1/h}s))\alpha^2(c_2^{1/h}s)\gamma^0(q(c_1^{1/h}s))x.$$

Introduce parameters $s_1 = s_2 = s_3 = s$ into $\alpha(s; c)x$ as in the proof of Lemma 3.3. Then for any C^∞-function φ,

$$\frac{d^l}{ds^l}\varphi(\alpha(0; c)x^1) = \sum \binom{l}{i, j, k} c_1^{i/h} c_2^{j/h} c_1^{k/h} \frac{\partial}{\partial s_1^i} \frac{\partial}{\partial s_2^j} \frac{\partial^k}{\partial s_3^k} \varphi(\alpha(0, 0, 0; c)x^1).$$

If $1 \leqq l \leqq h-1$, this reduces as before to

$$\frac{d^l}{ds^l}\varphi(\alpha(0; c)x^1) = \sum_{i=0}^{l} \binom{l}{i} c_1^l \frac{\partial^i}{\partial s_1^i} \frac{\partial^{l-i}}{\partial s_3^{l-i}} \varphi(\alpha(0, 0, 0; c)x^1)$$

$$= c_1^l \frac{d^l}{ds^l}\varphi(\alpha^1(0)x^1) = 0,$$

since α^1 is of order h at x^1.

For $l = h$ we have

$$\frac{d^h}{ds^h}\varphi(\alpha(0; c)x^1) = c_1 \frac{d^h}{ds^h}\varphi(\alpha^1(0)x^1) + c_2 \frac{d^h}{ds^h}\varphi(\alpha^2(0)x^1). \qquad \text{Q.E.D.}$$

If a control variation $\alpha(s)x$ of order h is made at x^1, then the result is a family of trajectories whose locus of endpoints is given by

$$x(t^e; s) = \gamma^0(t^e - t^1)\alpha(s)x^1.$$

The first $h-1$ derivatives of $x(t^e; s)$ are zero and h derivative is given by

$$\frac{d^h}{ds^h}x(t^e; 0) = \left(\frac{\partial}{\partial x}\gamma^0(t^e - t^1)x^1\right)\frac{d^h}{ds^h}\alpha(0)x^1. \tag{3.2}$$

This is applied to (1.4) to obtain (2.9) of the HMP, but first we must show that we can "add" the effect of control variations made at differing times.

LEMMA 3.5. *Suppose $\alpha^1(s)x, \cdots, \alpha^r(s)x$ are control variations to $u^0(t)$ at $x^1 = x(t^1), \cdots, x^r = x(t^r)$ of order $h_1, \cdots, h_r$ respectively. Let $c = (c_1, \cdots, c_r)$ be a vector of nonnegative real numbers. Then there exists a family of admissible trajectories indexed by small $s \geqq 0$ and c whose locus of endpoints is given by $x(t^e; s; c)$ such that*

$$\frac{d^j}{ds^j}x(t^e; 0; c) = 0$$

for $j = 1, \cdots, h-1$ where $h =$ least common multiple $\{h_i\}$ and

$$\frac{d^h}{ds^h}x(t^e; 0; c) = \sum_{i=1}^{r} c_i\left(\frac{\partial}{\partial x}\gamma^0(t^e - t^i)x^i\right)\frac{d^{h_i}}{ds^{h_i}}\alpha^i(0)x^i. \tag{3.3}$$

Moreover $x(t^e; s; c)$ is continuous in c for small $s \geqq 0$.

Proof. Using Lemma 3.4, we can assume that the t^i are distinct. For simplicity, assume $r = 2$, $t^1 < t^2$ and $h_1 = h_2 = h$. The general case follows by a similar argument. Consider the family of trajectories whose locus of endpoints is given by

$$x(t^e; s; c) = \gamma^0(t^e - t^2)\alpha^2(c_2^{1/h}s)\gamma^0(t^2 - t^1)\alpha^1(c_1^{1/h}s)\gamma^0(t^1 - t^0)x^0.$$

Suppose φ is a C^∞-function at $x^e = x(t^e)$. Then using the chain rule technique,

$$\frac{d^l}{ds^l}\varphi(x(t^e;0;c)) = \sum_{i=0}^{l} \binom{l}{i} c_1^{i/h} c_2^{(l-i)/h} \frac{\partial^i}{\partial s_1^i} \frac{\partial^{l-i}}{\partial s_2^{l-i}} \varphi(x(t^e;0;c)).$$

Let

$$\psi(x) = \frac{\partial^{l-i}}{\partial s_2^{l-i}} \varphi(\gamma^0(t^e - t^2)\alpha^2(c_2^{1/h}0)\gamma^0(t^2 - t^1)x).$$

Then $\psi(x)$ is a C^∞-function at $x(t^1)$. Since α^1 is of order h, for $1 \leqq i < h$,

$$\frac{\partial^i}{\partial s_1^i} \psi(\alpha^1(c_1^{1/h}0)x^1) = 0.$$

Therefore if $1 \leqq l < h$, then

$$\frac{d^l}{ds^l}\varphi(x(t^e;0;c)) = c_2^{l/h} \frac{\partial^l}{\partial s_2^l}(\varphi \circ \gamma^0(t^e - t^2))\alpha^2(c_2^{1/h}0)x^2 = 0$$

since α^2 is of order h. A similar argument proves (3.3). Q.E.D.

In light of this lemma, we define a cone K in the tangent space at $x^e = x(t^e)$ as the convex hull of all vectors of the form (3.2). This cone is a measure of the controllability at x^e available through higher order control variations made all along the reference trajectory. The completion of the proof of the HMP follows Halkin's proof of the PMP [5] using a fixed point argument. Intuitively for $u^0(t)$ to be minimal, the cone K of controllability must be separable by a hyperplane from the cone of L of directions which satisfy the boundary conditions and decrease y_0. Formally L is defined to be the cone of all tangent vectors τ at x^e such that

$$\left(\frac{\partial}{\partial x} y_0(x^e)\right)\tau \leqq 0$$

and for $i = 1, \cdots, m$,

$$\left(\frac{\partial}{\partial x} y_i(x^e)\right)\tau = 0.$$

THEOREM 3.6 (HMP). *Suppose there exists no nontrivial adjoint variable satisfying* (2.6)–(2.9). *Then $u^0(t)$ is not minimal.*

Proof. If $\lambda^e = (\lambda_0^e, \cdots, \lambda_n^e)$ defines a hyperplane separating K and L in the tangent space of x^e, i.e.,

$$\lambda^e \tau \leqq 0 \quad \forall\, \tau \in K,$$

$$\lambda^e \tau \geqq 0 \quad \forall\, \tau \in L,$$

then define

$$\lambda(t) = \lambda^e \frac{\partial}{\partial x} \gamma^0(t^e - t)x^0.$$

It is easy to verify that $\lambda(t)$ satisfies (2.6)–(2.9).

On the other hand, if no such λ^e exists, then it follows that there exists no hyperplane separating K^* and L^* where these are the cones in $\mathbb{R}^{m+1}$ defined by

$$K^* = \left\{ \frac{\partial}{\partial x} y(x^e)\tau : \tau \in K \right\},$$

$$L^* = \left\{ \frac{\partial}{\partial x} y(x^e)\tau : \tau \in L \right\}.$$

(Recall that $y = (y_0, \cdots, y_m)$ and the $(m+1)\times(n+1)$ matrix $\partial y/\partial x(x^e)$ is assumed to be of full rank, $m+1$.)

From the definition of L, the cone L^* is generated by the vector $(-1, 0, \cdots, 0)$, hence this vector must be in the interior of K^*. Suppose $\sigma^0, \cdots, \sigma^m$ are linearly independent vectors in K^* such that

$$(-1, 0, \cdots, 0) = \sum_{i=0}^{m} \sigma^i.$$

Let $\tau^0, \cdots, \tau^m$ be vectors in K such that

$$\sigma^i = \frac{\partial}{\partial x} y(x^e)\tau^i.$$

For some h and for each $i = 0, \cdots, m$, there is a control variation $\alpha^i(s)x$ made at some $x(t^i)$ such that (3.2) equals τ^i. These variations can be used to construct a family of admissible trajectories whose locus of endpoints is given by $x(t^e; s; c)$ as in Lemma 3.5.

The vectors $\sigma^0, \cdots, \sigma^m$ form a basis for $\mathbb{R}^{m+1}$ and we use $\|\cdot\|$ to denote the L^1 norm relative to this basis, i.e. $\|\sum d_i\sigma^i\| = \sum |d_i|$. In particular if $r \geqq 0$, $c_i \geqq 0$ and $\sum c_i = 1$, then $\|r \sum c_i\sigma^i\| = r$.

By Taylor's theorem and compactness, there exists a constant M and an $\varepsilon > 0$ such that

$$\left\| y(x(t^e; s; c)) - y(x^e) - \frac{s^h}{h!} \sum_{i=0}^{m} c_i\sigma^i \right\| \leqq Ms^{h+1} \tag{3.4}$$

for all $\{(s, c): 0 \leqq s \leqq \varepsilon, c_i \geqq 0, \sum c_i = 1\}$.

For some $\varepsilon_1 > 0$, let

$$S = \left\{ r \sum_{i=0}^{m} c_i\sigma^i : 0 \leqq r \leqq \varepsilon_1, c_i \geqq 0 \text{ and } \sum c_i = 1 \right\}$$

and

$$\sigma^* = (-\varepsilon_1/2, 0, \cdots, 0).$$

Clearly $\sigma^* \in$ interior $S \subseteq K^*$. Define a map $g : S \to \mathbb{R}^{m+1}$ by

$$g\left(r \sum_{i=0}^{m} c_i\sigma^i \right) = y(x(t^e; (h!r)^{1/h}; c)).$$

Then from (3.4) we see that if ε_1 is small enough, there exists a constant M_1 such that

$$\left\| g\left(r \textstyle\sum_{i=0}^{m} c_i\sigma^i\right) - \left(y(x^e) + r \sum_{i=0}^{m} c_i\sigma^i\right)\right\| \leqq M_1 r^{1+1/h}.$$

Let $N(\sigma^*, \delta)$ denote the closed ball of radius δ around σ^* in the norm $\|\cdot\|$. Choose δ small enough so that this neighborhood is contained in S and choose $0<\theta<1$ such that

$$M_1\theta^{1+1/h}(\delta+\varepsilon_1/2)^{1+1/h} < \theta\delta. \tag{3.5}$$

Since S is a convex set containing both 0 and $N(\sigma^*, \delta)$, it follows that it contains $N(\theta\sigma^*, \theta\delta)$. Finally define

$$g_1\left(r \sum_{i=0}^{m} c_i\sigma^1\right) = y(x^e) + r \sum_{i=0}^{m} c_i\sigma^i - g\left(r \sum_{i=0}^{m} c_i\sigma^i\right) + \theta\sigma^*.$$

Clearly g_1 is continuous and we claim that g_1 maps $N(\theta\sigma^*, \theta\delta)$ into itself. To see this, suppose $r\sum c_i\sigma^i \in N(\theta\sigma^*, \theta\delta)$. Then

$$\begin{aligned}\left\| g_1\left(r \sum_{i=0}^{m} c_i\sigma^i\right) - \theta\sigma^*\right\| &= \left\| y(x^e) + r \sum_{i=0}^{m} c_i\sigma^i - g\left(r \sum_{i=0}^{m} c_i\sigma^i\right)\right\| \\ &\leqq M_1 r^{1+1/h}.\end{aligned} \tag{3.6}$$

By the triangle inequality,

$$r = \left\| r \sum_{i=0}^{m} c_i\sigma^i\right\| \leqq \|\theta\sigma^*\| + \theta\delta = \theta(\varepsilon_1/2 + \delta).$$

Putting these two inequalities together with (3.5) we obtain

$$\left\| g_1\left(r \sum_{i=0}^{m} c_i\sigma^i\right) - \theta\sigma^*\right\| < \theta\delta$$

as desired.

By the Brouwer fixed point theorem there exists an $r\sum c_i\sigma^i$ such that

$$g_1\left(r \sum_{i=0}^{m} c_i\sigma^i\right) = r \sum_{i=0}^{m} c_i\sigma^i$$

or

$$g\left(r \sum_{i=0}^{m} c_i\sigma^i\right) = y(x^e) + \theta\sigma^*.$$

This implies that $x(t^e; (h!\, r)^{1/h}; c)$ is the endpoint of an admissible trajectory satisfying the boundary conditions with a smaller y_0 value, hence $u^0(t)$ is not optimal.

Actually $x(t^e)$ is not even a local minimum for we can choose θ as close to 0 as we choose subject to (3.5). Q.E.D.

4. Linear conditions for scalar controls. Suppose the control of (2.1) is a scalar and the set Ω is a subinterval of $\mathbb{R}$. The PMP characterizes the optimal

control as one where the Hamiltonian achieves its maximum, and therefore we need only consider the endpoints of Ω and any interior points where (2.13) and (2.14) are satisfied. Typically for each x and λ this means considering only a finite number of discrete values of u. However, there is at least one important exception, namely, if the dynamics are linear in the control

$$\dot{x} = a_0(x) + ua_1(x). \tag{4.1}$$

Systems like this frequently arise in diverse applications because the assumption of linearity is so convenient in the formulation of mathematical models. Moreover, in Example 4.2, we show how necessary conditions developed for (4.1) can be easily extended to systems where the control enters nonlinearly.

If the dynamics is linear in u, then so is the Hamiltonian, H, and $\partial H/\partial u$ does not explicitly depend on u. If it is not zero, then the extremal control is bang-bang, i.e., at an endpoint of Ω. However, if it is zero, then the extremal control is singular since (2.14) is trivially satisfied. Moreover, (2.13) and (2.14) do not isolate the extremal control, and so we must consider the behavior of the system over an interval of time.

Suppose $u^0(t)$ and $x(t)$ are extremal for $t \in [t^0, t^e]$ for some choice of $\lambda(t)$. Assume that for $t \in (t^1, t^2)$, $u^0(t)$ is C^∞ and in the interior of Ω, hence singular. The Hamiltonian is given by

$$H(\lambda, x, u) = \lambda a_0(x) + u\lambda a_1(x),$$

and (2.13) reduces to

$$\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) = \lambda(t) a_1(x(t)) = 0 \tag{4.2}$$

for $t \in [t^1, t^2]$. Since $\lambda(t)$ annihilates H, this implies that

$$H(\lambda(t), x(t), u^0(t)) = \lambda(t) a_0(x(t)) = 0 \tag{4.3}$$

for $t \in [t^1, t^2]$.

It is straightforward to verify that given an arbitrary vector field $b(x)$ and any solution $\lambda(t)$ of the adjoint differential equation along the trajectory $x(t)$ which is generated by the control $u^0(t)$,

$$\frac{d}{dt} \lambda(t) b(x(t)) = \lambda(t)[a_0, b](x(t)) + u^0(t)\lambda(t)[a_1, b](x(t)), \tag{4.4}$$

where the Lie bracket is defined by

$$[a_i, b](x) = \left(\frac{\partial}{\partial x} b(x)\right) a_i(x) - \left(\frac{\partial}{\partial x} a_i(x)\right) b(x).$$

Repeated differentiation of (4.2) yields

$$\frac{d^k}{dt^k} \frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) = 0 \tag{4.5}$$

for $t \in [t^1, t^2]$ and $k = 0, \cdots, \infty$. In particular,

$$\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) = \lambda(t) a_1(x(t)) = 0, \tag{4.6}$$

$$\frac{d}{dt}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) = \lambda(t)[a_0, a_1](x(t)) = 0, \tag{4.7}$$

$$\begin{aligned}\frac{d^2}{dt^2}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) \\ = \lambda(t)[a_0[a_0, a_1]](x(t)) + u^0(t)\lambda(t)[a_1[a_0, a_1]](x(t)) = 0,\end{aligned} \tag{4.8}$$

$$\begin{aligned}\frac{d^3}{dt^3}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) = \lambda(t)[a_0[a_0[a_0, a_1]]](x(t)) \\ + 2u^0(t)\lambda(t)[a_0[a_1[a_0, a_1]]](x(t)) + (u^0(t))^2\lambda(t)[a_1[a_1[a_0, a_1]]](x(t)) \\ + \dot{u}^0(t)\lambda(t)[a_1[a_0, a_1]](x(t)) = 0,\end{aligned} \tag{4.9}$$

and so on. (In the next section we show that $[a_0[a_1[a_0, a_1]]] = [a_1[a_0[a_0, a_1]]]$.) One could also differentiate (4.3), however, no new conditions result.

Since $u^0(t) \in$ interior Ω for $t \in (t^1, t^2)$ we can, without loss of generality, assume that $u^0(t) = 0$ for $t \in [t^1, t^2]$ and $\pm 1 \in \Omega$ by redefining a_0 and a_1 as $a_0 + u^0 a_1$ and ca_1 for some constant c and by choosing a slightly smaller interval $[t^1, t^2]$ and a new Ω so that every admissible trajectory of the new system is also admissible for the old. Then (4.5) simplifies to

$$\frac{d^k}{dt^k}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) = \lambda(t) ad^k(a_0) a_1(x(t)) = 0 \tag{4.10}$$

for $t \in [t^1, t^2]$, where $ad^0(a_0)a_1 = a_1$ and $ad^k(a_0)a_1 = [a_0, ad^{k-1}(a_0)a_1]$.

Equation (4.5) (or (4.10)) is sometimes referred to as the *linear necessary condition* for an optimal control because it is precisely this condition that one would obtain by linearizing (4.1) around the reference trajectory and considering the effect of a sequence of first order control variations at properly chosen times. This is the McShane–Pontryagin approach. Moreover in the case of (4.10), it involves brackets of a_0 and a_1 which are linear in the controllable vector field a_1.

Equation (4.3) is a constant necessary condition, i.e., it is zero order with respect to the controllable vector field a_1 of (4.1). It follows from the first order control variations

$$\alpha^{\pm}(s)x = \gamma^0(\pm s)x \tag{4.11}$$

whose derivative applied to (2.9) yields

$$\lambda(t)\frac{d}{ds}\alpha^{\pm}(0)x(t) = \pm\lambda(t)(a_0(x(t)) + u^0(t)a_1(x(t))) \leqq 0.$$

Since $x_0 = t$, this condition is independent of (4.5) (or 4.10)). Therefore $u^0(t) = 0$ is extremal for $t \in [t^1, t^2]$ if and only if the rank of $\{ad^k(a_0)a_1(x(t)): k = 0, \cdots, \infty\}$ is less than n at each $t \in [t^1, t^2]$. (If the rank is n, then (4.3) and (4.10) supply $n+1$ linearly independent conditions, and hence only $\lambda(t) = 0$ satisfies them.)

Notice that (4.10) was not obtained directly from a control variation, but rather by differentiating (4.2). As a first application of the HMP, we would like to develop (4.10) directly via high order control variations. In the next section, these same control variations are used to obtain conditions which are quadratic in a_1.

Before we start, perhaps a word or two is required about terminology. When we speak of high order control variations, the order is with respect to the parameter s of the variations which is a time-like parameter. On the other hand, when we speak of linear or quadratic conditions, we mean relative to the controllable part of (4.1), i.e., of first or second order with respect to the integral of the absolute variation in control. In particular, when $u^0(t)=0$, these conditions can be expressed using brackets which are linear or quadratic in a_1.

Suppose $\gamma^{\pm 1}$ and γ^0 are the flows of $u^{\pm 1}(t)=\pm 1$ and $u^0(t)=0$. Then define the control variations

$$\alpha^{\pm 0}(s)x=\gamma^{\pm 1}(s)\gamma^0(-s)x.$$

Computing the first derivative,

$$\frac{d}{ds}\alpha^{\pm 0}(0)x=\pm a_1(x)=\pm ad^0(a_0)a_1(x),$$

and so these control variations yield (4.10) for $k=0$.

Next define

$$\alpha^{\pm 1}(s)x=\gamma^{\pm 1}(s)\gamma^0(s)\gamma^{\mp 1}(s)\gamma^0(-3s)x,$$

which are variations of order two, since $(d/ds)\alpha^{\pm 1}(0)x=0$.

To compute the second derivative, it is convenient to use the chain rule technique and allow $\alpha^{\pm 1}$ to operate on an arbitrary C^∞-function φ.

$$\begin{aligned}\frac{d^2}{ds^2}\varphi(\alpha^{\pm 1}(0)x)&=\pm 4(a_0a_1(\varphi(x))-a_1a_0(\varphi(x)))\\&=\pm 4[a_0,a_1]\varphi(x)=\pm 4ad(a_0)a_1(\varphi(x)).\end{aligned}$$

When applied to the HMP this yields (4.10) with $k=1$.

We generalize the above for any integer $r\geqq 1$. Define

$$\alpha_r^{\pm 1}(s)x=\gamma^{\pm 1}(s^r)\gamma^0(s)\gamma^{\mp 1}(s^r)\gamma^0(-s-2s^r)x.$$

If s^r is replaced by s_1 and s by s_2, then the chain rule implies that at $s=0$,

$$\begin{aligned}&\frac{d^j}{ds^j}=\frac{\partial^j}{\partial s_2^j},\\&\frac{d^{r+j}}{ds^{r+j}}=\frac{(r+j)!}{j!}\frac{\partial^j}{\partial s_2^j}\frac{\partial}{\partial s_1}+\frac{\partial^{r+j}}{\partial s_2^{r+j}},\\&\frac{d^{2r+j}}{ds^{2r+j}}=\frac{(2r+j)!}{j!}\frac{(2r+j)!}{j!2!}\frac{\partial^j}{\partial s_2^j}\frac{\partial^2}{\partial s_1^2}+\frac{(2r+j)!}{j!}\frac{\partial^{r+j}}{\partial s_2^{r+j}}\frac{\partial}{\partial s_1}+\frac{\partial^{2r+j}}{\partial s_2^{2r+j}}\end{aligned}\tag{4.12}$$

for $j=0,\cdots,r-1$.

From this it follows that

$$\frac{d^j}{ds^j}\varphi(\alpha_r^{\pm 1}(0)x)=0$$

for $j=1,\cdots,r$ and

$$\frac{d^{r+1}}{ds^{r+1}}\varphi(\alpha_r^{\pm 1}(0)x)=\pm(r+1)!ad(a_0)a_1(\varphi(x)).$$

Of course $\alpha_r^{\pm 1}(s)x$ does not lead to a new condition, but its generalizations $\alpha_r^{\pm k}(s)x$ do, where for k odd,

$$\alpha_r^{\pm k}(s)x=\gamma^{\pm 1}\left(\binom{k}{0}s^r\right)\gamma^0(s)\gamma^{\mp 1}\left(\binom{k}{1}s^r\right)\gamma^0(s)$$

(4.13a)

$$\cdots\gamma^0(s)\gamma^{\mp 1}\left(\binom{k}{k}s^r\right)\gamma^0(-ks-2^ks^r)x,$$

and for k even,

$$\alpha_r^{\pm k}(s)x=\gamma^{\pm 1}\left(\binom{k}{0}s^r\right)\gamma^0(s)\gamma^{\mp 1}\left(\binom{k}{1}s^r\right)\gamma^0(s)$$

(4.13b)

$$\cdots\gamma^0(s)\gamma^{\pm 1}\left(\binom{k}{k}s^r\right)\gamma^0(-ks-2^ks^r)x.$$

Using (4.12) it can be shown that

$$\frac{d^j}{ds^j}\varphi(\alpha_r^{\pm k}(0)x)=0$$

for $j=1,\cdots,r-1$ and

$$\frac{d^{r+j}}{ds^{r+j}}\varphi(\alpha_r^{\pm k}(0)x)=\pm\frac{(r+j)!}{j!}\sum_{l=0}^{k}\sum_{i=0}^{j}\binom{j}{i}l^i\binom{k}{l}(-l)^{j-i}a_0^{j-i}(a_0\pm(-1)^la_1)a_0^i(\varphi(x))$$

$$=\pm\frac{(r+j)!}{j!}\sum_{i=0}^{j}\binom{j}{i}(-1)^{j-i}c_{j,k}a_0^{j-i}a_1a_0^i(\varphi(x)),$$

where $c_{j,k}=\sum_{l=0}^{k}(-1)^l\binom{k}{l}l^j$.

In the next lemma we show that $c_{j,k}=0$ if $0\leqq j<k$ and $c_{k,k}=(-1)^kk!$ From this it follows that if $k<r$, then $\alpha_r^{\pm k}$ is a control variation of order $r+k$ and

$$\frac{d^{r+k}}{ds^{r+k}}\varphi(\alpha_r^{\pm k}(0)x)=\pm(-1)^k(r+k)!\sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}a_0^{k-i}a_1a_0^i(\varphi(x)),$$

which by induction on k can be shown to equal

$$\pm(-1)^k(r+k)!ad^k(a_0)a_1(\varphi(x)).$$

Applying these variations to the HMP yields all the linear necessary conditions (4.10).

LEMMA 4.1.[1] *For any integers* $0 \leqq j \leqq k$, *let*

$$c_{j,k} = \sum_{l=0}^{k} (-1)^l \binom{k}{l} l^j.$$

Then $c_{j,k} = 0$ *if* $j < k$ *and* $c_{k,k} = (-1)^k k!$

Proof. By the binomial formula,

$$(e^t - 1)^k = \sum_{l=0}^{k} (-1)^{k-l} \binom{k}{l} e^{lt}.$$

Expanding e^{lt} in a Taylor series yields

$$(e^t - 1)^k = \sum_{j=0}^{\infty} \sum_{l=0}^{k} (-1)^{k-l} \binom{k}{l} l^j \frac{t^j}{j!}.$$

For $j = 0, \cdots, k-1$, the coefficient of t^j on the left is clearly 0 so

$$0 = \sum_{l=0}^{k} (-1)^{k-l} \binom{k}{l} \frac{l^j}{j!} = \frac{(-1)^k}{j!} c_{j,k}.$$

The coefficient of t^k is clearly 1 so

$$1 = \sum_{l=0}^{k} (-1)^{k-l} \binom{k}{l} \frac{l^k}{k!} = \frac{(-1)^k}{k!} c_{k,k}. \qquad \text{Q.E.D.}$$

Remark. In constructing $\alpha_r^{\pm k}(s)x$, we used the flows $\gamma^{\pm 1}$ of $a_0 \pm a_1$ to obtain a high order variation whose first nonzero derivative is a *multiple* of $\pm ad^k(a_0)a_1$ $(= ad^k(a_0)(a_0 \pm a_1))$. Suppose $\beta^{\pm}(s)x$ are control variations of order h along $x(t)$ whose h derivatives are $\pm b(x(t))$ for some vector field $b(x)$. Let $\beta^{\pm}(s)x = \gamma^{\pm j}(p_{\pm j}(s)) \cdots \gamma^{\pm 1}(p_{\pm 1}(s))\gamma^0(q_{\pm}(s))x$, where $\gamma^{\pm i}$ are flows of admissible controls and $p_{\pm i}(s) \geqq 0$ for small s. Define $\zeta^{\pm 1}(s)x = \gamma^{\pm j}(p_{\pm j}(s)) \cdots \gamma^{\pm 1}(p_{\pm 1}(s))x$, and construct $\alpha_r^{\pm k}(s)x$ as in (4.13) but with $\zeta^{\pm 1}$ replacing $\gamma^{\pm 1}$. If $k < r \cdot h$, the result is a control variation of order $k + r \cdot h$ whose $k + r \cdot h$ derivative is a multiple of $\pm ad^k(a_0)b(x(t))$ along $x(t)$.

Example 4.1. Consider the linear system

$$\dot{x} = A(t)x + ub(t),$$

where $x(0) = x0$ and $|u| \leqq 1$. Introduce time as a state variable, $x_0 = t$, so that the system is autonomous and define $\mathbf{x} = (x_0, x)$ such that

$$a_0(\mathbf{x}) = \begin{pmatrix} 1 \\ A(x_0)x \end{pmatrix}, \qquad a_1(\mathbf{x}) = \begin{pmatrix} 0 \\ b(x_0) \end{pmatrix}.$$

Then

$$[a_0, a_1](\mathbf{x}) = \begin{pmatrix} 0 \\ \dfrac{d}{dx_0} b(x_0) - A(x_0)b(x_0) \end{pmatrix},$$

$$ad^2(a_0)a_1(\mathbf{x}) = \begin{pmatrix} 0 \\ \dfrac{d^2}{dx_0^2} b(x_0) - \left(\dfrac{d}{dx_0} A(x_0)\right) b(x_0) - 2A(x_0) \dfrac{d}{dx_0} b(x_0) + A^2(x_0)b(x_0) \end{pmatrix},$$

1 The author is indebted to H. Hermes for the proof of Lemma 4.1.

and so on. For autonomous systems this simplifies to

$$ad^k(a_0)a_1(\mathbf{x}) = \begin{pmatrix} 0 \\ (-1)^k A^k b \end{pmatrix}.$$

Any bracket which is homogeneous of degree two or more in $a_1(\mathbf{x})$ is identically zero. Therefore $\partial H/\partial u$ and all its time derivatives are independent of u, and (4.5) reduces to (4.10) regardless of whether $u^0(t) = 0$ or not.

Suppose the system is *controllable*, i.e., at each x there exists a k such that $a_0(\mathbf{x}), a_1(\mathbf{x}), \cdots, ad^k(a_0)a_1(\mathbf{x})$ is of full rank, $n+1$. Then there exists no nontrivial $\boldsymbol{\lambda}(t)$ satisfying (4.3) and 4.10) and any extremal control must be bang-bang, $|u(t)| = 1$. A similar analysis is given by Hermes and La Salle in § 9 of [19].

Example 4.2. Consider a nonlinear system which is *not necessarily* linear in the control

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

where $x(0) = x^0$ and $u \in \Omega$. Given a reference control $u^0(t) \in$ interior Ω for $t \in (t^1, t^2)$, we can put the system in the form (4.1) by *prolonging* the control. Define a new state $x_{n+1} = u - u^0(t)$ and a new control $v = \dot{x}_{n+1}$. Let $\mathbf{x} = (x, x_{n+1})$ and

$$a_0(\mathbf{x}) = \begin{pmatrix} f(x, u^0(x_0) + x_{n+1}) \\ 0 \end{pmatrix}, \qquad a_1(\mathbf{x}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

On the hypersurface $x_{n+1} = 0$, which includes the reference trajectory of the original problem

$$ad^k(a_0)a_1(\mathbf{x}) = \begin{pmatrix} -ad^{k-1}(f_0)f_u(x) \\ 0 \end{pmatrix}, \tag{4.14}$$

where $f_0(x) = f(x, u^0(x_0))$ and $f_u(x) = (\partial/\partial u)f(x, u^0(x_0))$. Notice that prolongation introduces a new linear direction $a_1(\mathbf{x}(t))$ and shifts the other linear directions by one $-a_0$ factor (4.14). In particular, if the original problem is linear in the control, $\dot{x} = f_0(x) + uf_u(x)$, and $u^0(t) = 0$, then prolongation essentially shifts $ad^{k-1}(f_0)f_u$ to $-ad^k(a_0)a_1$.

Consider the necessary conditions (4.3) and (4.5) for the prolonged problem where $\boldsymbol{\lambda} = (\lambda, \lambda_{n+1})$ and $\mathbf{H}(\boldsymbol{\lambda}, \mathbf{x}, v) = \boldsymbol{\lambda}(a_0(\mathbf{x}) + va_1(\mathbf{x}))$. The reference control is $v^0(t) = 0$ and for $k = 1$,

$$\frac{\partial}{\partial v}\mathbf{H}(\boldsymbol{\lambda}(t), \mathbf{x}(t), v^0(t)) = \boldsymbol{\lambda}(t)a_1(\mathbf{x}(t)) = 0,$$

which implies that $\lambda_{n+1}(t) = 0$, i.e., the prolonged adjoint variable lives on the original state space.

For $k > 1$,

$$\begin{aligned} 0 = \frac{d^k}{dt^k}\frac{\partial}{\partial v}\mathbf{H}(\boldsymbol{\lambda}(t), \mathbf{x}(t), v^0(t)) &= \boldsymbol{\lambda}(t)ad^k(a_0)a_1(\mathbf{x}(t)) \\ &= -\lambda(t)ad^{k-1}(f_0)f_u(x(t)) \\ &= -\frac{d^{k-1}}{dt^{k-1}}\frac{\partial}{\partial u}H(\lambda(t), x(t), u^0(t)), \end{aligned} \tag{4.15}$$

and so prolongation also shifts, by one time derivative, the linear necessary conditions of the original problem. Moreover (4.3) for the prolonged problem reduces to (4.3) for the original:

$$\begin{aligned} 0 = \mathbf{H}(\boldsymbol{\lambda}(t), \mathbf{x}(t), v^0(t)) &= \boldsymbol{\lambda}(t) a_0(\mathbf{x}(t)) \\ &= \lambda(t) f(x(t), u^0(t)) \\ &= H(\lambda(t), x(t), u^0(t)). \end{aligned}$$

Prolongation increases the dimension of the state space by one, but also introduces linear controllability in that direction $(a_1(x(t)))$ along the reference trajectory, so the codimension of linear controllability remains constant, and extremal trajectories remain extremal. Prolongation can also be viewed as restricting the class of admissible control variations. In the original problem the variation in u was required to be piecewise C^∞; in the prolonged problem the variation in u is continuous and piecewise C^∞. It is interesting to note that this smaller class yields the same necessary conditions for $u^0(t) \in$ interior Ω. This is a consequence of the infinite differentiability of the original problem and might not hold if it were only finitely differentiable.

Perhaps a word or two about the form of the control variations $\alpha_r^{\pm k}(s)x$ is in order. They somewhat resemble the variations of Kelley, Kopp and Moyer [8]. The derivatives of $\alpha_r^{\pm k}(s)x$ can be conveniently thought of as polynomials in the noncommuting variables a_0 and a_1. Parametrizing $\gamma^{\pm 1}$ by s^r and γ^0 by s has the following net effect; one must differentiate r times to obtain an a_1 factor, but only once for an a_0 factor. This allows us to control the relative degrees of a_0 and a_1.

The binomial coefficients and signs of $\gamma^{\pm 1}$ give $\alpha_r^{\pm k}(s)x$ the appearance of a kth order difference operator where γ^0 is the shift operator and $\gamma^{\pm 1}$ are the positive and negative evaluation operators. Needless to say, this is no coincidence since $ad^k(a_0)a_1(x)$ is precisely the kth time derivative of $a_1(x(t))$ in any coordinate system where $a_0(x)$ is a constant vector field. There are numerous other kth order difference operators and if one were to use them in an analogous fashion as models for constructing high order control variations, then the same necessary conditions would result.

It is not surprising that these necessary conditions can be expressed in terms of brackets of a_0 and a_1 for, as is well known [10], these brackets span all the directions in which the system (4.1) can evolve. However, it is a bit surprising that the $k+r$ derivative of $\alpha_r^{\pm k}$ should be exactly equal to a bracket of a_0 and a_1 when viewed as a formal polynomial in a_0 and a_1. There is a fundamental reason for this. Consider the real algebra of all formal polynomials in two noncommuting indeterminates, a_0 and a_1. The bracket is defined as before, $[a_0, a_1] = a_0 a_1 - a_1 a_0$. Then certain of these polynomials can be constructed from a_0 and a_1 via bracketing and forming linear combinations. Such polynomials are called *Lie elements* and they are characterized by *Friedrich's criterion* (see Jacobson [6, p. 170]) which, in our present context, can be described as follows. A formal polynomial in a_0 and a_1 is a Lie element if and only if whenever a_0 and a_1 are replaced by arbitrary C^∞-vector fields, the result is a first order partial differential operator on smooth functions, i.e., it involves only the first partial derivatives of the functions.

Since the first $r+k-1$ derivatives of $\alpha_r^{\pm k}$ are zero, Lemma (3.2) states that the $r+k$ derivative is a first order operator. Moreover, this is independent of the choice of a_0 and a_1, and so this derivative must be a Lie element. Because the degree of homogeneity in a_0 and a_1 is determined by the parametrization, the $r+k$ derivative can only be a multiple of $ad^k(a_0)a_1$, the only bracket which is homogeneous in a_0 and a_1 of the appropriate degrees.

A first order partial differential operator is characterized by the Liebnitz rule for the first derivative of a product of functions. (Friedrich's criterion is merely an abstract form of this.) The following will prove useful in the next section.

LEMMA 4.2. *Suppose $\alpha(s)x$ is a control variation of order h at x^1. Then the first $2h-1$ derivatives of $\alpha(s)x$ are first order partial differential operators on smooth functions at x^1.*

Proof. By Lemma (3.2) the first through h derivatives are first order operators at x^1. As for the others, let $\varphi(x)$ and $\psi(x)$ be smooth functions around x^1. Then by the generalized Liebnitz rule for higher derivatives,

$$\frac{d^j}{ds^j}\varphi\cdot\psi(\alpha(0)x^1)=\sum_{i=0}^{j}\binom{j}{i}\frac{d^i}{ds^i}\varphi(\alpha(0)x^1)\frac{d^{j-i}}{ds^{j-i}}\psi(\alpha(0)x^1).$$

Since $\alpha(s)x$ is of order h at x^1, only two terms of the right side are possibly nonzero if $0<j<2h$, so this reduces to the Liebnitz rule for first derivatives,

$$\frac{d^j}{ds^j}\varphi\cdot\psi(\alpha(0)x^1)=\varphi(x^1)\frac{d^j}{ds^j}\psi(\alpha(0)x^1)+\frac{d^j}{ds^j}\varphi(\alpha(0)x^1)\psi(x^1).\quad\text{Q.E.D.}$$

COROLLARY 4.3. *Suppose $\alpha(s)x$ is a control variation which is of order h at x^1 independent of the choice of a_0 and a_1. Then the first $2h-1$ derivatives must be Lie elements when viewed as formal polynomials in the indeterminates a_0 and a_1.*

Although the control variations considered in this section have not led to new necessary conditions, they are useful because their higher derivatives do, as we shall see in the next section. Another important aspect of these variations is that they allow us to make instantaneous control modifications to move in any linear direction. This property will allow us to cancel out undesirable lower order effects of other variations via Lemma 3.4, and thus arrive at higher order variations. We formalize this property in the following.

LEMMA 4.4. *Suppose $c_0(t),\cdots,c_k(t)$ are bounded C^∞-real-valued functions for $t\in(t^1,t^2)$. Define a vector field along $x(t)$ by*

$$b(t)=\sum_{i=0}^{k}c_i ad^i(a_0)a_1(x(t)).$$

Then for any $h>2k$, there exists a control variation $\beta(s)x$ of order h such that

$$\frac{d^h}{ds^h}\beta(0)x(t)=b(t)$$

for $t\in(t^1,t^2)$.

Proof. Proceed by induction on k. If $k=1$, *choose a constant c* large enough so that $|c_0(t)|\leqq c$ for $t\in(t^1,t^2)$. Let $u^1(t)=c_0(t)/c$ and construct the control variations $\alpha_h^{+0}(s)x$ as before using $a_0+u^1a_1$ instead of a_0+a_1. This is a control

variation of order h and

$$\frac{d^h}{ds^h}\alpha_h^{+0}(0)(x(t)) = h!ad^0(a_0)(u^1a_1)(x(t))$$
$$= h!u^1(t)a_1(x(t))$$
$$= \frac{h!}{c}c_0(t)a_1(x(t)).$$

The desired variation is $\beta(s)x = \alpha_h^{+0}((c/h!)^{1/h}s)x$.

Now suppose the lemma is true for $k-1$. Then define $u^k(t) = c_k(t)/c$ where $c \geqq |c_k(t)|$ for $t \in (t^1, t^2)$. Construct $\alpha_r^{+k}(s)x$ using $a_0 \pm (-1)^k u^k a_1$ instead of $a_0 \pm a_1$ where $r = h-k > k$. This is a variation of order h and

$$\frac{d^h}{ds^h}\alpha_r^{+k}(0)(x(t)) = h!ad^k(a_0)(u^k a_1)(x(t))$$
$$= h!u^k(t)ad^k(a_0)a_1(x(t)) + \text{linear combination}$$
$$\text{of } ad^i(a_0)a_1(x(t)) \quad \text{for } i = 0, \cdots, k-1.$$

Define $\beta^k(s)x = \alpha_r^k((c/h!)^{1/h}s)x$. By induction there exist $\beta^{k-1}(s)x$ of order h such that

$$\frac{d^h}{ds^h}\beta^{k-1}(0)x(t) = b(t) - \frac{d^h}{ds^h}\beta^k(0)x(t).$$

The desired variation is obtained by "adding" β^k and β^{k-1} as described in Lemma 3.4. Q.E.D.

Remark. It is important to note that in the construction of these variations, the flow of $a_0 \pm u^i a_1$ is parametrized by a multiple of s^r where $r = h-i$ for $i = 0, \cdots, k$. Therefore should we continue to differentiate, the first derivative that could possibly involve a term quadratic in a_1 is the $2(h-k)$ derivative.

COROLLARY 4.5. *Consider the nonlinear system of Example* 4.2 *which is not necessarily linear in the control. Suppose* $u^0(t)$ *is* C^∞ *and in the interior of* Ω *for* $t \in (t^1, t^2)$. *Given any bounded* C^∞*-real-valued functions* $c_0(t), \cdots, c_k(t)$ *define a vector field along* $x(t)$ *by*

$$b(t) = \sum_{i=0}^{k} c_i ad^i(f_0)f_u(x(t)).$$

Then for any $h > 2k+2$, *there exists a control variation* $\beta(s)x$ *of order h such that*

$$\frac{d^h}{ds^h}\beta(0)x(t) = b(t)$$

for $t \in (t^1, t^2)$.

Proof. Prolong the problem as before and apply Lemma 4.4.

5. Quadratic conditions for scalar controls. For (4.1) assume that $u^0(t) = 0 \in$ interior Ω is an extremal control for $|t - t^1| < \varepsilon$. Then (2.14) is trivially satisfied so that $u^0(t)$ is singular. Moreover since $u^0(t)$ is an extremal control, (4.3) and (4.10)

are satisfied for some $\lambda(t)$. Because these are equality constraints rather than inequality constraints, replacing $\lambda(t)$ with $-\lambda(t)$ does not alter them. Therefore they do not distinguish between minimizing and maximizing singular extremals.

To clear up this ambiguity, quadratic necessary conditions were developed by Kelley [7], Kopp and Moyer [9], Kelley, Kopp and Moyer [8], Tait [17], Goh [3], [4], Robbins [14], [15] and others. We refer the reader to the survey articles of Gabasov and Kirillova [2], Bell [1] and Jacobson [18] for extensive bibliographies. These conditions are sometimes referred to as the generalized Legendre–Clebsch conditions (GLC) because they resemble the Legendre–Clebsch condition (2.14) when expressed in terms of the Hamiltonian. Generally the proofs of the GLC ignore the problem of terminal constraints either by assuming there are not any, or by a normality assumption, a sometimes vague concept in the literature. Essentially, normality means that there exists sufficient local controllability around the reference trajectory to meet any terminal constraints that might be imposed without affecting the validity of the GLC. We give a more precise definition later.

In this section, using the HMP, we develop quadratic necessary conditions which generalize the GLC to problems with terminal constraints without using a normality assumption.

Let D_i^j denote the linear space of Lie elements which are homogeneous of degree i and j in the indeterminates a_0 and a_1, respectively, and let

$$D_i = \operatorname{span} \bigcup_{j=0}^{\infty} D_i^j,$$

$$D^j = \operatorname{span} \bigcup_{i=0}^{\infty} D_i^j,$$

$$D = \operatorname{span} \bigcup_{i,j=0}^{\infty} D_i^j.$$

Let $D_i^j(x)(D_i(x), D^j(x), D(x))$ denote the linear subspace of a tangent vector at x obtained by substituting the vector fields of (4.1) in the Lie elements and evaluating at x.

Suppose $u^0(t) = 0$ and $x(t)$ are a singular extremal control and trajectory on $[t^1, t^2]$. Following Robbins [15], we say that the control is *singular of degree* $h+1$ on this interval if h is the smallest integer such that for some $t \in (t^1, t^2)$,

$$[a_1, ad^h(a_0)a_1](x(t)) \notin D^1(x(t)).$$

The next theorem describes the quadratic necessary conditions for such a control to be minimal.

THEOREM 5.1. *Assume that $u^0(t)$ and $x(t)$ are defined for* (4.1) *on $[t^0, t^e]$. Suppose $u^0(t) = 0 \in$ interior Ω on the subinterval (t^1, t^2). If u is singular of degree $h+1$ on this subinterval and h is finite, then h is odd. If $u^0(t)$ is minimal, then there exists a $\lambda(t)$ satisfying the PMP on $[t^0, t^e]$ such that*

$$(-1)^{(h+1)/2}\lambda(t)[a_1, ad^h(a_0)a_1](x(t)) \leqq 0$$

on the subinterval $[t^1, t^2]$.

Note that the theorem does not imply that the degree of singularity is finite, just that if $h < \infty$, then h must be odd, whether the extremal trajectory is minimal or not. Later we give an example where the degree of singularity is infinite. There may exist several subintervals of $[t^0, t^e)$ on which the degree of singularity varies. Before proving the theorem, we state a generalization and a corollary which do not assume linearity in the control or $u^0(t) = 0$. First we make a generalized definition. Suppose $u^0(t)$ and $x(t)$ are a singular extremal control and trajectory for (2.1) on $[t^1, t^2]$. The control is *singular of degree* $h+1$ on this interval if h is the smallest integer for which there exists $\lambda(t)$ satisfying the adjoint differential equation (2.6) and the constant and linear necessary conditions

$$H(\lambda(t), x(t), u^0(t)) = 0, \tag{5.1}$$

$$\frac{d^k}{dt^k}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) = 0 \tag{5.2}$$

for $k = 0, \cdots, \infty$ on any nontrivial subinterval of $[t^1, t^2]$ such that for some t in this subinterval,

$$\frac{\partial}{\partial u}\frac{d^{h+1}}{dt^{h+1}}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) \neq 0.$$

Notice that a control could be singular of degree $h+1 = 0$ or ∞.

THEOREM 5.2. *Assume that $u^0(t)$ and $x(t)$ are defined for* (2.1) *on $[t^0, t^e]$. Suppose $u^0(t) \in$ interior Ω on the subinterval (t^1, t^2). If u is singular of degree $h+1$ on this subinterval and h is finite, then h is odd. If $u^0(t)$ is minimal, then there exists a $\lambda(t)$ satisfying the PMP on $[t^0, t^e]$ such that*

$$(-1)^{(h+1)/2}\frac{\partial}{\partial u}\frac{d^{h+1}}{dt^{h+1}}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) \leqq 0$$

on the subinterval $[t^1, t^2]$.

A singular extremal control $u^0(t) \in$ interior Ω and trajectory $x(t)$ *are normal* on (t^1, t^2) if for each $t \in (t^1, t^2)$ there exists only one linearly independent $\lambda(t)$ satisfying the constant and linear necessary conditions (5.1) and (5.2). Since $x = (x_0, \cdots, x_n)$ this is equivalent to the assumption that the variations $\alpha^{\pm}(s)x$ of (4.11) and $\alpha_r^{\pm k}(s)x$ of (4.13) supply exactly n-dimensional local controllability at each $x(t)$ for $t \in (t^1, t^2)$.

In particular for (4.1) and $u^0(t) = 0$, this is equivalent to the assumption that the dimension of $D^1(x(t))$ is $n-1$ for each $t \in (t^1, t^2)$.

COROLLARY 5.3 (Kelley, Kopp and Moyer [8]). *Assume that $u^0(t)$ and $x(t)$ are defined for* (2.1) *on $[t^0, t^e]$. Suppose $u^0(t) \in$ interior Ω and is normal on the subinterval $[t^1, t^2]$. If $u^0(t)$ is minimal, then there exists a $\lambda(t)$ satisfying the PMP on $[t^0, t^e]$ which is unique to the scalar multiple by normality. Let h be the smallest integer such that*

$$\frac{\partial}{\partial u}\frac{d^{h+1}}{dt^{h+1}}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) \neq 0$$

for some $t \in (t^1, t^2)$. *If* h *is finite, then* h *is odd and on the subinterval* $[t^1, t^2]$,

$$(-1)^{(h+1)/2} \frac{\partial}{\partial u} \frac{d^{h+1}}{dt^{h+1}} \frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) \leqq 0$$

must hold.

Proof. We now prove Theorem 5.1. In Example (5.2) we show that Theorem 5.2 is equivalent to Theorem 5.1 by prolongation. The corollary follows immediately from Theorem 5.2.

Except for a nowhere dense set, every $t \in (t^1, t^2)$ is contained in an open interval where $D^1(x(t))$ is of constant dimension with a basis consisting of

$$\{a_1(x(t)), ad(a_0)a_1(x(t)), \cdots, ad^l(a_0)a_1(x(t))\}$$

for some l. Without loss of generality we can assume that the open interval is all of (t^1, t^2), for at other points the theorem follows by taking continuous limits.

By repeated application of the Jacobi identity for Lie elements, $[b_i[b_j, b_k]] = [[b_i, b_j]b_k] + [b_j[b_i, b_k]]$, and the skew symmetry relation, $[b_i, b_j] = -[b_j, b_i]$, it is easy to see that

$$\begin{aligned}[a_1, ad^i(a_0)a_1] = &\sum_{j=0}^{l-1} (-1)^j [a_0[ad^j(a_0)a_1, ad^{i-j-1}(a_0)a_1]] \\ &+ (-1)^l [ad^l(a_0)a_1, ad^{i-l}(a_0)a_1].\end{aligned} \tag{5.3}$$

If i is even and $l = i/2$, then skew symmetry implies that the last term on the right side is zero so

$$[a_1, ad^i(a_0)a_1] = \sum_{j=0}^{(i/2)-1} (-1)^j [a_0[ad^j(a_0)a_1, ad^{i-j-1}(a_0)a_1]]. \tag{5.4}$$

From (5.3) and (5.4) it can be shown that a basis for the linear space D_i^2 of Lie elements consists of

(5.5a) $\{[a_1, ad^i(a_0)a_1], ad^2(a_0)[a_1, ad^{i-2}(a_0)a_1], \cdots, ad^{i-1}(a_0)[a_1[a_0, a_1]]\}$

if i is odd, and

(5.5b) $\{[a_0[a_1, ad^{i-1}(a_0)a_1]], ad^3(a_0)[a_1, ad^{i-3}(a_0)a_1], \cdots, ad^{i-1}(a_0)[a_1[a_0, a_1]]\}$

if i is even.

Now suppose that u is singular of degree $h+1$, i.e.,

$$[a_1, ad^i(a_0)a_1](x(t)) \in D^1(x(t))$$

on the subinterval $[t^1, t^2]$ for $i = 1, \cdots, h-1$, but not for $i = h$. Bracketing both sides with a_0 yields

$$ad^j(a_0)[a_1, ad^i(a_0)a_1](x(t)) \in D^1(x(t)) \tag{5.6}$$

for $j = 0, \cdots, \infty$, $i = 1, \cdots, h-1$ and $t \in [t^1, t^2]$. In particular,

$$D_i^2(x(t)) \in D^1(x(t)) \tag{5.7}$$

on $[t^1, t^2]$ for $i = 1, \cdots, h-1$.

If h is even, (5.5b) and (5.6) imply that $[a_1, ad^h(a_0)a_1](x(t)) \in D^1(x(t))$ which contradicts the definition of h, hence h must be odd. Notice also that (5.6) implies that the only bracket which keeps (5.7) from being true for $i = h$ is $[a_1, ad^h(a_0)a_1](x(t))$.

To prove the rest of the theorem we must construct an appropriate high order control variation. We start with $\alpha_r^{+k}(s)x$ for any $k \geqq (h+1)/2$, $r > k$ and $r > h$. We have already computed the first $k+r$ derivatives of this variation and we know by Corollary 4.3 that the first $2(k+r)-1$ derivatives are Lie elements. We wish to study the jth derivative where $k+r < j \leqq h+2r \leqq 2(k+r)-1$.

It is easy to see that this variation causes no displacement in the time direction and so the jth derivative cannot possibly contain an a_0 term. Moreover, from the parametrization of the components of $\alpha_r^{+k}(s)x$ we know that the jth derivative is a sum of elements of D^1 if $r \leqq j < 2r$ and a sum of elements of D^1 and D^2 if $2r \leqq j < 3r$.

We already have control variations in the directions of $D^1(x(t))$ so we are only interested in the part of the jth derivative that lies in D^2 for $2r < j \leqq 3r$. Again from the parametrization we know that the part from D^2 is more precisely from D^2_{j-2r} and hence a linear combination of (5.5). Moreover from (5.7) we know that

$$D^2_{j-2r}(x(t)) \in D^1(x(t))$$

for $1 \leqq j-2r < h$, so the first derivative that could possibly furnish a new test is $j = h+2r$. (Note that by the choice of r and k, it follows that $j < 3r$ and $j < 2(k+r)$ as desired.) This derivative can be expanded in the basis (5.5) but we are really only interested in the coefficient of $[a_1, ad^h(a_0)a_1]$ for this is the part of the derivative that lies outside $D^1(x(t))$.

To compute the coefficient of $[a_1, ad^h(a_0)a_1]$, we need only compute the coefficient of the monomial $a_1 a_0^h a_1$ in the jth derivative for this is the only bracket of (5.5) that contains that monomial. We defer to a later lemma the computation that shows that the sign of this coefficient is $(-1)^{(h+1)/2}$.

In summary, we know the following. The first $k+r-1$ derivatives of $\alpha_r^{+k}(s)(x(t))$ are zero, derivatives $k+r$ through $h+2r-1$ lie in $D^1(x(t))$ and the $h+2r$ derivative consists of some parts from $D^1(x(t))$ plus a positive multiple of $(-1)^{(h+1)/2}[a_1, ad^h(a_0)a_1](x(t))$. To complete the proof we must make $\alpha_r^{+k}(s)x$ into a control variation of order $h+2r$ at $x(t)$ by canceling out all the lower derivatives for $t \in (t^1, t^2)$.

To do this we must apply Lemma (4.4) using the fact that $\{a_1(x(t)), \cdots, ad^l(a_0)a_1(x(t))\}$ spans $D^1(x(t))$. The lemma allows us to construct a control variation of any order $>2l$ whose first nonzero derivative is any vector field along $x(t)$ which lies in $D^1(x(t))$. Therefore we must choose k and r such that $k+r > 2l$ and by "adding" new variations to $\alpha_r^{+k}(s)x$, we can cancel out its lower derivatives from $k+r$ through $h+2r-1$. Call the resulting variation $\beta(s)x$.

We must be careful in doing this, for it is possible that the sign of $[a_1, ad^h(a_0)a_1](x(t))$ in the $h+2r$ derivative of $\beta(s)x$ differs from its sign in the $h+2r$ derivative of $\alpha_r^{+k}(s)x$, and this would change the test. Recall that the parameters of the flows of $a_0 \pm a_1$ in $\alpha_r^{+k}(s)x$ are s^r and, on the other hand, the variations of Lemma 4.4 used to cancel derivatives $r+k$ through $2r+h-1$ are composed of the flows of $a_0 \pm u^i a_1$ parametrized by s^{r+k-l} or higher powers of s.

So, if $k > l$, then these "added" variations cannot possibly change the coefficient of $[a_1, ad^h a_0 a_1](x(t))$, although it could change the part of the $h+2r$ derivative which lies in $D^1(x(t))$. However this is not important for the test since (4.10) implies that $\lambda(t)$ annihilates $D^1(x(t))$.

In closing, we emphasize that this necessary condition is an inequality precisely because the bracket involved is quadratic in a_1. If we used $\alpha_r^{-k}(s)x$ instead of $\alpha_r^{+k}(s)x$ as a base for our high order variation, the same necessary condition would result because this is equivalent to replacing $a_1(x)$ by $-a_1(x)$ which leaves invariant the brackets quadratic in a_1. Using either of these variations, we have controllability in the direction $(-1)^{(h+1)/2}[a_1, ad^h(a_0)a_1](x(t))$, but not its negative. Q.E.D.

LEMMA 5.4. *Let $c_{k,h}$ be the coefficient of $a_1 a_0^h a_1(\varphi(x))$ in*

$$\frac{d^{2r+h}}{ds^{2r+h}}\varphi(\alpha_r^{+k}(0)x)$$

where $k \geqq (h+1)/2$, $r > k$ and $r > h$. Then

$$c_{k,h} = 0 \quad \text{if } h = 2, 4, \cdots, 2k-2$$

and

$$(-1)^{(h+1)/2} c_{k,h} > 0 \quad \text{if } h = 1, 3, \cdots, 2k-1.$$

Proof. By direct computation,

$$c_{k,h} = (2r+h)! \sum_{0 \leqq i < j \leqq k} (-1)^{i+j} \binom{k}{i}\binom{k}{j} \frac{(j-i)^h}{h!}.$$

If $h > 0$ and is even, then

$$2c_{k,h} = (2r+h)! \sum_{i,j=0}^{k} (-1)^{i+j} \binom{k}{i}\binom{k}{j} \frac{(j-i)^h}{h!}.$$

Expand $(e^{-t}-1)^k(e^t-1)^k$ by the binomial formula:

$$(e^{-t}-1)^k(e^t-1)^k = \sum_{i,j=0}^{k} (-1)^{i+j} \binom{k}{i}\binom{k}{j} e^{(j-i)t}.$$

Expand $e^{(j-i)t}$ in a Taylor series:

$$(e^{-t}-1)^k(e^t-1)^k = \sum_{h=0}^{\infty} \sum_{i,j=0}^{k} (-1)^{i+j} \binom{k}{i}\binom{k}{j} \frac{(j-i)^h}{h!} t^h.$$

For $h = 2, 4, \cdots, 2k-2$, the coefficient of t^h on the left side is clearly 0 so

$$0 = \sum_{i,j=0}^{k} (-1)^{i+j} \binom{k}{i}\binom{k}{j} \frac{(j-i)^h}{h!} = \frac{2}{(2r+h)!} c_{k,h},$$

and the first claim of the lemma has been shown.

If we assume h is a real variable, then for fixed k, $c_{k,h}$ is a sum of k exponentials. Therefore it has at most $k-1$ zeros, which we have just shown to be $h = 2, 4, \cdots, 2k-2$. It follows that $c_{k,h}$ alternates signs at $h = 1, 3, \cdots, 2k-1$

and, in particular, the sign of $c_{k,2k-1}$ must be the same as the coefficient of the largest exponential which is $(-1)^k = (-1)^{(h+1)/2}$. Q.E.D.

Example 5.1. Consider the problem of minimizing $x_4(t^e)$ subject to $x^0 = 0$, $x_0(t^e) = 1$ and

$$\dot{x}_0 = 1, \qquad \dot{x}_3 = x_1^2/2,$$
$$\dot{x}_1 = u, \qquad \dot{x}_4 = -x_2^2/2,$$
$$\dot{x}_2 = x_1,$$

Clearly the trajectory determined by $u^0(t) = 0$ is not optimal, but let us apply the previous theorems and corollary. This trajectory, $x(t) = (t, 0, 0, 0, 0)$ for $t \in [0, 1]$, is a singular extremal since $\lambda(t) = (0, 0, 0, 0, -1)$ satisfies the PMP (uniquely to scalar multiple) and $\partial^2/\partial u^2 H = 0$.

A straightforward calculation shows that $[a_1, ad(a_0)a_1](x(t)) \notin D^1(x(t))$, so the degree of singularity $h + 1 = 2$ and we apply the test

$$-\lambda(t)[a_1, ad(a_0)a_1](x(t)) \leqq 0$$

which is trivially satisfied. Therefore Theorem 5.1 does not rule out $u^0(t) = 0$.

To understand the relationship between them, let us apply the other theorem and corollary. For some $t \in [0, 1]$ (in fact, every t), the adjoint vector $\mu(t) = (0, 0, 0, 1, 0)$ satisfies the necessary conditions (5.1) and (5.2) and the adjoint differential equation on $[0, 1]$. For any $t \in [0, 1]$,

$$\frac{\partial}{\partial u}\frac{d^2}{dt^2}\frac{\partial}{\partial u} H(\mu(t), x(t), u^0(t)) \neq 0,$$

so again $h + 1 = 2$ and Theorem 5.2 only allows us to test if

$$-\frac{\partial}{\partial u}\frac{d^2}{dt^2}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) \leqq 0,$$

which of course is trivially satisfied.

As for Corollary 5.3, it does not apply since the trajectory is not normal (dimension of $D^1(x(t))$ is $2 < n - 1 = 3$).

These quadratic necessary conditions failed to rule out an obviously nonminimal (in fact, maximal) trajectory because the problem was not given as a minimal realization (see Sussmann [16]). We are only interested in $x_0(t^e)$ and $x_4(t^e)$, so we define $y(x) = (y_0(x), y_1(x)) = (x_4, x_0)$ as our output. It is clear that the x_3 coordinate is superfluous to the input-output description of the problem and may be dropped. Then $[a_1, ad(a_0)a_1](x(t)) = 0$ and so $h + 1 = 4$. Since

$$\lambda(t)[a_1, ad^3(a_0)a_1](x(t)) = 1 \not\leqq 0,$$

the trajectory is nonoptimal. Similarly, when applying Theorem 5.2, we find $h = 3$ and the corresponding test rules out $u^0(t) = 0$. Moreover since the dimension $n + 1$ of the state space is now 4 rather than 5, and the dimension of $D^1(x(t))$ is still 2, the trajectory is normal by the comments following Theorem 5.2. Corollary 5.3 also rules out the trajectory.

Notice that if an additional terminal constraint, $x_3(t^e) = 0$, is added to the original problem, then the output map must be expanded to include x_3 and this

coordinate cannot be eliminated. Therefore the quadratic necessary conditions no longer rule out the control $u^0(t) = 0$, but this is as it should be for this control generates the only trajectory satisfying the terminal constraints.

Example 5.2. Once again consider a nonlinear system which is not necessarily linear in the control as in Example 4.2. Prolong the system as before by introducing new state and control variables. On the hypersurface $x_{n+1} = 0$, which includes the reference trajectory

$$[a_1, ad^h(a_0)a_1](\mathbf{x}) = \begin{pmatrix} -ad^{h-1}(f_0)f_{u,u}(x) - \sum_{i=0}^{h-2} ad^{h-2-i}(f_0)[f_u, ad^i(f_0)f_u](x) \\ 0 \end{pmatrix} \tag{5.8}$$

when f_0, f_u are as before and $f_{u,u}(x) = (\partial^2/\partial u^2)f(x, u^0(x_0))$. This shows that the GLC for $h = 1$ of the prolonged problem is equivalent to the Legendre–Clebsch condition of the original.

If the original problem is linear in the control, $\dot{x} = f_0(x) + uf_u(x)$, and $u^0(t) = 0$, then the prolongation shifts $D^2_{h-2}(x(t))$ to $D^2_h(\mathbf{x}(t))$. It also shifts the GLC for $h-2$ to the GLC for h; the lower order GLC is satisfied with equality and therefore multiplication by $\boldsymbol{\lambda}(t)$ cancels all but $\lambda(t)[f_u, ad^{h-2}(f_0)f_u](x(t))$ on the right side of (5.8).

Prolongation can be used to show that Theorem 5.1 implies Theorem 5.2 for an arbitrary nonlinear system (2.1). A straightforward calculation shows that for $h \geqq 1$:

$$\begin{aligned} &\frac{\partial}{\partial u}\frac{d^{h-1}}{dt^{h-1}}\frac{\partial}{\partial u}H(\lambda(t), x(t), u^0(t)) \\ &\quad = \lambda(t)ad^{h-1}(f_0)f_{u,u}(x) + \sum_{i=0}^{h-2} \lambda(t)ad^i(f_0)[f_u, ad^{h-2-i}(f_0)f_u](x(t)) \\ &\quad = -\boldsymbol{\lambda}(t)[a_1, ad^h(a_0)a_1](\mathbf{x}(t)). \end{aligned}$$

Many proofs of the GLC are based on the reverse of prolongation. In the literature this is known as a "transformation of control variable" or "passing to the accessory minimum problem". By using the integral of the old control as the new control variable, the GLC for h is converted into the GLC for $h-2$. Repeated application reduces the GLC for h to the Legendre–Clebsch condition which has been previously demonstrated. The principal difficulty in applying this technique is that in effect, one is dropping the dimension of the state by changing a state variable to a control variable. Another way of looking at this is to say that one is allowing impulse controls. One must justify the claim that necessary conditions developed using this wider class of controls are also necessary conditions for the original class. These problems are usually ignored in the literature and instead normality is assumed to be sufficient to overcome any difficulties that arise in this fashion and also to meet the terminal constraints.

Example 5.3. To see that the degrees of singularity $h+1$ can be infinite, consider the problem of minimizing $x_2(t^e)$ subject to $x(0) = 0$, $x_0(t^e) = 1$, $|u| \leqq 1$ and

$$\dot{x}_0 = 1, \qquad \dot{x}_2 = x_1^3/6,$$
$$\dot{x}_1 = u + x_1^2.$$

A straightforward computation shows that along the trajectory $x(t) = (t, 0, 0)$ generated by the control $u^0(t) = 0$ for the adjoint variable $\lambda(t) = (0, 0, -1)$,

$$\frac{d^h}{dt^h}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) = 0$$

and

$$\frac{\partial}{\partial u}\frac{d^h}{dt^h}\frac{\partial}{\partial u} H(\lambda(t), x(t), u^0(t)) = 0$$

for all h. Therefore this trajectory is a singular extremal, and the quadratic tests are inconclusive. To show the nonoptimality of this trajectory, the HMP must be used to construct a necessary condition which is particularly suited to the problem at hand, i.e., a cubic or higher test.

Let $\alpha^{\pm}(s)x$ and $\alpha_1^{\pm 0}(s)x$ be defined by (4.11) and (4.13) respectively, such that

$$\frac{d}{ds}\alpha^{\pm}(0)x = \pm(1, 0, 0),$$

$$\frac{d}{ds}\alpha_1^{\pm 0}(s)x = \pm(0, 1, 0),$$

and so the trajectory is normal. Consider $\alpha_1^{\pm 1}(s)x$. It is easy to show that this is a variation of order two and

$$\frac{d^2}{ds^2}\alpha_1^{\pm 1}(s)x = \pm 4[a_0, a_1](x),$$

but this bracket is zero along $x(t)$. Therefore, in this case, $\alpha_1^{\pm 1}$ are of order at least 3. Computing the next derivative which also must be a Lie element (since $\alpha_1^{\pm 1}$ is of order 2 for all a_0 and a_1),

$$\frac{d^3}{ds^3}\alpha_1^{\pm 1}(s)x = \pm 6[a_0[a_0, a_1]](x) - 10[a_1[a_0, a_1]](x).$$

The first of these brackets is zero and the second is $(0, -2, 0)$ along $x(t)$. Therefore this derivative can be canceled by "adding" $\alpha_1^{+0}(2^{1/3}s^3/6)x$ via Lemma 3.4. Call the resulting variation $\beta^{\pm}(s)x$; it is of order 4 along $x(t)$. Because of the parametrization of α_1^{+0}, it follows that the fourth derivatives of $\beta^{\pm}$ and $\alpha_1^{\pm 1}$ are identical. This derivative is not a Lie element, but it is in the span of $D(x(t))$. By direct computation,

$$\frac{d^4}{ds^4}\beta^{\pm}(s)x = \pm(0, 0, 6).$$

Applying this to the HMP yields the nonoptimality of $u^0(t) = 0$.

This is actually a third order test because it is a multiple of $[a_1[a_1[a_0, a_1]]]x) = (0, 0, 1)$ which is cubic in a_1. A similar conclusion can be obtained by the method of Hermes [20].

This example demonstrates how one constructs necessary conditions which are adapted to a particular problem. First one applies the standard linear and

quadratic tests developed in the last two sections. If these are inconclusive, i.e., they are satisfied with equality rather than strict inequality, then one must construct cubic and higher tests ad hoc. The basic building blocks are the family of variations $\alpha_r^{\pm k}(s)x$. One works with as small a k and r as possible and considers the higher derivatives. If these can be canceled using linear or quadratic controllability, then one "adds" the appropriate variations to do so. Higher derivatives are considered and hopefully a definitive test is eventually realized.

6. Quadratic conditions for vector controls. In this section, we generalize the results of the last to a system with vector controls,

$$\dot{x} = a_0(x) + \sum_{i=1}^{l} u_i a_i(x), \tag{6.1}$$

where $u = (u_1, \cdots, u_l)$ is constrained to lie in Ω, a subset of $\mathbb{R}^l$ with nonempty interior.

By fixing all but one of the controls and varying the other, one obtains the previously discussed linear and quadratic necessary conditions for each control.

These are the same linear necessary conditions involving $(d^h/dt^h)(\partial/\partial u_i)H$ (or if $u^0(t) = 0$, $ad^h(a_0)a_i$) for $i = 1, \cdots, l$, and the same quadratic conditions involving $(\partial/\partial u_i)(d^h/dt^h)(\partial/\partial u_i)H$ (or if $u^0(t) = 0$, $[a_i, ad^h(a_0)a_i]$) for $i = 1, \cdots, l$. There are, however, new quadratic necessary conditions associated with the mixed partials $(\partial/\partial u_i)(d^h/dt^h)(\partial/\partial u_j)H$ (or if $u^0(t) = 0$, $[a_i, ad^h(a_0)a_j]$) for $i, j = 1, \cdots, l$.

These conditions were first developed by Goh [4] using a sequence of accessory minimum problems under an assumption of normality. We use the HMP to prove these results and extend them to problems with terminal constraints without normality.

Let D_i^j denote the linear space of Lie elements which are homogeneous of degree i in the indeterminate a_0 and homogeneous of degree j in the vector of indeterminates $(a_1, \cdots, a_l)$, and let D_i, D^j, D, $D_i^j(x)$, $D_i(x)$, $D^j(x)$ and $D(x)$ be as before. Since there is more than one controllable vector field, there are some significant differences. For example, D_0^2 which previously was $\{0\}$ since $[a_1, a_1] = 0$, now contains the nontrivial Lie elements $[a_i, a_j]$ where $i \neq j$.

Suppose the reference control is $u^0(t) = 0$. Associated with each control u_i, there is a *degree of singularity* $h_i + 1$ defined as before; h_i is the smallest integer such that for some $t \in (t^1, t^2)$,

$$[a_i, ad^{h_i}(a_0)a_i](x(t)) \notin D^1(x(t)).$$

We wish to emphasize the fact that $D^1(x(t))$ contains $ad^k(a_0)a_j(x(t))$ where $j \neq i$, but the arguments of Theorem 5.1 are still valid, so that if $h_i < \infty$, it must be odd.

THEOREM 6.1. *Assume that $u^0(t)$ and $x(t)$ are defined for* (6.1) *on $[t^0, t^e]$. Suppose $u^0(t) = 0 \in$ interior Ω and each u_i is singular of degree $h_i + 1$ on the subinterval (t^1, t^2). If $u^0(t)$ is minimal, then there exists a $\lambda(t)$ satisfying the PMP on $[t^0, t^e]$ such that on the subinterval $[t^1, t^2]$,*

$$\lambda(t)[a_i, ad^k(a_0)a_j](x(t)) = 0 \tag{6.2}$$

for $k = 0, \cdots, (h_i + h_j)/2 - 1$, $1 \leqq i, j \leqq l$. *Moreover, if* $h_i < \infty$ *for* $i = 1, \cdots, k \leqq l$, *then the* $k \times k$ *matrix whose* i, j *entry is*

$$(-1)^{(h_j+1)/2} \lambda(t)[a_i, ad^{(h_i+h_j)/2}(a_0)a_j](x(t)), \tag{6.3}$$

where $1 \leqq i, j \leqq k$ *must be symmetric and nonpositive definite.*

The following theorem generalizes the above to an arbitrary nonlinear system (2.1) where $u^0(t)$ is not necessarily zero. Recall that the control u_i is singular of degree $h_i + 1$ on $[t^1, t^2]$ if h_i is the smallest integer such that for some $t \in (t^1, t^2)$ there exists $\lambda(t)$ satisying the adjoint differential equation (2.6) and the constant and linear necessary condition

$$H(\lambda(t), x(t), u^0(t)) = 0, \tag{6.4}$$

$$\frac{d^k}{dt^k} \frac{\partial}{\partial u_j} H(\lambda(t), x(t), u^0(t)) = 0 \tag{6.5}$$

for $k = 0, \cdots, \infty$ and $j = 1, \cdots, l$ on any nontrivial subinterval of $[t^1, t^2]$ such that for some t in this subinterval,

$$\frac{\partial}{\partial u_i} \frac{d^{h+1}}{dt^{h+1}} \frac{\partial}{\partial u_i} H(\lambda(t), x(t), u^0(t)) \neq 0.$$

Again we wish to emphasize that (6.5) must hold for every u_j and if $h_i < \infty$, it must be odd.

THEOREM 6.2. *Assume that* $u^0(t)$ *and* $x(t)$ *are defined for* (2.1) *on* $[t^0, t^e]$. *Suppose* $u^0(t) \in$ interior Ω *and each* u_i *is singular of degree* $h_i + 1$ *on the subinterval* (t^1, t^2). *If* $u^0(t)$ *is minimal, then there exists a* $\lambda(t)$ *satisfying the PMP on* $[t^0, t^e]$ *such that on the subinterval* $[t^1, t^2]$,

$$\frac{\partial}{\partial u_i} \frac{d^k}{dt^k} \frac{\partial}{\partial u_j} H(\lambda(t), x(t), u^0(t)) = 0$$

for $k = 0, \cdots, (h_i + h_j)/2$, $1 \leqq i, j \leqq l$. *Moreover if* $h_i < \infty$ *for* $i = 1, \cdots, k \leqq l$, *then the* $k \times k$ *matrix whose* i, j *entry is*

$$(-1)^{(h_j+1)/2} \frac{\partial}{\partial u_i} \frac{d^{(h_i+h_j)/2+1}}{dt^{(h_i+h_j)/2+1}} \frac{\partial}{\partial u_j} H(\lambda(t), x(t), u^0(t)),$$

where $1 \leqq i, j \leqq k$, *must be symmetric and nonpositive definite.*

As before, a singular extremal control $u^0(t) \in$ interior Ω and trajectory $x(t)$ are *normal* on (t^1, t^2) if for each $t \in (t^1, t^2)$ there exists only one linearly independent vector $\lambda(t)$ satisfying the constant and linear necessary conditions (6.4) and (6.5).

COROLLARY 6.3 (Goh [4]). *Assume that* $u^0(t)$ *and* $x(t)$ *are defined for* (2.1) *on* $[t^0, t^e]$. *Suppose* $u^0(t) \in$ interior Ω *and is normal on the subinterval* (t^1, t^2). *If* $u^0(t)$ *is minimal, then there exists a* $\lambda(t)$ *satisfying the PMP on* $[t^0, t^e]$, *which is unique to the scalar multiple by normality. Let* h_i *be the smallest integer such that*

$$\frac{\partial}{\partial u_i} \frac{d^{h_i+1}}{dt^{h_i+1}} \frac{\partial}{\partial u_i} H(\lambda(t), x(t), u^0(t)) \neq 0$$

for some $t \in (t^1, t^2)$. *Then each finite* h_i *must be odd and on the subinterval* $[t^1, t^2]$ *such that*

$$\frac{\partial}{\partial u_i}\frac{d^k}{dt^k}\frac{\partial}{\partial u_j}H(\lambda(t), x(t), u^0(t)) = 0$$

must hold for $k = 0, \cdots, (h_i + h_j)/2$, $1 \leqq i, j \leqq l$. *Moreover if* $h_i < \infty$ *for* $i = 1, \cdots, k \leqq l$, *then the* $k \times k$ *matrix whose i, j entry is*

$$(-1)^{(h_j+1)/2}\frac{\partial}{\partial u_i}\frac{d^{(h_i+h_j)/2+1}}{dt^{(h_i+h_j)/2+1}}\frac{\partial}{\partial u_j}H(\lambda(t), x(t), u^0(t)),$$

where $1 \leqq i, j \leqq k$ *must be symmetric and nonpositive definite.*

Remark 1. Goh [4] does not express his necessary conditions in above form, but they are equivalent. The Hamiltonian formulation of Corollary 6.3 makes the conditions easier to describe and apply. They are closer to Robbins [15], but his results are weaker for they do not include quadratic conditions involving two controls which are singular of differing degrees.

Remark 2. The above results make it desirable to choose coordinates in the control space $\Omega \subseteq \mathbb{R}^l$ so that the controls $u_1, \cdots, u_l$ are singular of as high a degree as possible. We discuss how this is done in Examples 6.1 and 6.2.

Proof. We now prove Theorem 6.1, Theorem 6.2 follows by prolongation as before, and the corollary follows immediately from Theorem 6.2.

By repeated application of the Jacobi identity,

$$\begin{aligned}[a_i, ad^k(a_0)a_j] = &\sum_{\sigma=0}^{\rho-1}(-1)^\sigma[a_0[ad^\sigma(a_0)a_i, ad^{k-\sigma-1}(a_0)a_j]] \\ &+(-1)^\rho[ad^\rho(a_0)a_i, ad^{k-\rho}(a_0)a_j].\end{aligned} \tag{6.6}$$

Letting $\rho = k$, we obtain

$$\begin{aligned}[a_i, ad^k(a_0)a_j] = &(-1)^{k+1}[a_j, ad^k(a_0)a_i] \\ &+\sum_{\sigma=0}^{k-1}(-1)^\sigma[a_0[ad^\sigma(a_0)a_i, ad^{k-\sigma-1}(a_0)a_j]].\end{aligned} \tag{6.7}$$

These equations imply that a basis for the linear space D_k^2 of Lie elements consists of the union of (5.5) with

$$\{ad^{k-\sigma}(a_0)[a_i, ad^\sigma(a_0)a_j]: 0 \leqq \sigma \leqq k, 1 \leqq i < j \leqq k\}. \tag{6.8}$$

Note the presence of terms like $[a_i, ad^k(a_0)a_j]$ in this basis even when k is even.

If $i = j$, then (6.2) follows immediately because u_i is assumed singular of degree $h_i + 1$. Suppose $i \neq j$ and $k \leqq (h_i + h_j)/2 - 1$. Choose any k_i and k_j such that $k_i + k_j = k$ and $k_i \leqq (h_i - 1)/2$, $k_j \leqq (h_j - 1)/2$. Choose $r_i > k_i$, $r_j > k_j$ such that $k_i + r_i = k_j + r_j$. Define a pair of control variations

$$\beta^{\pm}(s)x = \zeta_{r_i}^{\pm k_i}(s)\xi_{r_j}^{+k_j}(s)\zeta_{r_i}^{\mp k_i}(s)\xi_{r_j}^{-k_j}(s)\gamma^0(-2p(s) - 2q(s))x,$$

where

$$\zeta_{r_i}^{\pm k_i}(s)x = \alpha_{r_i}^{\pm k_i}(s)\gamma^0(p(s))x$$

using the control a_i instead of a_1;

$$\xi_{r_j}^{\pm k_j}(s)x = \alpha_{r_j}^{\pm k_j}(s)\gamma^0(q(s))x$$

using the control a_j and $p(s) = k_i s + 2^{k_i} s^{r_i}$, $q(s) = k_j s + 2^{k_j} s^{r_j}$. In other words, $\beta^{\pm}(s)x$ are constructed by "adding" $\alpha_{r_i}^{+k_i}(s)x$ and $\alpha_{r_i}^{-k_i}(s)x$ made with a_i to $\alpha_{r_j}^{+k_j}(s)x$ and $\alpha_{r_j}^{-k_j}(s)x$ made with a_j.

From the definitions of $\zeta_{r_i}^{\pm k_i}$ and $\xi_{r_j}^{\pm k_j}(s)x$ and the chain rule we have

$$\frac{d^\rho}{ds^\rho}\zeta_{r_i}^{\pm k_i}(0)x = \frac{d^\rho}{ds^\rho}\gamma^0(p(0))x,$$

$$\frac{d^\rho}{ds^\rho}\xi_{r_j}^{\pm k_j}(0)x = \frac{d^\rho}{ds^\rho}\gamma^0(q(0))x$$

if $1 \leqq \rho < k_i + r_i$, and

$$\frac{d^{k_i+r_i}}{ds^{k_i+r_i}}\zeta_{r_i}^{\pm k_i}(0)x = \frac{d^{k_i+r_i}}{ds^{k_i+r_i}}\gamma^0(p(0))x \pm (-1)^{k_i}\frac{(r_i+k_i)!}{k_i!}ad^{k_i}(a_0)a_i(x),$$

$$\frac{d^{k_i+r_i}}{ds^{k_i+r_i}}\xi_{r_j}^{\pm k_j}(0)x = \frac{d^{k_i+r_i}}{ds^{k_i+r_i}}\gamma^0(q(0))x \pm (-1)^{k_j}\frac{(r_j+k_j)!}{k_j!}ad^{k_j}(a_0)a_j(x).$$

Using this we see that

$$\begin{aligned}\frac{d^\rho}{ds^\rho}\beta^{\pm}(0)x = {} & \frac{d}{ds^\rho}\alpha_{r_i}^{\pm k_i}(0)x + \frac{d}{ds^\rho}\gamma^0(p(0))\alpha_{r_j}^{+k_j}(s)\gamma^0(-p(0))x \\ & + \frac{d}{ds^\rho}\gamma^0(p(0)+q(0))\alpha_{r_i}^{\mp k_i}(0)\gamma^0(-p(0)-q(0))x \\ & + \frac{d}{ds^\rho}\gamma^0(2p(0)+q(0))\alpha_{r_j}^{-k_j}(0)\gamma^0(-2p(0)-q(0))x\end{aligned} \tag{6.9}$$

if $1 \leqq \rho < 2(k_i + r_i)$. If $\rho = 2(k_i + r_i)$ we have the above plus the extra terms on the right,

$$\mp(-1)^{k_i+k_j}\frac{((k_i+r_i)!)^2}{k_i!k_j!}[ad^{k_i}(a_0)a_i, ad^{k_j}(a_0)a_j]. \tag{6.10}$$

We wish to make $\beta^{\pm}(s)x$ into a control variation of order $2(k_i + r_i)$ by "adding" other variations to cancel out the right side of (6.9) for $\rho = 1, \cdots, 2(k_i + r_i)$.

Since $\alpha_{r_i}^{\pm k_i}(s)x$ and $\alpha_{r_j}^{\pm k_j}(s)x$ are "added" to make $\beta^{\pm}(s)x$ and the former are of order $k_i + r_i$, so must be the latter. Moreover the $k_i + r_i$ derivative of $\beta^{\pm}(s)x$ is just the sum of the corresponding derivatives of the former and hence zero. Therefore $\beta^{\pm}(s)x$ are control variations of order at least $k_i + r_i + 1$ independent of a_0, a_i and a_j and so all their derivatives up to $2(k_i + r_i) + 1$ are Lie elements.

Studying the right side of (6.9) for $\rho \leqq 2(k_i + r_i)$ we see that it contains no cross terms, i.e., terms with both an a_1 and an a_2 factor. It involves only linear brackets and brackets either quadratic in a_i or quadratic in a_j. By the relationship of k_i *and* k_j to the degrees of singularity h_i and h_j, it follows that the right side of (6.9) along $x(t)$ is in $D^1(x(t))$. Therefore it can be canceled using Lemma (4.4).

The result is a pair of control variations of order $2(k_i+r_i)$ whose $2(k_i+r_i)$ derivatives are given by (6.10). Using (2.9) of the HMP it follows that if $u^0(t)$ is minimal, then

$$\lambda(t)[ad^{k_i}(a_0)a_i, ad^{k_j}(a_0)a_j](x(t))=0. \tag{6.11}$$

Differentiating (6.11) with respect to time yields

$$\lambda(t)ad^{\rho}(a_0)[ad^{k_i}(a_0)a_i, ad^{k_j}(a_0)a_j](x(t))=0,$$

and so (6.6) and induction on $k=k_i+k_j$ implies (6.2).

To show (6.3) we first assume that each $h_i=h$ for $i=1,\cdots,k$. Let $M(t)$ be the $k\times k$ matrix whose i,j entry is

$$(-1)^{(h+1)/2}\lambda(t)[a_i, ad^h(a_0)a_j](x(t)).$$

It follows from (6.2) and (6.7) for odd h that $M(t)$ is symmetric. To show that $M(t)$ is nonpositive definite along minimal trajectories is equivalent to showing that for any $u(t)=(u_1(t),\cdots,u_k(t))$,

$$u(t)^T M(t)u(t)\leqq 0$$

along minimal trajectories.

Given any such $u(t)$, define a new system with scalar control v by

$$\dot{x}=b_0(x)+vb_1(x), \tag{6.12}$$

where $b_0(x)=a_0(x)$ and $b_1(x)=\sum u_i(x_0)a_i(x)$. Applying Theorem 5.1 to (6.12) we see that the control v is singular of degree $h+1$ and we obtain the necessary condition

$$(-1)^{(h+1)/2}\lambda(t)[b_1, ad^h(b_0)b_1](x(t))\leqq 0. \tag{6.13}$$

Now

$$\begin{aligned}[b_1, ad^h(b_0)b_1](x(t)) &= \sum_{i,j} u_i(x_0)u_j(x_0)[a_i, ad^h(a_0)a_j](x(t))\\ &\quad +\text{terms from } D_\rho^2(x(t)) \quad \text{for } \rho<h.\end{aligned} \tag{6.14}$$

So, applying (6.13) and (6.2) to (6.14) yields the desired result.

Suppose the degrees of singularity of the various controls are not the same. If we prolong u_i, i.e., define a new state $x_{n+1}=u_i$ and new control $v=\dot{x}_{n+1}$, then we obtain a system of the form

$$\dot{\mathbf{x}}=b_0(\mathbf{x})+vb_i(\mathbf{x})+\sum_{j\neq i} u_j b_j(\mathbf{x}),$$

where $\mathbf{x}=(x, x_{n+1})$ and

$$b_0(\mathbf{x})=\begin{pmatrix} a_0(x)+x_{n+1}a_i(x)\\ 0\end{pmatrix},\qquad b_i(\mathbf{x})=\begin{pmatrix}0\\1\end{pmatrix},\qquad b_j(\mathbf{x})=\begin{pmatrix}a_j(x)\\0\end{pmatrix}.$$

Along $x_{n+1}=0$ for $j\neq i$,

$$[b_i, ad^{\rho}(b_0)b_i](\mathbf{x})=\begin{pmatrix}-\sum_{\sigma=0}^{\rho-2} ad^{\rho-\sigma-2}(a_0)[a_i, ad^{\sigma}(a_0)a_i](x)\\ 0\end{pmatrix},$$

$$[b_j, ad^{\rho}(b_0)b_j](\mathbf{x})=\begin{pmatrix}[a_j, ad^{\rho}(a_0)a_j(x)]\\ 0\end{pmatrix},$$

$$[b_i, ad^{\rho}(b_0)b_j](\mathbf{x})=\begin{pmatrix}\sum_{\sigma=0}^{\rho-1} ad^{\rho-\sigma-1}(a_0)[a_i, ad^{\sigma}(a_0)a_j](x)\\ 0\end{pmatrix},$$

$$[b_j, ad^{\rho}(b_0)b_i](\mathbf{x})=\begin{pmatrix}-[a_j, ad^{\rho-1}(a_0)a_i](x)\\ 0\end{pmatrix}.$$

Therefore the degree of singularity of v us h_i+3. In this way, all the controls can be made singular of the same degree and (6.3) follows from repeated use of the above identities. Q.E.D.

Example 6.1. Suppose $u^0(t)$ generates a normal singular extremal for (2.1) on $[t^1, t^2]$ and we wish to apply Corollary 6.3. To obtain as many necessary conditions as possible it is desirable to make a time-dependent change of coordinates of the control space $\Omega\subseteq\mathbb{R}^l$.

Start with the symmetric $l\times l$ matrix

$$M_0(t)=\left(\frac{\partial^2}{\partial u_i\partial u_j}H(\lambda(t), x(t), u^0(t))\right),$$

where $1\leqq i, j\leqq l$ and $\lambda(t)$ is uniquely determined to the scalar multiple by normality. By passing to a subinterval if necessary, we can assume that rank $M_0(t)$ is constant and equal to $l-l_1$ where $l_1\neq 0$ by assumption. There exists an orthonormal basis $e_0^1(t), \cdots, e_0^l(t)$ for $\mathbb{R}^l$ such that

$$M_0(t)e_0^i(t)=0$$

for $i=1, \cdots, l_1$. Make the change of coordinates

$$u^{(0)}=E_0(t)u^{(1)},$$

where $u^{(0)}=(u_1^{(0)}, \cdots, u_l^{(0)})$ are the original coordinates, $u^{(1)}=(u_1^{(1)}, \cdots, u_l^{(1)})$ are the new coordinates and

$$E_0(t)=(e_0^1(t)\vdots\cdots\vdots e_0^l(t)).$$

On the subspace $\mathbb{R}^{l_1}\subseteq\mathbb{R}^l$ spanned by $e_0^1(t), \cdots, e_0^{l_1}(t)$, we continue to change coordinates. Consider the $l_1\times l_1$ matrix

$$M_1(t)=\left(\frac{\partial}{\partial u_i}\frac{d^2}{dt^2}\frac{\partial}{\partial u_j}H(\lambda(t), x(t), u^0(t))\right),$$

where $1\leqq i, j\leqq l_1$. This matrix is symmetric, and by passing to a subinterval if necessary, we can assume it is of rank l_1-l_2 for each $t\in(t^1, t^2)$. Choose an orthonormal basis $e_1^1(t), \cdots, e_1^{l_1}(t)$ for $\mathbb{R}^{l_1}$ such that

$$M_1(t)e_1^i(t)=0$$

for $i = 1, \cdots, l_2$. Make the change of coordinates

$$u^{(2)} = E_2(t)u^{(1)},$$

where

$$E_2(t) = \left(\begin{array}{c|cc} e_1^1(t) \cdots e_1^{l_1}(t) & 0 & \\ \hline 0 & 1 & 0 \\ & 0 & 1 \end{array}\right)$$

We continue on in this fashion until some $M_k(t)$ is of full rank or it becomes clear that some controls are singular of infinite degree. Then apply the corollary.

Example 6.2. Suppose $u^0(t)$ generates a singular extremal for (2.1) on $[t^1, t^2]$ which is not normal and we wish to apply Theorem 6.2. Once again, to obtain as many necessary conditions as possible, we must make a change of coordinates in the control space.

Since the trajectory is not normal at each $t \in (t^1, t^2)$, there exists more than one linearly independent λ satisfying the constant and linear necessary conditions (6.4) and (6.5). By passing to a subinterval if necessary, we can assume that the dimension of the space of such λ is constant, say ρ, at each $t \in (t^1, t^2)$, and therefore there exist ρ linearly independent solutions $\lambda^1(t), \cdots, \lambda^\rho(t)$ of the adjoint differential equation (2.6) satisfying (6.4) and (6.5).

Define ρ symmetric $l \times l$ matrices:

$$M_0^\sigma(t) = \left(\frac{\partial^2}{\partial u_i \partial u_j} H(\lambda^\sigma(t), x(t), u^0(t))\right)$$

where $1 \leqq \sigma \leqq \rho$ and $1 \leqq i, j \leqq l$. By passing to a subinterval if necessary we can assume that the rank of the $\rho \cdot l \times l$ matrix,

$$M_0(t) = \begin{pmatrix} M_0^1(t) \\ \vdots \\ M_0^\rho(t) \end{pmatrix},$$

is constant, say $l - l_1$. Choose an orthonormal basis $e_0^1(t), \cdots, e_0^l(t)$ for $\mathbb{R}^l$ such that

$$M_0(t)\, e_0^i(t) = 0$$

for $i = 1, \cdots, l_1$. Make the change of coordinates

$$u^{(0)} = E_0(t)u^{(1)}$$

where $u^{(0)}$, $u^{(1)}$ and $E_0(t)$ are as in Example 6.1.

On the subspace $\mathbb{R}^{l_1}$ spanned by $e_0^1(t), \cdots, e_0^{l_1}(t)$, we continue to change coordinates. Consider the ρ symmetric $l_1 \times l_1$ matrices

$$M_1^\sigma(t) = \left(\frac{\partial}{\partial u_i} \frac{d^2}{dt^2} \frac{\partial}{\partial u_j} H(\lambda^\sigma(t), x(t), u^0(t))\right),$$

where $1 \leqq \sigma \leqq \rho$ and $1 \leqq i, j \leqq l_1$. Define

$$M_1(t) = \begin{pmatrix} M_1^1(t) \\ \vdots \\ M_1^\rho(t) \end{pmatrix}$$

and so on until some $M_k(t)$ is of full rank or it becomes clear that some controls are singular of infinite degree. Then apply Theorem 6.2.

7. Conclusion. The purpose of this paper was to introduce the HMP as a useful tool for constructing high order necessary conditions for optimal control problems with terminal constraints. The HMP is a natural extension of the PMP based on a generalized form of the Pontryagin–Weierstrass condition.

We used the HMP to rigorously demonstrate the GLC for problems with terminal constraints with or without normality. Heretofore the proofs of the GLC relied on a blanket assumption of normality to guarantee their validity.

The HMP can also be used to develop necessary conditions specifically tailored for the problem of interest as in Example 5.3. These special conditions might involve cubic or higher effects of control variations. Further research is needed to discover whether they can be put in a systematic form.

REFERENCES

[1] D. J. BELL, *Singular problem in optimal control—A survey*, Control Systems Centre Report 249, University of Manchester Institute of Science and Technology, Manchester, England, 1974.

[2] R. GABASOV AND F. M. KIRILLOVA, *High order necessary conditions for optimality*, this Journal, 10 (1972), pp. 127–168.

[3] B. S. GOH, *The second variation for the singular Bolza problem*, this Journal, 4 (1966), pp. 309–325.

[4] ———, *Necessary conditions for singular extremals involving multiple control variables*, this Journal, 4 (1966), pp. 716–731.

[5] H. HALKIN, *A maximum principle of the Pontryagin type for systems described by nonlinear difference equations*, this Journal, 4 (1966), pp. 90–111.

[6] N. JACOBSON, *Lie Algebras*, Wiley-Interscience, New York, 1962.

[7] H. J. KELLEY, *A second variation test for singular extremals*, AIAA J., 2 (1964), pp. 1380–1382.

[8] H. J. KELLEY, R. E. KOPP AND H. G. MOYER, *Singular extremals*, Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967, pp. 63–101.

[9] R. E. KOPP AND H. G. MOYER, *Necessary conditions for singular extremals*, AIAA J., 3 (1965), pp. 1439–1444.

[10] A. J. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems*, this Journal, 12 (1974), pp. 43–52.

[11] ———, *The high order maximal principle*, Geometric Methods in System Theory, D. Q. Mayne and R. W. Brockett, eds., D. Reidel, Dordrecht, Holland, 1973, pp. 174–184.

[12] J. P. MCDANELL AND W. F. POWERS, *Necessary conditions for joining optimal singular and nonsingular subarcs*, this Journal, 9 (1971), pp. 161–173.

[13] H. MAURER, *An example of a continuous junction for a singular control problem of even order*, this Journal, 13 (1975), pp. 899–903.

[14] H. M. ROBBINS, *Optimality of intermediate thrust arcs of rocket trajectories*, AIAA J., 3 (1965). pp. 1094–1098.

[15] ———, *A generalized Legendre–Clebsch condition for the singular cases of optimal control*, IBM J. Res. Develop., 11 (1967), pp. 361–372.

[16] H. J. SUSSMANN, *Minimal realizations of nonlinear systems*, Geometric Methods in Systems Theory, D. Q. Mayne and R. W. Brockett, eds., D. Riedel, Dordrecht, Holland, 1973, pp. 243–252.

[17] K. TAIT, *Singular problems in optimal control*, Ph.D. thesis, Harvard University, Cambridge, MA, 1965.

[18] D. H. JACOBSON, *Totally singular quadratic minimization problems*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 651–658.

[19] H. HERMES AND J. P. LA SALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[20] H. HERMES, *High order algebraic conditions for controllability*, Proc. Undine Conference on Algebraic System Theory, 1975.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 2, February 1977

DIFFERENTIAL STABILITY IN NONLINEAR PROGRAMMING*

JACQUES GAUVIN† AND JON W. TOLLE‡

Abstract. This paper consists of a study of stability and differential stability in nonconvex programming. For a program with equality and inequality constraints, upper and lower bounds are estimated for the potential directional derivatives of the perturbation function (or the extremal-value function). These results are obtained with the help of a constraint qualification which is shown to be necessary and sufficient to have bounded multipliers. New results on the continuity of the perturbation function are also obtained.

Introduction. This paper is concerned with the differential properties of the extremal-value function for a nonconvex nonlinear program under perturbations of the right-hand side. In particular, conditions will be given under which bounds can be derived for the directional derivatives of this function. In order to obtain these results, conditions for the stability of the solution set, the multiplier set, and the extremal-value function are derived.

Stability and differential stability have been well-studied for convex programs. Rockafellar [15], [16] has established and utilized the connection between stability and duality. For a survey of this and related work see Geoffrion [6]. Gol'stein [7] has obtained a strong result on the directional differentiability of the extremal-value function under weak assumptions. Hogan [10] also considers these same directional derivatives for use in constructing an optimization algorithm. Williams [18] and Robinson [14] have investigated differential stability for the more special case of linear systems.

In the nonconvex case, Evans and Gould [4] have given sufficient conditions for stability (continuity of the perturbation function) of a program with inequality constraints for "right-hand side" perturbations. Greenberg and Pierskalla [8] have obtained some extensions of these results to functional perturbations and to programs with equality constraints. In reference [9] Hogan gives extensions to parameterized nonlinear programs and Robinson [13] has estimated changes in the solution sets for quite general parameterized nonlinear programs. For differential stability in nonconvex programming fewer results have been published. The standard theorem seems to be of Fiacco and McCormick [5] in which the optimal objective value in a neighborhood of a local optimum is shown, under strong assumptions, to be differentiable with its gradient equal to the Kuhn–Tucker multiplier vector corresponding to that local optimum.

Recently there have appeared some papers which have investigated the stability of nonconvex programs through the use of penalty functions and augmented Lagrangian functions. Rockafellar [17] shows the important role played by differential stability of degree 2 in nonconvex duality theory and for the existence of a saddle point for a certain augmented Lagrangian. Conditions, which include the classical second order sufficient conditions to have an isolated

* Received by the editors March 17, 1975, and in revised form May 18, 1976.

† Ecole Polytechnique, Université de Montréal, Montréal, Québec H3C 3A7 Canada. This work was supported in part by the National Research Council of Canada under Grant A-9273.

‡ Department of Mathematics and Curriculum in Operations Research and Systems Analysis, University of North Carolina, Chapel Hill, North Carolina 27514.

optimum, are shown to be sufficient for this stability of degree 2. Armacost and Fiacco [1] study functionally perturbed nonconvex programs through the use of penalty functions. They compute first and second order changes in the local extremal-value function via these penalty functions.

In § 1 of this paper, the basic terminology and notation used as well as some fundamental results from matrix theory, point-to-set maps, and the theory of differential equations are presented. In § 2, some relations between stability and the set of Kuhn–Tucker vectors for a mixed nonlinear program are established. A known constraint qualification is shown to be necessary and sufficient to have bounded Kuhn–Tucker vectors. This constraint qualification is then used to derive the continuity of the extremal-value function. In § 3, lower and upper bounds for the potential directional derivatives of the extremal-value function are given as well as conditions for the existence of the directional derivatives.

1. Preliminaries. In the paper we will consider the mixed nonlinear program

$$\begin{aligned} &\text{maximize } f(x), \qquad x \in R^n, \\ &\text{subject to} \\ &g_i(x) \leqq 0, \qquad i = 1, \cdots, m, \\ &h_j(x) = 0, \qquad j = 1, \cdots, p, \end{aligned} \tag{1.1}$$

and its perturbed program

$$\begin{aligned} &\text{maximize } f(x), \qquad x \in R^n, \\ &\text{subject to} \\ &g_i(x) \leqq b_i, \qquad i = 1, \cdots, m, \\ &h_j(x) = c_j, \qquad j = 1, \cdots, p, \end{aligned} \tag{1.2}$$

where the vectors b and c represent "small perturbations". The functions f, $\{g_i\}$, and $\{h_j\}$ will be assumed to be continuously differentiable functions defined on R^n unless otherwise specified. The gradients of these functions will be denoted by ∇f, $\{\nabla g_i\}$, and $\{\nabla h_j\}$ respectively. All vectors will be used interchangeably as row and column vectors and the inner product of any two vectors x and y will be denoted $x \cdot y$.

For fixed b and c the feasible region for (1.2) will be denoted by $U(b, c) = S(b) \cap T(c)$ where $S(b) = \{x : g_i(x) \leqq b_i,\ i = 1, \cdots, m\}$ and $T(c) = \{x : h_j(x) = c_j,\ j = 1, \cdots, p\}$. The feasible perturbations are $D = \{(b, c) : U(b, c) \neq \varnothing\}$ and the extremal-value function is the function $f_{\sup}: R^m \times R^p \to R$ defined by

$$f_{\sup}(b, c) = \begin{cases} \sup\{f(x) : x \in U(b, c)\}, & U(b, c) \neq \varnothing, \\ -\infty, & U(b, c) = \varnothing. \end{cases}$$

For $(b, c) \in D$, we define the optimal set $P(b, c)$ as follows:

$$P(b, c) = \{x \in U(b, c) : f(x) = f_{\sup}(b, c)\}.$$

We shall sometimes consider $U(b, c)$ and $P(b, c)$ to be point-to-set mappings from $R^m \times R^p$ into the subsets of R^n.

For $\bar{x} \in U(b, c)$ we denote the active inequality constraints for (1.2) as

$$I(\bar{x}; b) = \{i : g_i(\bar{x}) = b_i\}.$$

Also we denote by $K(\bar{x}; b, c)$ the set of Kuhn–Tucker vectors corresponding to $\bar{x}$, that is, the set of $(u, w) \in R^m \times R^p$ such that

$$\begin{aligned} &\sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) + \sum_{j=1}^{p} w_j \nabla h_j(\bar{x}) = \nabla f(\bar{x}), \\ &u_i \geqq 0, \qquad i = 1, \cdots, m, \\ &u_i(g_i(\bar{x}) - b_i) = 0, \qquad i = 1, \cdots, m, \end{aligned} \tag{1.3}$$

and define

$$K(b, c) = \bigcup_{\bar{x} \in P(b, c)} K(\bar{x}; b, c).$$

In order to assure that $K(\bar{x}; b, c)$ be nonempty when $\bar{x}$ is a local maximum of (1.2) it is necessary to impose some type of constraint qualification on the functions $\{g_i\}$ and $\{h_j\}$ at $\bar{x}$. In this paper the following two qualifications will be used:

(i) There exists a $\tilde{y} \in R^n$ such that

$$\begin{aligned} &\nabla g_i(\bar{x}) \cdot \tilde{y} < 0, \qquad i \in I(\bar{x}; b), \\ &\nabla h_j(\bar{x}) \cdot \tilde{y} = 0, \qquad j = 1, \cdots, p. \end{aligned} \tag{CQ1)*}$$

(ii) The gradients $\{\nabla h_j(\bar{x})\}$, $j = 1, \cdots, p$ are linearly independent.

The condition (CQ1)* is the Mangasarian–Fromowitz constraint qualification (see [11]).

$$\begin{aligned} &\text{The gradients } \{\nabla g_i(\bar{x}), \nabla h_j(\bar{x})\}, i \in I(\bar{x}; b), \\ &j = 1, \cdots, p, \text{ are linearly independent.} \end{aligned} \tag{CQ2)*}$$

In the absence of equality constraints (CQ1)* is equivalent to the Cottle constraint qualification: the system

$$\sum_{i \in I(\bar{x}; b)} u_i \nabla g_i(\bar{x}) = 0, \qquad u_i \geqq 0,$$

has no nonzero solution. If the g_i are convex and the h_j affine, (CQ1)* is the well-known Slater condition.

In this paper we shall have occasion to use the pseudoinverse of a matrix. If A is $m \times n$ with rank m then AA^T is nonsingular and the matrix $A^{\#} = A^T(AA^T)^{-1}$ is called the pseudoinverse of A. The following properties of $A^{\#}$ will be useful.

PROPOSITION 1.1. *If A is $m \times n$ with rank m then*

(i) $AA^{\#} = I_m$,

(ii) $A^{\#}b$ *is a solution of* $Ax = b$,

(iii) *for any* $\hat{x}$ *which is a solution of* $Ax = 0$ *there exists a* $\hat{z} \in R^n$ *such that*

$$\hat{x} = (I_n - A^{\#}A)\hat{z},$$

(iv) *any* $\bar{x}$ *which is a solution of* $Ax = b$ *can be written*

$$\bar{x} = A^{\#}b + (I_n - A^{\#}A)\bar{x}.$$

For completeness, we include some important concepts from the theory of point-to-set maps. For a more detailed exposition the reader is referred to Hogan [9] or Berge [2]. In the following, F will represent a point-to-set mapping of $E \subseteq R^k$ into the subsets of R^l.

DEFINITION 1.1. The map F is *upper semi-continuous* at $\bar{x} \in E$, if for any open set $\mathcal{O}$ containing $F(\bar{x})$ there is an open neighborhood $N(\bar{x})$ of $\bar{x}$ such that $F(x) \subseteq \mathcal{O}$ for each $x \in N(\bar{x}) \cap E$.

The concept of upper semi-continuity will be more suited to our purposes if phrased in terms of sequences.

DEFINITION 1.2. F is said to be *closed* at $\bar{x} \in E$ if $\{x_n\} \subset E$, $x_n \to \bar{x}$, $y_n \in F(x_n)$, and $y_n \to \bar{y}$ imply that $\bar{y} \in F(\bar{x})$.

DEFINITION 1.3. F is *uniformly compact* near $\bar{x} \in E$ if there is a neighborhood $N(\bar{x})$ of $\bar{x}$ such that the closure of the set $\bigcup_{x \in N(\bar{x})} F(x)$ is compact.

With the above definitions we can relate the concepts of closed maps and upper semi-continuity. The proof of the following proposition is found in [9].

PROPOSITION 1.2. *Let F be uniformly compact near $\bar{x} \in E$. Then F is closed at $\bar{x}$ if and only if $F(\bar{x})$ is compact and F is upper semi-continuous at $\bar{x}$.*

Finally we state two results on the solution of nonlinear differential equations which will be useful in the sequel.

Let K and L be open subsets of R^n and R^p respectively and let ϕ be a continuous mapping from $K \times L$ into a bounded subset of R^n. Consider the two-parameter family of ordinary differential equations with specified initial values

$$\psi'(t) = \phi(\psi(t), \xi),$$

$$\psi(t_0) = \eta$$

where $\xi \in L$ and $\eta \in K$. The following proposition is an easy generalization of Peano's existence theorem (see, for example, [3]).

PROPOSITION 1.3. *Let $\bar{L}$ be any compact subset of L. Then for $\eta^0 \in K$ there exists a $\delta > 0$ such that for every η such that $\|\eta - \eta^0\| \leqq \delta$ and every $\xi \in \bar{L}$ there exists at least one solution to* (1.4). *Moreover, all solutions to* (1.4) *for $\xi \in \bar{L}$ and $\|\eta - \eta^0\| \leqq \delta$ exist and form an equicontinuous and uniformly bounded family on some fixed interval $[t_0, t_1]$, $t_1 > t_0$.*

By using the Ascoli–Arzela theorem together with the integral equation form of (1.4) the following corollary to Proposition 1.3 can be deduced.

PROPOSITION 1.4. *Let ξ^n be a sequence of vectors in $\bar{L}$ converging to ξ^0 and η^n be a sequence of vectors in K converging to $\eta^0 \in K$. If $\psi_n(t)$ is a solution of* (1.4)

corresponding to $\xi = \xi^n$ *and* $\eta = \eta^n$ *then the sequence* $\{\psi_n(t)\}$ *contains a subsequence,* $\{\psi_m(t)\}$, *which converges uniformly to* $\psi_0(t)$ *on* $[t_0, t_1]$. *Moreover* $\psi_0(t)$ *is a solution to* (1.4) *for* $\xi = \xi^0$ *and* $\eta = \eta^0$ *and* $\psi'_m(t)$ *converges to* $\psi'_0(t)$ *pointwise on* $[t_0, t_1]$.

2. Stability and the Kuhn–Tucker vectors. In this section we investigate some stability properties of the program (1.2). The stability of a nonlinear program is traditionally measured in terms of the semi-continuity properties of the point-to-set maps $U(b, c)$ and $P(b, c)$ and the function $f_{\sup}(b, c)$. Here we show how these properties for $f_{\sup}(b, c)$ derive from the basic assumptions of uniform compactness of $U(b, c)$ and the existence of optimal points where the constraint qualification (CQ1)* holds. In addition, the concept of stability is extended to include the behavior of the set of multiplier vectors, $K(\bar{x}; b, c)$ for $\bar{x} \in P(b, c)$. Besides their intrinsic interest, many of these results will have important consequences in the developments in § 3.

The first result, due to Evans and Gould [4] and Greenberg and Pierskalla [8], shows the importance of the uniform compactness assumption on the feasible set.

LEMMA 2.1. *If* $U(0, 0)$ *is nonempty and* $U(b, c)$ *is uniformly compact near* $(0, 0)$ *then* $U(b, c)$ *and the extremal-value function* $f_{\sup}(b, c)$ *are upper semi-continuous at* $(0, 0)$.

Proof. Because the functions g_i and h_j are continuous $U(b, c)$ is closed at $(0, 0)$ and therefore, by Proposition 1.2 it is upper semi-continuous at $(0, 0)$. The upper semi-continuity of $f_{\sup}$ at $(0, 0)$ now follows from the results in [8]. □

In this study we shall find it necessary that the Kuhn–Tucker vectors for a given optimal point form a compact set. The next result shows that the constraint qualification (CQ1)* is necessary as well as sufficient for this property. This theorem complements the recent work of Robinson [13] in which (CQ1)* is shown to be equivalent to a type of local stability for the set $U(b, c)$. Taken together these results strongly suggest that (CQ1)* is the natural constraint qualification to invoke when studying questions of stability for a nonlinear program.

THEOREM 2.2. *Let* $\bar{x}$ *be a local maximum for program* (1.1). *Then* $K(\bar{x}; 0, 0)$ *is a nonempty, compact, and convex set if and only if* (CQ1)* *is satisfied at* $\bar{x}$.

Proof. If $K(\bar{x}; 0, 0)$ is nonempty then it is clearly closed and convex. So it will suffice to show that (CQ1)* is equivalent to $K(\bar{x}; 0, 0)$ being nonempty and bounded.

Consider the linear program

$$\text{minimize} \sum_{i \in I(\bar{x};0)} (-u_i)$$

subject to

$$\sum_{i \in I(\bar{x};0)} u_i \nabla g_i(\bar{x}) + \sum_{j=1}^{p} w_j \nabla h_j(\bar{x}) = \nabla f(\bar{x}),$$

$$u_i \geqq 0, \qquad i \in I(\bar{x}; 0),$$

$$w_j \text{ unrestricted}, \qquad j = 1, \cdots, p,$$

and its dual

$$\begin{aligned}
&\text{maximize } \nabla f(\bar{x}) \cdot y \\
&\text{subject to} \\
&\nabla g_i(\bar{x}) \cdot y \leqq -1, \qquad i \in I(\bar{x}; 0), \\
&\nabla h_j(\bar{x}) \cdot y = 0, \qquad j = 1, \cdots, p, \\
&y \text{ unrestricted.}
\end{aligned}$$

Let $\nabla H(\bar{x})$ be the matrix whose rows are the gradients $\nabla h_j(\bar{x})$. If (CQ1)* holds, the dual is feasible. For any y which is dual feasible there exists a z such that

$$y = [I_n - \nabla H(\bar{x})^{\#} \nabla H(\bar{x})]z$$

where $\nabla H(\bar{x})^{\#}$ is the pseudoinverse of $\nabla H(\bar{x})$ (see Proposition 1.1). Define the arc $\alpha(\lambda)$ by the differential equation

$$\begin{aligned}
&\alpha'(\lambda) = [I_n - \nabla H(\alpha(\lambda))^{\#} \nabla H(\alpha(\lambda))]z, \\
&\alpha(0) = \bar{x}.
\end{aligned}$$

It is easily shown that for small λ this arc is feasible for program (1.1) (see the proof of Lemma 2.4 for a similar argument). Thus for λ sufficiently small, $f(\alpha(\lambda)) \leqq f(\alpha(0)) = f(\bar{x})$ and $\nabla f(\bar{x}) \cdot y = \nabla f(\alpha(0)) \cdot \alpha'(0) \leqq 0$. Hence the dual program is bounded and the primal is feasible and bounded. It follows that the set of feasible u vectors is bounded. From part (ii) of (CQ1)* it follows immediately that the set of feasible w vectors is bounded.

Conversely, if we assume $K(\bar{x}; 0, 0)$ to be nonempty and bounded then the primal problem is feasible and bounded, so the dual is feasible and part (i) of (CQ1)* is satisfied. If the gradients $\{\nabla h_j(\bar{x})\}, j = 1, \cdots, p$ are not linearly independent then for a fixed u the solutions of the system

$$\sum_{j=1}^{p} w_j \nabla h_j(\bar{x}) = \left[\nabla f(\bar{x}) - \sum_{i=1}^{m} u_i \nabla g_i(\bar{x})\right]$$

define an unbounded linear manifold. But this contradicts the assumption and therefore part (ii) of (CQ1)* must hold. □

It should be observed that (CQ1)* does not imply that the Kuhn–Tucker vector is unique as do stronger constraint qualifications, e.g., (CQ2)*.

A standard goal in the study of stability is to assure the continuity of the perturbation function $f_{\sup}(b, c)$. The next set of lemmas culminate with the proof that this continuity is a consequence of the uniform compactness assumption and the constraint qualification (CQ1)* thus extending results given in [4] and [8]. In the process of establishing this theorem a constructive means of obtaining points in $U(b, c)$ for (b, c) near $(0, 0)$ will be given.

Henceforth $z = (z^1, z^2)$ will represent a unit vector in $R^m \times R^p$. Such a z will be called a direction vector. The next two lemmas establish that if (CQ1)* holds then $U(\lambda z^1, \lambda z^2)$ is nonempty for $0 \leqq \lambda \leqq \bar{\lambda}$, for some $\bar{\lambda} > 0$.

LEMMA 2.3. *If* (CQ1)* *is satisfied at some* $\bar{x} \in P(0, 0)$, *then for each direction* z *there exists a* $\bar{y} \in R^n$ *such that*

(i) $\nabla g_i(\bar{x}) \cdot \bar{y} \leqq z_i^1$, $i \in I(\bar{x}; 0)$,

(ii) $\nabla h_j(\bar{x}) \cdot \bar{y} = z_j^2$, $j = 1, \cdots, p$,

(iii) $\nabla f(\bar{x}) \cdot \bar{y} = \min_{(u,w) \in K(\bar{x};0,0)} \{u \cdot z^1 + w \cdot z^2\}$.

Proof. Consider the following linear program

$$
\text{(P)}^* \qquad
\begin{aligned}
&\min \{u \cdot z^1 + w \cdot z^2\} \\
&\text{subject to} \\
&\sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) + \sum_{j=1}^{p} w_j \nabla h_j(\bar{x}) = \nabla f(\bar{x}), \\
&u_i g_i(\bar{x}) = 0, \qquad i = 1, \cdots, m, \\
&u \geqq 0, \\
&w \text{ unrestricted},
\end{aligned}
$$

and its dual

$$
\text{(D)}^* \qquad
\begin{aligned}
&\max \{\nabla f(\bar{x}) \cdot y\} \\
&\text{subject to} \\
&\nabla g_i(\bar{x}) \cdot y + g_i(\bar{x}) v_i \leqq z_i^1, \qquad i = 1, \cdots, m, \\
&\nabla h_j(\bar{x}) \cdot y = z_j^2, \qquad j = 1, \cdots, p, \\
&y, v \text{ unrestricted}.
\end{aligned}
$$

From Theorem 2.2, the primal problem, (P)*, is bounded and feasible. Thus by the dual theorem of linear programming there is an optimal dual solution $\bar{y}$ such that (i)–(iii) hold. □

LEMMA 2.4. *If* (CQ1)* *is satisfied at some* $\bar{x} \in P(0, 0)$ *then, for every direction* z, *there exists a* $\bar{\lambda} > 0$ *and a continuously differentiable function* $\psi(\cdot\,; z) : [0, \bar{\lambda}] \to R^n$ *such that* $\psi(0; z) = \bar{x}$ *and* $\psi(\lambda; z) \in U(\lambda z^1, \lambda z^2)$ *for* $\lambda \in [0, \bar{\lambda}]$.

Proof. Let z be given. Let $\bar{y}$ satisfy (i) and (ii) of Lemma 2.3 and $\tilde{y}$ given by (CQ1)* at $\bar{x}$. Then $\nabla H(\bar{x})(\bar{y} + \tilde{y}) = z^2$. From the properties of the pseudoinverse matrix given in Proposition 1.1 we have

$$\bar{y} + \tilde{y} = \nabla H(\bar{x})^{\#} z^2 + [I_n - \nabla H(\bar{x})^{\#} \nabla H(\bar{x})](\bar{y} + \tilde{y}).$$

Now define $\psi(\lambda; z)$ by the differential equation

$$
\begin{aligned}
&\psi'(\lambda; z) = \nabla H(\psi(\lambda; z))^{\#} z^2 + [I_n - \nabla H(\psi(\lambda; z))^{\#} \nabla H(\psi(\lambda; z))](\bar{y} + \tilde{y}), \\
&\psi(0; z) = \bar{x}
\end{aligned}
\tag{2.1}
$$

and note that

$$\psi'(0; z) = \bar{y} + \tilde{y}.$$

By Proposition 1.3, there exists a solution to this initial value problem in a positive interval, say $|\lambda| \leqq \alpha$, where α is independent of z.

Since

$$\begin{aligned} H(\psi(\lambda; z)) &= \int_0^1 \frac{d}{dt} H(\psi(\lambda t; z))\, dt \\ &= \lambda \int_0^1 \nabla H(\psi(\lambda t; z))\psi'(\lambda t; z)\, dt \\ &= \lambda \int_0^1 z^2\, dt = \lambda z^2 \end{aligned}$$

we have that $h_j(\psi(\lambda; z)) = \lambda z_j^2$, $j = 1, \cdots, p$, for all λ, $0 \leqq \lambda \leqq \alpha$. Moreover, for $i \in I(\bar{x}; 0)$,

$$\begin{aligned} g_i(\psi(\lambda; z)) &= g_i(\bar{x}) + \nabla g_i(\bar{x}) \cdot (\bar{y} + \tilde{y})\lambda + o(\lambda) \\ &< \lambda z_i^1 + o(\lambda) \\ &\leqq \lambda z_i^1 \end{aligned}$$

for λ such that $0 \leqq \lambda \leqq \lambda_i$ for some λ_i, $0 < \lambda_i \leqq \alpha$. Also, for $i \notin I(\bar{x}; 0)$, $g_i(\psi(0; z)) = g_i(\bar{x}) < 0$ implies $g_i(\psi(\lambda; z)) \leqq \lambda z_i^1$ for λ such that $0 \leqq \lambda \leqq \lambda_i$ where $0 < \lambda_i \leqq \alpha$. We may conclude that $\psi(\lambda; z) \in U(\lambda z^1, \lambda z^2)$ for $0 \leqq \lambda \leqq \bar{\lambda} = \min\{\lambda_i\}$. □

The next lemma demonstrates that there exists a $\bar{\lambda} > 0$ which is valid for all direction z and therefore shows that $U(b, c)$ is not empty for (b, c) near $(0, 0)$.

LEMMA 2.5. *If* (CQ1)* *is satisfied at some* $\bar{x} \in U(0, 0)$ *then there exists a* $\bar{\lambda} > 0$ *and for every direction* z *a continuously differentiable function* $\psi(\cdot; z): [0, \bar{\lambda}] \to R^n$ *such that* $\psi(\lambda; z) \in U(\lambda z^1, \lambda z^2)$ *for each* $\lambda \in [0, \bar{\lambda}]$.

Proof. Let $z_0 = (z_0^1, z_0^2)$ be a given direction vector and let $\bar{y}_0 = \nabla H(\bar{x})^{\#} z_0^2$. Let $\tilde{y}$ given by (CQ1)* such that

$$\nabla g_i(\bar{x}) \cdot (\bar{y}_0 + \tilde{y}) < -1, \qquad i \in I(\bar{x}; 0). \tag{2.2}$$

For any direction z in some fixed neighborhood N_0 of z_0 let $\bar{y} = \nabla H(\bar{x})^{\#} z^2$. If N_0 is sufficiently small, then (2.2) holds for $\bar{y}_0$ replace by $\bar{y}$. For each $z \in N_0$, define the function $\psi(\lambda; z)$ by the differential equation (2.1). As in Lemma 2.4, we have $H(\psi(\lambda; z)) = \lambda z^2$ for all λ, $0 \leqq \lambda \leqq \alpha$ and all z and

$$g_i(\psi(\lambda; z)) \leqq \lambda z_i^1, \qquad \lambda \in [0, \bar{\lambda}(z)]$$

for some $\bar{\lambda}(z) > 0$.

We next show that if the neighborhood of z_0 is sufficiently small then the $\bar{\lambda}(z)$ can be chosen independently of z in that neighborhood. Suppose not; then there exists a sequence of direction vectors, $\{z^n\}$, converging to z_0 such that the corresponding sequence $\{\bar{\lambda}_n\} = \{\bar{\lambda}(z^n)\}$ converges to zero. This means that there is an $i \in \{1, \cdots, m\}$, a subsequence $\{z^m\}$ tending to z_0, and a positive sequence $\{\lambda_m\} \searrow 0$ such that

$$g_i(\psi(\lambda_m; z^m)) > \lambda_m (z_i^1)^m. \tag{2.3}$$

By applying Proposition 1.4 we can assume that the sequence $\psi(\lambda; z^m)$ converges uniformly to $\psi(\lambda; z_0)$ on the interval $[0, \alpha]$ where $\psi(\lambda; z_0)$ is a solution to (2.2) for $\bar{y} = \bar{y}_0$. We show, by considering two cases, that (2.3) leads to a contradiction.

Case 1. If $i \notin I(\bar{x}; 0)$ then, by the uniform convergence, (2.3) leads to

$$g_i(\psi(0; z_0)) = g_i(\bar{x}) \geqq 0,$$

which is a contradiction since g_i is inactive at $\bar{x}$.

Case 2. If $i \in I(\bar{x}; 0)$ then $g_i(\psi(0; z^m)) = 0$ and

$$g_i(\psi(\lambda_m; z^m)) - g_i(\psi(0; z^m)) > \lambda_m (z_i^1)^m$$

By the mean value theorem there exists a $\tilde{\lambda}_m \in [0, \lambda_m]$ such that

$$\nabla g_i(\psi(\tilde{\lambda}_m; z^m)) \cdot \psi'(\tilde{\lambda}_m; z^m) > (z_i^1)^m$$

Replacing $\psi'(\tilde{\lambda}_m; z^m)$ by the corresponding right-hand side in (2.1) and using the uniform convergence we obtain

$$\nabla g_i(\bar{x}) \cdot (\bar{y}_0 + \tilde{y}_0) > z_{0_i}^1 \geqq -1,$$

which is a contradiction to (2.2).

We have now shown that in some neighborhood of z_0 there is a $\bar{\lambda}(z_0)$ such that $\psi(\lambda; z) \in U(\lambda z^1, \lambda z^2)$ for all λ, $0 \leqq \lambda \leqq \bar{\lambda}(z_0)$. Applying this result to each direction vector we obtain a cover of the unit sphere in R^{m+p}. By choosing a finite subcover and taking the minimum of the corresponding $\bar{\lambda}(z)$ we obtain the desired result. □

We now establish the continuity of $f_{\sup}$ at $(b, c) = (0, 0)$.

THEOREM 2.6. *Suppose that $U(0, 0)$ is nonempty, $U(b, c)$ is uniformly compact near $(0, 0)$, and* (CQ1)* *holds at some $\bar{x} \in P(0, 0)$. Then $f_{\sup}(b, c)$ is continuous at $(0, 0)$.*

Proof. $f_{\sup}$ is upper semi-continuous at $(0, 0)$ by Lemma 2.1. To show that $f_{\sup}$ is lower semi-continuous, let $\{(b^n, c^n)\}$ be a sequence of points such that

$$\liminf_{(b,c)\to(0,0)} f_{\sup}(b, c) = \lim_{n\to\infty} f_{\sup}(b^n, c^n).$$

Writing (b^n, c^n) as $\lambda^n(\hat{b}^n, \hat{c}^n)$, $\lambda^n \searrow 0$, where $(\hat{b}^n, \hat{c}^n) = z^n$ is a unit vector, we obtain (possibly using a subsequence) a sequence $\{(\hat{b}^n, \hat{c}^n)\}$ converging to the unit vector $(b^0, c^0) = z^0$.

Using Lemma 2.5 we have that $\psi(\lambda^n; z^n) \in U(b^n, c^n)$ for n large enough to insure that $\lambda^n \leqq \bar{\lambda}$. Thus

$$f_{\sup}(b^n, c^n) \geqq f(\psi(\lambda^n; z^n))$$

and by the uniform convergence argument of Lemma 2.5

$$\begin{aligned} \lim_{n\to\infty} f_{\sup}(b^n, c^n) &\geqq \lim_{n\to\infty} f(\psi(\lambda^n; z^n)) \\ &= f(\psi(0; z^0)) \\ &= f(\bar{x}) = f_{\sup}(0, 0). \end{aligned}$$

Consequently, $f_{\sup}$ is lower semi-continuous and hence continuous at $(0, 0)$. □

In the next theorem it is demonstrated that the constraint qualification (CQ1)* is preserved under perturbation. More general constraint qualifications need not have this property.

THEOREM 2.7. *Assume* (CQ1)* *holds at* $\bar{x} \in P(0, 0)$. *Let* $\{(b^n, c^n)\}$ *and* $\{x_n\}$ *be sequences such that* $(b^n, c^n) \to (0, 0)$, $x_n \in P(b^n, c^n)$, *and* $x_n \to \bar{x}$. *Then for* n *sufficiently large,* (CQ1)* *holds at* x_n *and there exist subsequences* $\{(u^m, w^m)\}$ *and* $\{x_m\}$ *with* $(u^m, w^m) \in K(x_m; b^m, c^m)$ *such that* $(u^m, w^m) \to (\bar{u}, \bar{w})$ *for some* $(\bar{u}, \bar{w}) \in K(\bar{x}; 0, 0)$.

Proof. Since $x_n \to \bar{x}$, for n sufficiently large it follows that $I(x_n; b^n) \subseteq I(\bar{x}; 0)$ and that $\{\nabla h_j(x_n)\}$, $j = 1, \cdots, p$, are linearly independent. Since (CQ1)* holds at $\bar{x}$, there exists a $\tilde{y}$ such that

$$\nabla g_i(\bar{x}) \cdot \tilde{y} < 0, \qquad i \in I(\bar{x}; 0),$$

$$\nabla h_j(\bar{x}) \cdot \tilde{y} = 0, \qquad j = 1, \cdots, p.$$

As in the proof of Theorem 2.2 there exists a $z \in R^n$ such that

$$\tilde{y} = [I_n - \nabla H(\bar{x})^{\#} \nabla H(\bar{x})]z.$$

Define

$$\tilde{y}_n = [I_n - \nabla H(x_n)^{\#} \nabla H(x_n)]z.$$

Then $\tilde{y}_n \to \tilde{y}$, so for large n

$$\nabla g_i(x_n) \cdot \tilde{y}_n < 0, \qquad i \in I(x_n; b^n), \tag{2.4}$$

and by definition

$$\nabla h_j(x_n) \cdot \tilde{y}_n = 0, \qquad j = 1, \cdots, p. \tag{2.5}$$

Hence (CQ1)* holds at x_n for n sufficiently large.

Let $(u^n, w^n) \in K(x_n; b^n, c^n)$. Then for each n

$$w^n \nabla H(x_n) = \nabla f(x_n) - \sum_{i=1}^{m} u_i^n \nabla g_i(x_n)$$

and thus

$$w^n = [\nabla f(x_n) - \sum_{i=1}^{m} u_i^n \nabla g_i(x_n)] \nabla H(\bar{x}_n)^{\#}. \tag{2.6}$$

From (2.4) and (2.5) we have

$$\begin{aligned} \nabla f(x_n) \cdot \tilde{y}_n &= \sum_{i=1}^{m} u_i^n \nabla g_i(x_n) \cdot \tilde{y}_n \\ &\leqq u_j^n \nabla g_j(x_n) \cdot \tilde{y}_n \end{aligned}$$

for any $j \in I(\bar{x}; 0)$. Consequently

$$u_j^n \leqq \frac{\nabla f(x_n) \cdot \tilde{y}_n}{\nabla g_j(x_n) \cdot \tilde{y}_n}, \qquad j \in I(\bar{x}; 0),$$

$$u_j^n = 0, \qquad j \notin I(\bar{x}; 0).$$

These inequalities together with (2.6) imply that the sequence $\{(u^n, w^n)\}$ is bounded. Any convergent subsequence must clearly converge to some $(\bar{u}, \bar{w}) \in K(\bar{x}; 0, 0)$. □

From the proof of Theorem 2.7 it is apparent that the sequence $\{x_n\}$ and $\bar{x}$ can just as well be taken to be local maxima rather than global.

As a consequence of Theorem 2.7 we obtain

COROLLARY 2.8. *If $U(0, 0)$ is nonempty and compact, and if* (CQ1)* *holds at each $\bar{x} \in P(0, 0)$ then $K(0, 0)$ is compact.*

Proof. Choose any sequences $\{(u^n, w^n)\}$, $(u^n, w^n) \in K(0, 0)$. Take $x_n \in P(0, 0)$ such that $(u^n, w^n) \in K(x^n; 0, 0)$. Since $P(0, 0)$ is compact, there exists a subsequence $\{x_m\}$ and an $\bar{x} \in P(0, 0)$ such that $x_m \to \bar{x}$. From Theorem 2.7, choosing $(b^m, c^m) = (0, 0)$, there exists a subsequence $\{(u^k, w^k)\}$, $(u^k, w^k) \in K(x_k; 0, 0)$ and a $(\bar{u}, \bar{w}) \in K(\bar{x}; 0, 0)$ such that $(u^k, w^k) \to (\bar{u}, \bar{w})$. □

The next theorem generalizes Theorem 2.7 in that it gives conditions under which sequences $\{x_n\}$ and $\{(u^n, w^n)\}$, with $(u^n, w^n) \in K(x_n; b^n, c^n)$, have convergent subsequences as $(b^n, c^n) \to (0, 0)$ without assuming that $x_n \to \bar{x} \in P(0, 0)$.

THEOREM 2.9. *Suppose $U(0, 0)$ is nonempty and $U(b, c)$ is uniformly compact near* $(0, 0)$ *and that* (CQ1)* *holds at each $\bar{x} \in P(0, 0)$. Then for any sequences $\{(b^n, c^n)\}$, $\{x_n\}$, with $(b^n, c^n) \to (0, 0)$, and $x_n \in P(b^n, c^n)$ there exist subsequences $\{x_m\}$ and $\{(u^m, w^m)\}$ with $(u^m, w^m) \in K(x_m; b^m, c^m)$ such that $x_m \to \bar{x}$ and $(u^m, w^m) \to (\bar{u}, \bar{w})$. Moreover $\bar{x} \in P(0, 0)$ and $(\bar{u}, \bar{w}) \in K(\bar{x}; 0, 0)$.*

Proof. Since $U(b, c)$ is compact, $\{x_n\}$ has a subsequence $\{x_k\}$ such that $x_k \to \bar{x}$. $\bar{x}$ is clearly in $U(0, 0)$. By Theorem 2.6 $f_{\sup}$ is continuous at $(0, 0)$ so

$$f(\bar{x}) = \lim f(x_k) = \lim f_{\sup}(b^k, c^k) = f_{\sup}(0, 0).$$

Hence $\bar{x} \in P(0, 0)$. The theorem is now proved by applying Theorem 2.7 to $\{x_k\}$ and $\{(b^k, c^k)\}$. □

COROLLARY 2.10. *If $U(0, 0)$ is nonempty and $U(b, c)$ is uniformly compact near* $(0, 0)$, *and if* (CQ1)* *holds for every $\bar{x} \in P(0, 0)$, then for some $\delta > 0$* (CQ1)* *holds at each $x \in P(b, c)$ with $\|(b, c)\| \leqq \delta$, and the point-to-set map $K(b, c)$ is closed at* $(0, 0)$.

The proof is a direct consequence of Theorems 2.7 and 2.9.

3. Directional derivatives for the extremal-value function. Throughout this section $z = (z^1, z^2)$ will again represent a direction vector. The directional derivative of $f_{\sup}(b, c)$ at $(0, 0)$ in the direction z is given by

$$D_z f_{\sup}(0, 0) = \lim_{\lambda \to 0^+} \left[\frac{f_{\sup}(\lambda z^1, \lambda z^2) - f_{\sup}(0, 0)}{\lambda} \right] \tag{3.1}$$

providing, of course, that the limit exists. Our purpose is to determine conditions under which the directional derivatives will exist for all z and to compute their values. If the derivatives cannot be shown to exist then we will establish upper and lower bounds on the lim sup and lim inf of the difference quotient appearing in (3.1). These bounds enable us to gauge the rate of variation of $f_{\sup}$ in the direction z at $(0, 0)$ if the derivative is not known. The directional derivatives (or the bounds on the lim sup and lim inf) will be expressed in terms of the direction z and the set $K(0, 0)$, the Kuhn–Tucker vectors at $(0, 0)$. As special cases we will obtain the

known results for convex programs and for the program with constraint functions whose gradients are linearly independent.

The results contained herein can be considered an extension of the work of Gol'stein [7]. Gol'stein proved that for general perturbed convex programs the directional derivatives exist and satisfy a saddle-point condition. No assumptions of convexity are made here but the perturbations are restricted to be changes in the right-hand sides.

Rockafellar [17] has established, using second order conditions, the stability of degree 2 of $f_{\sup}(b)$ which is defined in the following manner: in a neighborhood of $b = 0$ there exists a function $\pi(\cdot)$, twice differentiable, such that $f_{\sup}(b) \leqq \pi(b)$ and $f_{\sup}(0) = \pi(0)$. This form of stability yields certain bounds on the potential directional derivatives of $f_{\sup}$. In this section we obtain sharp bounds under conditions not requiring second order derivatives.

THEOREM 3.1. *Suppose that $U(0, 0)$ is nonempty, $U(b, c)$ is uniformly compact near $(0, 0)$, and* (CQ1)* *holds at some $\bar{x} \in P(0, 0)$. Then for any direction z*

$$\liminf_{\lambda \to 0^+} \frac{f_{\sup}(\lambda z^1, \lambda z^2) - f_{\sup}(0, 0)}{\lambda} \geqq \min_{(u,w) \in K(\bar{x};0,0)} \{u \cdot z^1 + w \cdot z^2\}. \tag{3.2}$$

Proof. For a given z, let $\psi(\lambda; z)$ be the curve given in lemma 2.3. We have

$$\begin{aligned} &\liminf_{\lambda \to 0^+} \frac{f_{\sup}(\lambda x^1, \lambda z^2) - f_{\sup}(0, 0)}{\lambda} \\ &\quad \geqq \liminf_{\lambda \to 0^+} \frac{f(\psi(\lambda; z)) - f(\psi(0; z))}{\lambda} = \frac{d}{d\lambda} f(\psi(\lambda; z)) \\ &\quad = \nabla f(\psi(0; z)) \psi'(0; z) = \nabla f(\bar{x}) \cdot (\bar{y} + \tilde{y}) \end{aligned}$$

where $\bar{y}$ satisfies the conclusions of Lemma 2.3 and $\tilde{y}$ satisfies (CQ1)*. Since the left-hand side of the above inequality is independent of $\tilde{y}$ we can let $\tilde{y} \to 0$ and use (iii) of Lemma 2.3 to obtain the desired result. □

We can derive a sharper bound by strengthening the hypothesis of Theorem 3.1 as follows:

COROLLARY 3.2. *If the hypotheses of Theorem* 3.1 *are satisfied and, in addition,* (CQ1)* *holds at each $\bar{x} \in P(0, 0)$ we have*

$$\begin{aligned} &\liminf_{\lambda \to 0^+} \frac{f_{\sup}(\lambda z^1, \lambda z^2) - f_{\sup}(0, 0)}{\lambda} \\ &\quad \geqq \max_{\bar{x} \in P(0,0)} \min_{(u,w) \in K(\bar{x};0,0)} \{u \cdot z^1 + w \cdot z^2\}. \end{aligned} \tag{3.3}$$

Proof. Corollary 2.8 assures the compactness of $K(0, 0)$ and the result follows by applying Theorem 3.1 to each $\bar{x} \in P(0, 0)$. □

It can be shown that Theorem 3.1 and Corollary 3.2 can be valid for some specified direction z under a weaker constraint qualification than (CQ1)*. In particular we need that the linear program (P)* of Lemma 2.3 have a basic optimal solution satisfying strict complementary slackness for the specified z.

THEOREM 3.3. *Suppose that $U(0,0)$ is nonempty, $U(b,c)$ is uniformly compact near $(0,0)$ and (CQ1)* is satisfied for each $\bar{x} \in P(0,0)$. Then for any direction z*

$$\limsup_{\lambda \to 0^+} \left\{ \frac{f_{\sup}(\lambda z^1, \lambda z^2) - f_{\sup}(0,0)}{\lambda} \right\} \leqq \max_{\bar{x} \in P(0,0)} \max_{(\bar{u},\bar{w}) \in K(\bar{x};0,0)} (\bar{u} \cdot z^1 + \bar{w} \cdot z^2). \tag{3.4}$$

Proof. Let $\{\lambda_n\}$, $\lambda_n \to 0^+$, be any sequence such that

$$\limsup_{\lambda \to 0^+} \frac{f_{\sup}(\lambda z^1, \lambda z^2) - f_{\sup}(0,0)}{\lambda} = \lim_{n \to \infty} \frac{f_{\sup}(\lambda_n z^1, \lambda_n z^2) - f_{\sup}(0,0)}{\lambda_n}. \tag{3.5}$$

By Lemma 2.4, $U(\lambda_n z^1, \lambda_n z^2)$ is nonempty for λ_n small. Let $x_n \in P(\lambda_n z^1, \lambda_n z^2)$. Since $U(\lambda_n z^1, \lambda_n z^2)$ is uniformly compact near $(0,0)$, there is a subsequence, again indexed by n, and a $\bar{x}$ such that $x_n \to \bar{x}$. Since $f_{\sup}$ is continuous (by Theorem 2.6), $\bar{x} \in P(0,0)$. Then

$$\frac{f_{\sup}(\lambda_n z^1, \lambda_n z^2) - f_{\sup}(0,0)}{\lambda_n} = \frac{f(x_n) - f(\bar{x})}{\lambda_n}. \tag{3.6}$$

Consider the linear program

$$\begin{aligned} &\min \ -\{u \cdot z^1 + w \cdot z^2\} \\ &\text{subject to} \\ &\sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) + \sum_{j=1}^{p} w_j \nabla h_j(\bar{x}) = \nabla f(\bar{x}), \\ &u_i g_i(\bar{x}) = 0, \qquad i = 1, \cdots, m, \\ &u \geqq 0, \qquad w \text{ unrestricted}, \end{aligned} \tag{P}$$

and its dual

$$\begin{aligned} &\max \nabla f(\bar{x}) \cdot y \\ &\text{subject to} \\ &\nabla g_i(\bar{x}) \cdot y + g_i(\bar{x}) v_i \leqq -z_i^1, \qquad i = 1, \cdots, m, \\ &\nabla h_j(\bar{x}) \cdot y = -z_j^2, \qquad j = 1, \cdots, p, \\ &y, v \text{ unrestricted}. \end{aligned} \tag{D}$$

From Theorem 2.2, the primal problem (P) is feasible and bounded, thus the dual problem has an optimal solution $\bar{y}$ and

$$\nabla f(\bar{x}) \cdot \bar{y} = \min_{(u,w) \in K(\bar{x};0,0)} -\{u \cdot z^1 + w \cdot z^2\}. \tag{3.7}$$

From Proposition 1.1,

$$\bar{y} = -\nabla H(\bar{x})^{\#} z^2 + [I_n - \nabla H(\bar{x})^{\#} \nabla H(\bar{x})] \bar{y}.$$

For n sufficiently large, $\nabla H(x_n)$ still has full row rank and we can define

$$\bar{y}_n = -\nabla H(x_n)^{\#} z^2 + [I_n - \nabla H(x_n)^{\#} \nabla H(x_n)]\bar{y}.$$

Let $\tilde{y}$ given by (CQ1)*. By Proposition 1.1, there is a $\tilde{z} \in R^n$ such that

$$\tilde{y} = [I_n - \nabla H(\bar{x})^{\#} \nabla H(\bar{x})]\tilde{z}.$$

Define

$$\tilde{y}_n = [I_n - \nabla H(x_n)^{\#} \nabla H(x_n)]\tilde{z},$$

and for $\varepsilon > 0$, let $\hat{y}_n = \varepsilon \tilde{y}_n + \bar{y}_n \to \hat{y} = \varepsilon \tilde{y} + \bar{y}$ where, by definition,

$$\nabla g_i(\bar{x}) \cdot \hat{y} < -z_i^1, \qquad i \in I(\bar{x}; 0). \tag{3.8}$$

Since $\nabla H(x_n)\hat{y}_n = -z^2$, we have, by Proposition 1.1,

$$\hat{y}_n = -\nabla H(x_n)^{\#} z^2 + [I_n - \nabla H(x_n)^{\#} \nabla H(x_n)]\hat{y}_n.$$

As in Lemma 2.4, consider for each n the differential equation

$$\begin{aligned} \psi_n'(\lambda; z) &= -\nabla H(\psi_n(\lambda; z))^{\#} z^2 \\ &\quad + [I_n - \nabla H(\psi_n(\lambda; z))^{\#} \nabla H(\psi_n(\lambda; z))]\hat{y}_n, \\ \psi_n(0; z) &= x_n. \end{aligned} \tag{3.9}$$

Since $\hat{y}_n \to \hat{y}$ and $x_n \to \bar{x}$, Propositions (1.3) and (1.4) apply to the differential equations (3.9). Thus there exists a subsequence of solutions $\{\psi_m(\lambda; z)\}$ converging uniformly to $\psi_0(\lambda; z)$ on $[0, \alpha]$ where $\psi_0(\lambda; z)$ is a solution to (3.9) with x_n replaced by $\bar{x}$ and $\hat{y}_n$ by $\hat{y}$. In the same manner as in Lemma 2.4 we can show that for $\lambda \in [0, \alpha]$

$$\begin{aligned} h_j(\psi_m(\lambda; z)) &= h_j(\psi_m(0; z)) - \lambda z_j^2 \\ &= \lambda_m z_j^2 - \lambda z_j^2. \end{aligned}$$

Hence if m is sufficiently large

$$h_j(\psi_m(\lambda_m; z)) = 0, \qquad j = 1, \cdots, p.$$

Also we can show that for m large

$$g_i(\psi_m(\lambda_m; z)) \leqq 0, \qquad i = 1, \cdots, m.$$

Otherwise there exists an i and a subsequence, still indexed by m, such that

$$g_i(\psi_m(\bar{\lambda}_m; z)) > 0.$$

By uniform convergence, if $i \notin I(\bar{x}; 0)$,

$$\lim g_i(\psi_m(\bar{\lambda}_m; z)) = g_i(\psi_0(0; z)) = g_i(\bar{x}) \geqq 0$$

which contradicts the fact that i is inactive at $\bar{x}$. If $i \in I(\bar{x}; 0)$

$$g_i(\psi_m(\bar{\lambda}_m; z)) - g_i(\psi_m(0; z)) > -\bar{\lambda}_m z_i^1.$$

By the mean value theorem, for some $\tilde{\lambda}_m \in [0, \bar{\lambda}_m]$,

$$\nabla g_i(\psi_m(\tilde{\lambda}_m; z)) \cdot \psi_m'(\bar{\lambda}_m; z) > -z_i^1.$$

If we replace $\psi'_m(\tilde{\lambda}_n; z)$ by its right-hand side in (3.9), we obtain, by uniform convergence,

$$\nabla g_i(\bar{x}) \cdot \hat{y} \geqq -z_i^1$$

which is a contradiction of (3.8).

It now follows that for n sufficiently large,

$$\psi_n(\lambda_n; z) \in U(0, 0)$$

and hence $f(\bar{x}) \geqq f(\psi_n(\lambda_n; z))$. Thus

$$\begin{aligned}\frac{f(x_m)-f(\bar{x})}{\lambda_m} &\leqq -\left\{\frac{f(\psi_m(\lambda_m; z))-f(\psi_m(0; z))}{\lambda_m}\right\} \\ &= -\nabla f(\psi_m(\tilde{\lambda}_m; z)) \cdot \psi_m(\tilde{\lambda}_m; z),\end{aligned}$$

for some $\tilde{\lambda}_m \in [0, \lambda_n]$, by the mean value theorem. By (3.7), (3.9), and uniform convergence, we have

$$\begin{aligned}\lim_{\lambda_n \to 0} \frac{f(x_n)-f(\bar{x})}{\lambda_n} &\leqq -\nabla f(\bar{x}) \cdot \hat{y} \\ &= -\nabla f(\bar{x}) \cdot \varepsilon \tilde{y} - \min_{(u,w) \in K(\bar{x}; 0,0)} -\{u \cdot z^1 + w \cdot z^2\}.\end{aligned}$$

Since the left-hand side is independent of ε we can let $\varepsilon \to 0$ to obtain from (3.5) and (3.6)

$$\limsup_{\lambda \to 0^+} \frac{f_{\sup}(\lambda z^1, \lambda z^2)-f_{\sup}(0, 0)}{\lambda} \leqq \max_{(u,w) \in K(\bar{x}; 0,0)} \{u \cdot z^1 + w \cdot z^2\}$$

for some $\bar{x} \in P(0, 0)$. Since $P(0, 0)$ is compact, the result follows. □

The bounds given by (3.3) and (3.4) are sharp in the following sense: examples exist in which the directional derivatives of $f_{\sup}$ exist at (0, 0) with

$$D_z f_{\sup}(0, 0) = \max_{\bar{x} \in P(0, 0)} \max_{(\bar{u}, \bar{w}) \in K(\bar{x}; 0,0)} \{\bar{u} \cdot z^1 + \bar{w} \cdot z^2\}$$

and

$$D_{\hat{z}} f_{\sup}(0, 0) = \max_{\bar{x} \in P(0,0)} \min_{(\bar{u}, \bar{w}) \in K(\bar{x}; 0,0)} \{\bar{u} \cdot \hat{z}^1 + \bar{w} \cdot \hat{z}^2\}$$

for some $\hat{z} \neq z$. The following example illustrates this phenomenon as well as providing a situation in which neither bound is attained.

Example 3.1.

$$\begin{aligned}&\text{maximize } f(x) = x_2 \\ &\text{subject to} \\ &g_1(x) = x_2 + x_1^2 \leqq 0, \\ &g_2(x) = x_2 - x_1^2 \leqq 0, \\ &g_3(x) = -x_2 - 1 \leqq 0.\end{aligned}$$

The unique maximum point occurs at $\bar{x} = (0, 0)$ where (CQ1)* holds with $K(\bar{x}; 0) = \{u_1, u_2, 0): u_1 + u_2 = 1, u_1, u_2 \geqq 0\}$. For any $z \in R^3$, $S(\lambda z) \neq \varnothing$ and is compact for λ sufficiently small. For $z = (1/\sqrt{2}, 1/\sqrt{2}, 0)$, $P(\lambda z) = \{(0, \lambda/\sqrt{2}\}$, $f_{\sup}(\lambda z) = \lambda/\sqrt{2}$ and

$$D_z f_{\sup}(0) = \frac{1}{\sqrt{2}} = \max_{\bar{u} \in K(\bar{x}; 0)} \bar{u} \cdot z = \frac{1}{\sqrt{2}} = \max_{\substack{u_1 + u_2 = 1 \\ u_1 \geqq 0, u_2 \geqq 0}} \{u_1 + u_2\} \frac{1}{\sqrt{2}}.$$

For $z = (0, 1, 0)$, $P(\lambda z) = \{(0, 0)\}$, $f_{\sup}(\lambda z) = 0$, and

$$D_z f_{\sup}(0) = 0 = \min_{\bar{u} \in K(\bar{x}; 0)} u \cdot z = \min_{\substack{u_1 + u_2 = 1 \\ u_1 \geqq 0, u_2 \geqq 0}} \{u_2\}.$$

Finally, for $z = (1, 0, 0)$, $P(\lambda z) = \{(\sqrt{\lambda/2}, \lambda/2), (-\sqrt{\lambda/2}, \lambda/2)\}$, $f_{\sup}(\lambda z) = \lambda/2$, and $D_z f_{\sup}(0) = \frac{1}{2}$ so that

$$0 = \min_{\bar{u} \in K(\bar{x}; 0)} \{\bar{u} \cdot z\} < D_z f_{\sup}(0) < \max_{\bar{u} \in K(\bar{x}; 0)} \{\bar{u} \cdot z\} = 1.$$

We are now able to deduce the existence of the directional derivatives in a number of special cases.

COROLLARY 3.4. *If the hypotheses of Theorem* 3.3 *are satisfied with* (CQ2)* *replacing* (CQ1)* *then for all directions z, the directional derivative exists and is given by*

$$D_z f_{\sup}(0, 0) = \max_{\bar{x} \in P(0,0)} \{\bar{u} \cdot z^1 + \bar{w} \cdot z^2\}.$$

If in addition $\bar{x}$ is the only point of $P(0, 0)$, then $D_z f_{\sup}(0, 0)$ is the linear function of z given by

$$D_z f_{\sup}(0, 0) = \bar{u} \cdot z^1 + \bar{w} \cdot z^2.$$

Proof. The result follows immediately from Corollary 3.2 and Theorem 3.3 by noting that the constraint qualification (CQ2)* implies that the Kuhn–Tucker vector for $\bar{x}$ is unique. □

COROLLARY 3.5 (Gol'stein [7]). *Suppose the functions $-f$ and $\{g_i\}$, $i = 1, \cdots, m$, are convex and that the functions $\{h_j\}$, $j = 1, \cdots, p$, are affine. If $U(0, 0)$ is nonempty, $U(b, c)$ is uniformly compact near* $(0, 0)$ *and* (CQ1)* *is satisfied for each $\bar{x} \in P(0, 0)$ then $f_{\sup}$ has a directional derivative for the direction z at* $(0, 0)$ *and*

$$D_z f_{\sup}(0, 0) = \max_{\bar{x} \in P(0,0)} \min_{(\bar{u}, \bar{w}) \in K(\bar{x}; 0, 0)} \{\bar{u} \cdot z^1 + \bar{w} \cdot z^2\}. \tag{3.10}$$

Proof. Take the sequence $x_n \to \bar{x}$ as in the beginning of the proof of Theorem 3.3. For any $(\bar{u}, \bar{w}) \in K(\bar{x}; 0, 0)$,

$$f(x_n) - f(\bar{x}) \leqq L(x_n, \bar{u}, \bar{w}) - L(\bar{x}, \bar{u}, \bar{w}) + \lambda_n(\bar{u} \cdot z^1 + \bar{w} \cdot z^2)$$

where

$$L(x, \bar{u}, \bar{w}) = f(x) - \bar{u} \cdot g(x) - \bar{w} \cdot h(x)$$

is the Lagrangian function. From the convexity assumption, $\bar{x}$ is a local maximum of $L(x, \bar{u}, \bar{w})$ in R^n; hence

$$\lim_{\lambda_n \to 0} \frac{f(x_n)-f(x)}{\lambda_n} \leqq \min_{(\bar{u},\bar{w}) \in K(\bar{x};0,0)} \{\bar{u} \cdot z^1 + \bar{w} \cdot z^2\}.$$

From (3.5), (3.6) and Theorem 3.1, the result follows. □

We can also obtain the directional derivative $D_z f_{\sup}(0, 0)$ by placing a restriction on the rate of change of the point to set map $P(b, c)$ at $(0, 0)$.

THEOREM 3.6. *Assume the functions in the program* (1.1) *are twice continuously differentiable. Let* $U(0, 0)$ *be nonempty,* $U(b, c)$ *be uniformly compact near* $(0, 0)$ *and* (CQ1)* *be satisfied for each* $\bar{x} \in P(0, 0)$. *If, for given* z *and each* $\bar{x} \in P(0, 0)$, *every sequence* $x_n \in P(\lambda_n z^1, \lambda_n z^2)$ *with* $x_n \to \bar{x}$ *satisfies* $\|x_n - \bar{x}\|^2/\lambda_n \to 0$ *then* $D_z f_{\sup}(0, 0)$ *exists and is given by* (3.10).

Proof. Choose the sequence λ_n and x_n as in the beginning of the proof of Theorem 3.3. As in Corollary 3.5,

$$f(x_n)-f(x) \leqq L(x_n; \bar{u}, \bar{w})-L(\bar{x}; \bar{u}, \bar{w})+\lambda_n(\bar{u} \cdot z^1+\bar{w} \cdot z^2)$$

for any $(\bar{u}, \bar{w}) \in K(\bar{x}; 0, 0)$. Write $x_n - \bar{x} = \|x_n - \bar{x}\| S_n$, $\|S_n\| = 1$; then

$$\frac{f(x_n)-f(\bar{x})}{\lambda_n} \leqq \frac{\|x_n-\bar{x}\|^2}{2\lambda_n} S_n \cdot \nabla^2 L(\xi_n, \bar{u}, \bar{w}) S_n + \bar{u} \cdot z^1 + \bar{w} \cdot z^2$$

where $\xi_n = t_n x_n + (1-t_n)\bar{x}$, for some $t_n \in [0, 1]$. Therefore

$$\lim \frac{f(x_n)-f(\bar{x})}{\lambda_n} \leqq \min_{(\bar{u},\bar{w}) \in K(\bar{x};0,0)} \{\bar{u} \cdot z^1 + \bar{w} \cdot z^2\}$$

for some $\bar{x} \in P(0, 0)$. Thus (3.5), (3.6) together with Theorem 3.1 give the result. □

Finally, note that (3.10) holding for all z does not imply that $f_{\sup}$ is differentiable at $(0, 0)$. If (3.10) holds for all z and if $D_z f_{\sup}(0, 0) = -D_{-z} f_{\sup}(0, 0)$ then we have

$$\max_{\bar{x} \in P(0,0)} \min_{(\bar{u},\bar{w}) \in K(\bar{x};0,0)} \{\bar{u} \cdot z^1 + \bar{w} \cdot z^2\} = \min_{\bar{x} \in P(0,0)} \max_{(\bar{u},\bar{w}) \in K(x;0,0)} \{\bar{u} \cdot z^1 + \bar{w} \cdot z^2\} \tag{3.11}$$

for all z. Thus if (3.10) holds then (3.11) can be thought of as a necessary condition for the existence of $\nabla f_{\sup}(0, 0)$. Note that if there is a unique $(\bar{u}, \bar{w}) \in K(\bar{x}; 0, 0)$ for each $\bar{x} \in P(0, 0)$ and (3.11) holds then the Kuhn–Tucker vectors are the same for each $\bar{x} \in P(0, 0)$. Also if $P(0, 0)$ is a single vector $\bar{x}$ and (3.11) holds then $K(\bar{x}; 0, 0)$ is a singleton set.

Acknowledgment. The authors would like to express their gratitude to R. T. Rockafellar for his useful suggestions.

REFERENCES

[1] R. L. ARMACOST AND A. V. FIACCO, *Second-order parametric sensitivity analysis in NLP and estimates by penalty function methods*, Tech. Rep. T-324, School of Engineering and Appl. Sci., George Washington University, Washington, D.C., December 1975.

[2] C. BERGE, *Topological Spaces*, Oliver and Boyd, Edinburgh and London, 1963.

[3] G. BIRKHOFF AND G. C. ROTA, *Ordinary Differential Equations*, 2nd ed., John Wiley, New York, 1969.

[4] J. P. EVANS AND F. J. GOULD, *Stability in nonlinear programming*, Operations Res., 18 (1970), pp. 107–118.

[5] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[6] A. M. GEOFFRION, *Duality in nonlinear programming: A simplified application oriented development*, SIAM Rev., 13 (1971), pp. 1–37.

[7] E. G. GOL'STEIN, *Theory of Convex Programming*, Translations of Mathematical Monographs, vol. 36, American Mathematical Society, Providence, R.I., 1972.

[8] H. J. GREENBERG AND W. P. PIERSKALLA, *Extensions of the Evans–Gould stability theorems for mathematical programs*, Operations Res., 20 (1972), pp. 143–153.

[9] W. W. HOGAN, *Point-to-set maps in mathematical programming*, SIAM Rev., 15 (1973), pp. 591–603.

[10] ———, *Directional derivatives for external-value functions with applications to the completely convex case*, Operations Res., 21 (1973), pp. 188–209.

[11] O. V. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[12] S. M. ROBINSON, *Stability theory for systems of inequalities, Part I: Linear systems*, SIAM J. Numer. Anal., 12 (1975), pp. 754–769.

[13] ———, *Stability theory for systems of inequalities, Part II: Differentiable nonlinear system*, Tech. Rep. 1388, Mathematics Research Center, Madison, Wisconsin, 1975.

[14] ———, *A characterization of stability in linear programming*, Tech. Rep. 1542, Mathematics Research Center, Madison, Wisconsin, 1975.

[15] R. T. ROCKAFELLAR, *Duality in nonlinear programming*, Mathematics of the Decision Sciences, Part 1, G. G. Dantzeg and A. F. Veinot, eds., American Mathematical Society, Providence, R.I., 1968, pp. 401–422.

[16] ———, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[17] ———, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.

[18] A. C. WILLIAMS, *Marginal values in linear programming*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 82–99.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 2, February 1977

L_2-INSTABILITY CRITERIA FOR INTERCONNECTED SYSTEMS*

M. VIDYASAGAR†

Abstract. Various conditions are presented for a feedback interconnection of subsystems to be L_2-unstable. These results are of two types: (i) "small gain" type results that involve the formulation of a test matrix, and (ii) "passivity" type results that can be applied directly to the various subsystem and interconnection operators. The application of these results is illustrated through examples.

1. Introduction. Recently there has been a great deal of interest in analyzing the stability of "large-scale" systems by treating them as an interconnection of simpler subsystems. The results obtained using this approach are applicable both to Lyapunov stability [1]–[3] and input-output stability [4]. In this paper, we obtain some results relating to the L_2-instability of interconnected feedback systems. To the best of the author's knowledge, this is the first time that instability results have been presented for large-scale systems using the functional analysis approach.

In § 2, we derive some "small gain" type criteria for instability. These criteria are indirect, in the sense that they require the formulation of a "test" matrix, which is then checked for certain properties. The results of § 2 generalize some work of Takeda and Bergen [7], and can be thought of as the instability counterparts to those in [4]. In § 3, the criteria of § 2 are applied to some specific situations to illustrate how they can be used in practice. In § 4, we present some "direct" or "passivity" type of instability criteria. The adjective "direct" refers to the fact that the criteria do not involve the formulation of a test matrix, but can be applied directly to the various subsystem and interconnection operators. The criteria derived in § 4 are applied to some examples in § 5. Finally, § 6 contains the concluding remarks.

We now introduce the concepts of stability and instability that are used in this paper. Let $L_2 = L_2[0, \infty)$ denote the set of real-valued square integrable functions over the interval $[0, \infty)$. Given a function $f: [0, \infty) \to (-\infty, \infty)$, the function $f_T(\cdot): [0, \infty) \to (-\infty, \infty)$ defined by

$$f_T(t) = \begin{cases} f(t), & t \leqq T, \\ 0, & t > T, \end{cases} \tag{1}$$

is called the *truncation* of $f(\cdot)$ to the interval $[0, T)$. The linear space $L_{2e} = L_{2e}[0, \infty)$ is the set defined by

$$L_{2e} = \{f(\cdot) : f_T(\cdot) \in L_2 \ \forall T < \infty\}. \tag{2}$$

The class of interconnected systems studied in this paper consists of those of

* Received by the editors November 26, 1975, and in revised form May 27, 1976.

† Department of Electrical Engineering, Concordia University, Montreal, Quebec, Canada H3G 1M8. This work was supported by the National Research Council of Canada under Grant A-7790.

the form

$$e_i = u_i - \sum_{j=1}^{m} H_{ij} y_j, \qquad i = 1, \cdots, m, \tag{3a}$$

$$y_i = G_i e_i, \qquad i = 1, \cdots, m, \tag{3b}$$

where u_i is the input to the ith subsystem, y_i is the output and e_i is the "error". We assume that u_i, e_i, y_i belong to the product space $L_{2e}^{n_i}$ for some integer n_i, that $G_i: L_{2e}^{n_i} \to L_{2e}^{n_i}$, and that $H_{ij}: L_{2e}^{n_j} \to L_{2e}^{n_i}$. The system (3) represents an interconnection of m subsystems.

Throughout this paper, we make the following basic assumption:

Corresponding to each m-tuple $(u_1, \cdots, u_m)$ with $u_i \in L_2^{n_i}$, there exist $e_i, y_i \in L_{2e}^{n_i}$, $i = 1, \cdots, m$, such that (3) is satisfied. Conditions on the operators G_i, H_{ij} which insure that the above assumption holds can be found in [8, Chap. III].

In this set-up, we say that the system (3) is L_2-*stable* (or simply *stable*) if, for each m-tuple $(u_1, \cdots, u_m)$ with $u_i \in L_2^{n_i}$, any corresponding $e_i, y_i \in L_{2e}^{n_i}$ satisfying (3) actually belong to $L_2^{n_i}$. Similarly, the system (3) is L_2-*unstable* (or simply *unstable*) if there exists some m-tuple $(u_1, \cdots, u_m)$, with $u_i \in L_2^{n_i}$, such that the corresponding e_i or y_i do not belong to $L_2^{n_i}$ for some i.

Throughout this paper, we consistently deal with a particular type of L_2-unstable subsystem operator. In order to avoid endless repetition, we give a name to this class of operators. This is formalized below.

DEFINITION 1. The operator $G_i: L_{2e}^{n_i} \to L_{2e}^{n_i}$ is said to belong to *Class U* if

(i) G_i is linear.

(ii) The set M_i defined by.

$$M_i = \{x \in L_2^{n_i}: Gx \in L_2^{n_i}\} \tag{4}$$

is a proper subset of $L_2^{n_i}$.

(iii) There is a finite constant γ_{ci} such that

$$\|G_i x\|_i \leqq \gamma_{ci} \|x\|_i \quad \forall x \in M_i \tag{5}$$

where $\|\cdot\|_i$ denotes the norm on $L_2^{n_i}$.

(iv) There is a family of constants $\alpha_i(T)$ such that

$$\|(G_i x)_T\|_i \leqq \alpha_i(T) \|x_T\|_i, \quad \forall T \in [0, \infty), \qquad \forall x \in L_{2e}^{n_i}. \tag{6}$$

Remarks. Conditions (i)–(iv) imply, as shown in [7], that G_i represents an L_2-unstable system, that M_i is a closed subspace of $L_2^{n_i}$, and that the set

$$M_i^{\perp} = \{x \in L_2^{n_i}: \langle x, y \rangle_i = 0 \ \forall y \in M_i\}, \tag{7}$$

where $\langle \cdot, \cdot \rangle_i$ denotes the inner product on $L_2^{n_i}$, contains some nonzero elements. We refer to γ_{ci} as the *conditional gain* of G_i. Note that (5) does not define γ_{ci} uniquely, so that the conditional gain of G_i is not unique, as we define it.

2. Criteria involving test matrices. In this section, we present some sufficient conditions for the system (3) to be unstable. These criteria generalize the results in [7], and can be thought of as the instability counterparts to those in [4]. The

criteria presented in this section all involve the formulation of a "test" matrix, which is then checked for certain properties. We only present the general theorems in this section, and a discussion of the theorems is postponed to § 3, which also contains some applications.

THEOREM 1. *Suppose that, for all i, the operator G_i belongs to class U, and suppose there exist finite constants η_{ij} such that*

$$\|(H_{ij}x)_T\|_i \leqq \eta_{ij}\|x_T\|_j, \quad \forall T<\infty, \quad \forall x\in L_{2e}^{n_j}. \tag{8}$$

Define the $m\times m$ matrix P by

$$p_{ij} = \eta_{ij}\gamma_{cj} \tag{9}$$

(where γ_{cj} is the conditional gain of G_j). Under these conditions, the system (3) *is unstable if*

$$\rho(P)\leqq 1 \tag{10}$$

where $\rho(P)$ denotes the spectral radius of P. Specifically, if $u_i\in M_i^{\perp}/\{0\}$ for $i=1,\cdots,m$, then $y_i\notin L_2^{n_i}$ for some i.

Proof. Suppose by way of contradiction that $u_i\in M_i^{\perp}$, $u_i\neq 0\ \forall i$, and that $y_i\in L_2^{n_i}\ \forall i$, By (8) and (3a), it follows that $e_i\in L_2^{n_i}\ \forall i$. Since e_i, $y_i\in L_2^{n_i}$, this implies, by Definition 1, that $e_i\in M_i\ \forall i$. Let

$$z_i = u_i - e_i = \sum_{j=1}^{m} H_{ij}y_j = \sum_{j=1}^{m} H_{ij}G_je_j; \tag{11}$$

since $u_i\in M_i^{\perp}$ and $e_i\in M_i$, we have

$$\|z_i\|_i^2 = \|u_i\|_i^2 + \|e_i\|_i^2 \quad \forall i. \tag{12}$$

Since $u_i\neq 0$, (12) implies that

$$\|z_i\|_i > \|e_i\|_i. \tag{13}$$

On the other hand, from (11), we get

$$\begin{aligned}\|z_i\|_i &\leqq \sum_{j=1}^{m}\|H_{ij}y_j\|_j \leqq \sum_{j=1}^{m}\eta_{ij}\|y_j\|_j\\ &\leqq \sum_{j=1}^{m}\eta_{ij}\gamma_{cj}\|e_j\|_j = \sum_{j=1}^{m}p_{ij}\|e_j\|_j.\end{aligned} \tag{14}$$

Now, since P has all nonnegative entries, the Perron–Frobenius theorem [9, p. 66] states that (i) $\rho(P)$ is an eigenvalue of P, and (ii) one can find a row eigenvector $v=[v_i\cdots v_m]$ of P, corresponding to the eigenvalue $\rho(P)$, such that $v_i\geqq 0\ \forall i$. The fact that v is a row eigenvector of P corresponding to $\rho(P)$ means that

$$\sum_{i=1}^{m} v_ip_{ij} = \rho(P)v_j, \qquad j=1,\cdots,m. \tag{15}$$

We can now conclude the proof. On the one hand, from (14) we get

$$\begin{aligned}\sum_{i=1}^{m} v_i\|z_i\|_i &\leqq \sum_{i=1}^{m}\sum_{i=1}^{m} v_i p_{ij}\|e_j\|_j \\ &= \rho(P)\sum_{j=1}^{m} v_j\|e_j\|_j.\end{aligned} \tag{16}$$

On the other hand, from (13) we get

$$\sum_{i=1}^{m} v_i\|z_i\|_i > \sum_{i=1}^{m} v_i\|e_i\|_i, \tag{17}$$

where we use the fact that at least one v_i is positive. However, (16) and (17) are in contradiction, since $\rho(P)\leqq 1$. This contradiction shows that $y_i \notin L_2^{n_i}$ for some i. □

COROLLARY 1.1. *Under the conditions of Theorem* 1, *the system* (3) *is unstable if the leading principal minors of* $I-P$ *are all positive.*

Proof. It is shown in [4] that if the leading principal minors of $I-P$ are all positive, then $\rho(P)<1$. □

Theorem 1 states that, if all of the subsystem operators G_i are unstable and all of the interconnection operators H_{ij} are stable, then the overall system is unstable provided the test matrix P has a spectral radius less than or equal to one. In Theorems 2 and 3 below, the restriction that *all* of the subsystem operators are unstable is removed at the expense of some added conditions on the test matrix P.

THEOREM 2. *Suppose* $k<m$, *and that the operators* G_i, $i=1,\cdots,k$, *belong to class U. Suppose there exist finite constants for* γ_i, $i=k+1,\cdots,m$, *such that*

$$\|(G_i x)_T\|_i \leqq \gamma_i\|x_T\|_i, \quad \forall T<\infty, \quad \forall x\in L_{2e}^{n_i}. \tag{18}$$

Suppose there exist finite constants η_{ij} *such that* (8) *holds. Define the* $m\times m$ *matrix* P *by*

$$p_{ij}=\eta_{ij}\gamma_{cj}, \qquad i=1,\cdots,m, \quad j=1,\cdots,k, \tag{19a}$$

$$p_{ij}=\eta_{ij}\gamma_j, \qquad i=1,\cdots,m, \quad j=k+1,\cdots,m, \tag{19b}$$

and partition P as follows:

$$P=\begin{array}{c} \\ m-k \end{array}\overset{\begin{array}{cc} k & m-k \end{array}}{\left[\begin{array}{c|c} P_{uu} & P_{us} \\ \hline P_{su} & P_{ss} \end{array}\right]}. \tag{20}$$

Under these conditions, the system (3) is L_2-unstable if (i) $\rho(P)\leqq 1$, and (ii) at least one column of P_{su} contains all positive elements. Specifically, if $u_i^{\perp}\in M_i/\{0\}$ for $i=1,\cdots,k$, and $u_i=0$ for $i=k+1,\cdots,m$, then $y_i\notin L_2^{n_i}$ for some i in $\{1,\cdots,k\}$.

The proof of Theorem 2 requires the following simple result:

LEMMA 1. *Let* P *be as in Theorem* 2 *and suppose that at least one column of* P_{su} *contains all positive elements. Let* $v=[v_1,\cdots v_m]$ *be any nonnegative row eigenvector of* P *corresponding to the eigenvalue* $\rho(P)$. *Then* $v_i>0$ *for some* i *in* $\{1,\cdots,k\}$.

Proof of Lemma 1. Partition v as follows:

$$v = [\underset{k}{v_a} | \underset{m-k}{v_b}] \tag{21}$$

The fact that v is a row eigenvector corresponding to $\rho(P)$ means that

$$v_a P_{uu} + v_b P_{su} = \rho(P) v_a, \tag{22a}$$

$$v_a P_{us} + v_b P_{ss} = \rho(P) v_b. \tag{22b}$$

Now suppose by way of contradiction that $v_a = 0$. Then (22a) reduces to

$$v_b P_{su} = 0. \tag{23}$$

However, since v_b is nonnegative and at least one column of P_{su} contains all positive elements, (23) implies that $v_b = 0$. Hence $v = 0$, which contradicts the assumption that v is an eigenvector (and hence a nonzero vector). Therefore, some component of v_a is positive. □

Proof of Theorem 2. Suppose $u_i \in M_i^{\perp}/\{0\}$ for $i = 1, \cdots, k$ and $u_i = 0$ for $i = k+1, \cdots, m$, and assume by way of contradiction that $y_i \in L_2^{n_i}$ for $i = 1, \cdots, k$. Then $e_i \in M_i$ for $i = 1, \cdots, k$. Now, as in the proof of Theorem 1, define

$$z_i = u_i - e_i = \sum_{j=1}^{m} H_{ij} G_j e_j. \tag{24}$$

The hypotheses on u_i imply that

$$\|z_i\|_i > \|e_i\|_i \quad \text{for } i = 1, \cdots, k, \tag{25a}$$

$$\|z_{iT}\|_i = \|e_{iT}\|_i \quad \forall T < \infty, \quad \text{for } i = k+1, \cdots, m. \tag{25b}$$

From (25a), we see that there exists a $T < \infty$ such that

$$\|z_{iT}\|_i > \|e_{iT}\|_i \quad \text{for } i = 1, \cdots, k. \tag{26}$$

Let T be so chosen. Next, let $v = [v_1, \cdots v_m]$ be a nonnegative row eigenvector of P corresponding to the eigenvalue $\rho(P)$. Then on the one hand, we have, as in the proof of the Theorem 1, that

$$\begin{aligned} \sum_{i=1}^{m} v_i \|z_{iT}\|_i &\leqq \sum_{i=1}^{m} \sum_{j=1}^{m} v_i p_{ij} \|e_{jT}\|_j \\ &= \rho(P) \sum_{j=1}^{n} v_j \|e_{jT}\|_j, \end{aligned} \tag{27}$$

and on the other hand, from (26) and (25b),

$$\sum_{i=1}^{m} v_i \|z_{iT}\|_i > \sum_{i=1}^{m} v_i \|e_{iT}\|_i \tag{28}$$

In deriving (28), we have used the fact that $v_i > 0$ for some i in $\{1, \cdots, k\}$. Since (27) and (28) contradict each other if $\rho(P) \leqq 1$, it follows that $y_i \notin L_2^{n_i}$ for some i in $\{1, \cdots, k\}$. □

Theorem 2 shows that an interconnection of stable and unstable subsystems is itself unstable, provided (i) the test matrix P has a spectral radius less than or equal to one, and (ii) at least one of the unstable subsystems is connected to every stable subsystem. Theorem 3 below is actually a corollary to Theorem 2, but because of its significance, is stated as a separate theorem.

THEOREM 3. *Under the conditions of Theorem* 2, *the system* (3) *is unstable if* $\rho(P)<1$.

Proof. Suppose $\rho(P)<1$. Since the constants γ_{ci}, $i=1,\cdots,k$, and η_{ij}, $i,j=1,\cdots,m$, are only upper bounds, they can be replaced by larger numbers without affecting the validity of the bounds. In particular, one can replace γ_{c1} by $\gamma_{c1}+\varepsilon$ and η_{i1} by $\eta_{i1}+\varepsilon$ for $i=k+1,\cdots,m$, and choose $\varepsilon>0$ sufficiently small that the spectral radius of the resulting matrix P is still less than 1. Since the first column of the submatrix P_{su} contains all positive elements, the instability follows by Theorem 2. □

COROLLARY 3.1. *Under the conditions of Theorem* 2, *the system* (3) *is unstable if the leading principal minors of the matrix* $I-P$ *are all positive.* □

3. Applications of test matrix criteria. In this section, we apply Theorems 1 and 2 to various specific situations, to illustrate how they can be used in practice.

Application 1. We show that the "small gain" results of Takeda and Bergen [7] can be obtained as special cases of *both* Theorem 1 and Theorem 2. The system studied by Takeda and Bergen is described by the equations

$$e_1=u_1-y_2, \qquad e_2=u_2+y_1, \tag{29}$$

$$y_1=G_1e_1, \qquad y_2=G_2e_2. \tag{30}$$

They show that if (i) G_1 is linear, unstable, and has conditional gain $\gamma_c(G_1)$, and (ii) G_2 is stable, and has gain $\gamma(G_2)$, then the overall system is unstable provided

$$\gamma_c(G_1)\gamma(G_2)\leqq 1. \tag{31}$$

In particular, $y_1\notin L_2$ if $u_2=0$ and $u_1\in M_1^{\perp}/\{0\}$, where M_1 is defined as in (4).

To put the system (29)–(30) into the framework of Theorem 1, we observe that if $u_2=0$, the system under study can be described by

$$e_1=u_1-G_2y_1, \tag{32a}$$

$$y_1=G_1e_1. \tag{32b}$$

This is of the form (3) with $m=1$ (i.e., only one subsystem). Thus, Theorem 1 can be applied, and the test matrix P reduces to the scalar $\gamma_c(G_1)\gamma(G_2)$. Hence, the condition $\rho(P)\leqq 1$ reduces to (31).

To put the system (29)–(30) into the framework of Theorem 2, we let $m=2$ (i.e., two interconnected subsystems) and $k=1$, (i.e., one unstable subsystem). Then the constants γ_{c1} and γ_2 are given by

$$\gamma_{c1}=\gamma_c(G_1), \qquad \gamma_2=\gamma(G_2),$$

while the "interconnection matrix" H is given by

$$H=\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}.$$

Thus the test matrix P becomes

$$P = \begin{bmatrix} 0 & \gamma(G_2) \\ \gamma_c(G_1) & 0 \end{bmatrix}.$$

Now, if $\gamma_c(G_1) > 0$, the condition that one column of the submatrix P_{su} contains all positive elements is satisfied. Moreover, $\rho(P) \leqq 1$ if and only if (31) holds. In this case, the system is L_2-unstable by Theorem 2. On the other hand, if $\gamma_c(G_1) = 0$, then $\rho(P) = 0$, so that the system is L_2-unstable by Theorem 3.

Application 2. To illustrate the possibility of loop transformations in applying Theorems 1–3, we consider two interconnected feedback systems described by the equations

$$e_1(t) = u_1(t) - k_{11}(t)y_1(t) - k_{12}(t)y_2(t), \tag{33a}$$

$$e_2(t) = u_2(t) - k_{21}(t)y_1(t) - k_{22}(t)y_2(t), \tag{33b}$$

$$y_1(t) = (g_1 * e_1)(t), \tag{33c}$$

$$y_2(t) = (g_2 * e_2)(t) \tag{33d}$$

where $*$ denotes convolution. It is assumed that g_1 and g_2 are generalized functions belonging to the Banach algebra $\mathscr{A}$ [8, p. 26], and that $k_{11}(\cdot) - k_{22}(\cdot)$ are regulated functions satisfying the following assumptions, respectively:

$$0 < a_1 \leqq k_{11}(t) \leqq b_1 \quad \forall t, \tag{34a}$$

$$0 < a_2 < k_{22}(t) \leqq b_2 \quad \forall t, \tag{34b}$$

$$|k_{12}(t)| \leqq \beta_{12} \quad \forall t, \tag{34c}$$

$$|k_{21}(t)| \leqq \beta_{21} \quad \forall t. \tag{34d}$$

Let D_1 be the closed disk in the complex plane centered on the negative real axis and passing through $-1/a_1 + j0$, $-1/b_1 + j0$, and let D_2 be similarly defined. In order to set up the subsystems, we assume that

1. the locus $\omega \mapsto \hat{g}_1(j\omega)$ does not intersect D_1, but encircles it finitely many times;
2. the locus $\omega \mapsto \hat{g}_2(j\omega)$ neither intersects nor encircles the disk D_2,

where "$\hat{}$" denotes Laplace transformation.

In order to perform the loop transformations, we restate (33) in the form

$$e_1'(t) = u_1(t) - k_{11}'(t)y_1(t) - k_{12}(t)y_2(t), \tag{35a}$$

$$e_2'(t) = u_2(t) - k_{21}(t)y_1(t) - k_{22}'(t)y_2(t), \tag{35b}$$

$$y_1(t) = (g_1' * e_1')(t), \tag{35c}$$

$$y_2(t) = (g_2' * e_2')(t), \tag{35d}$$

where

(36a) $$k'_{11}(t) = k_{11}(t) - c_1,$$

(36b) $$k'_{22}(t) = k_{22}(t) - c_2,$$

(37a) $$\hat{g}'_1(s) = \hat{g}_1(s)/(1 + c_1\hat{g}_1(s)),$$

(37b) $$\hat{g}'_2(s) = \hat{g}_2(s)/(1 + c_2\hat{g}_2(s)),$$

and

(38a) $$c_1 = (a_1 + b_1)/2,$$

(38b) $$c_2 = (a_2 + b_2)/2.$$

Now, the system (35) is in a form suitable for the application of Theorem 2. Using by now well-known arguments, one can show that, if $k'_{11} = k_{12} = k_{21} = k'_{22} \equiv 0$, then the subsystem corresponding to e_1 and y_1 is unstable, whereas the subsystem corresponding to e_2 and y_2 is stable. Hence $m = 2$ and $k = 1$ in this case. Moreover, the conditional gain γ_{c1} is given by

(39) $$\gamma_{c1} = \sup_{\omega \in R} |\hat{g}'_1(j\omega)|$$

while the gain γ_2 is given by

(40) $$\gamma_2 = \sup_{\omega \in R} |\hat{g}'_2(j\omega)|.$$

Similarly, the bounds on the interaction operators are given by

(41a) $$\eta_{11} = (b_1 - a_1)/2 \triangleq r_1,$$

(41b) $$\eta_{12} = \beta_{12},$$

(41c) $$\eta_{21} = \beta_{21},$$

(41d) $$\eta_{22} = (b_2 - a_2)/2 \triangleq r_2.$$

In order to obtain suitable conditions for instability, it is necessary to introduce the concepts of a "margin of stability" and a "margin of instability". In the present context, we say that the subsystem 1 has a margin of instability δ_1 if

(42) $$\sup_{\omega \in R} |\hat{g}'_1(j\omega)| \cdot (\delta_1 + r_1) = 1$$

and similarly that the subsystem 2 has a margin of stability δ_2 if

(43) $$\sup_{\omega \in R} |\hat{g}'_2(j\omega)| \cdot (\delta_2 + r_2) = 1.$$

Note that δ_1 can be determined graphically by plotting $\hat{g}_1(j\omega)$ (the *untransformed* transfer function), as follows: Plot $\hat{g}_1(j\omega)$, and choose δ_1 such that the disk centered on the negative real axis and passing through $-1/(b_1 + \delta_1) + j0$ and $-1/(a_1 - \delta_1) + j0$ just touches the plot of $\hat{g}(j\omega)$. Also, since the conditions to be given below are anyway just sufficient conditions, one can always replace δ_1 by a suitable lower bound, if it is more readily obtainable. It is clear that similar considerations apply to δ_2 as well.

We are now in a position to state the stability criteria. From Theorem 2, the test matrix P is obtained as

$$P = \begin{bmatrix} \eta_{11}\gamma_{c1} & \eta_{12}\gamma_2 \\ \eta_{21}\gamma_{c1} & \eta_{22}\gamma_2 \end{bmatrix}. \tag{44}$$

Now, from (42) and (43), it is clear that

$$\eta_{11}\gamma_{c1} = r_1/(\delta_1 + r_1) < 1, \tag{45a}$$

$$\eta_{22}\gamma_2 = r_2/(\delta_2 + r_2) < 1. \tag{45b}$$

Also, we can safely assume that $\gamma_{c1} > 0$, since the problem is trivial otherwise. Returning to the test matrix P, if $\eta_{21} = 0$, i.e., if the unstable subsystem is not connected to the stable subsystem, then $\rho(P)$ is clearly less than one, so that overall instability follows by Theorem 3. On the other hand, if $\eta_{21} > 0$, then $\eta_{21}\gamma_{c1} > 0$, so that Theorem 2 is applicable. In this case, a sufficient condition for the overall system to be unstable is that $\rho(P) \geqq 1$, which reduces (after routine algebra) to

$$(1 - \eta_{11}\gamma_{C1}) \cdot (1 - \eta_{22}\gamma_2) - \eta_{12}\gamma_2 \cdot \eta_{21}\gamma_{C1} \geqq 0, \tag{46}$$

or in other words,

$$\beta_{12} \cdot \beta_{21} \leqq \delta_1 \cdot \delta_2. \tag{47}$$

The simple interpretation of the condition (47) is that overall instability results if the product of the interaction gains is less than or equal to the product of the margin of instability and the margin of stability.

In closing, it should be noticed that, if one assumes that the plot of $\hat{g}_1(j\omega)$ neither encircles nor intersects the disk D_1, then (47) with strict inequality is a sufficient condition for stability. Thus Theorems 1–3, together with earlier results [4], [5], form a complete package for testing the stability as well as instability of interconnected feedback systems.

4. Direct instability criteria. In this section, we derive some "direct" criteria for the instability of interconnected feedback systems. These criteria differ from those in § 2 in that they can be applied directly to the various subsystem and interconnection operators, and do not require the formulation of an auxiliary test matrix. The criteria presented in this section are of the "passivity" type, and can be thought of as the instability counterparts to the results given in [10]. They also generalize some results of Takeda and Bergen [7].

Throughout this section, we concentrate on an interconnection of some stable and some unstable subsystems. However, in order to avoid notational clutter, we lump all the stable subsystems into one big subsystem, and all the unstable subsystems into another. Thus the system under study is described by

$$e_1 = u_1 - H_{11}y_1 - H_{12}y_2, \tag{48a}$$

$$e_2 = u_2 - H_{21}y_1 - H_{22}y_2, \tag{48b}$$

$$y_1 = G_1 e_1, \tag{48c}$$

$$y_2 = G_2 e_2, \tag{48d}$$

where e_1, u_1, $y_1 \in L_{2e}^{n_1}$, e_2, u_2, $y_2 \in L_{2e}^{n_2}$, $G_i: L_{2e}^{n_i} \to L_{2e}^{n_i}$, and $H_{ij}: L_{2e}^{n_i} \to L_{2e}^{n_i}$. It is further assumed that the operator H_{ij} is specified by an $n_i \times n_j$ matrix of *constants*. Clearly there is no loss of generality in assuming the system description (48), because the operators G_i can be "block-diagonal"; moreover, the interconnection operators can always be chosen to be constants, by increasing the number of subsystems if necessary. We assume the following:

(A1) The operator G_1 belongs to class U (see Definition 1).

(A2) There exists a positive constant δ, such that

$$\langle e_1, G_1 e_1 \rangle \geqq \delta_1 \|e_1\|^2, \quad \forall e_1 \in M_1, \tag{49}$$

where $\langle \cdot, \cdot \rangle_T$ denotes the inner product, and

$$M_1 = \{e_1 \in L_2^{n_1} : G_1 e_1 \in L_2^{n_1}\} \tag{50}$$

and M_1 is a proper subspace of $L_2^{n_1}$.

(A3) The operator G_2 is passive; i.e.,

$$\langle e_2, G_2 e_2 \rangle_T \geqq 0 \quad \forall T, \quad \forall e_2 \in L_{2e}^{n_2}, \tag{51}$$

where $\langle \cdot, \cdot \rangle_T$ denotes the truncated inner product.

(A4) The $(n_1+n_2) \times (n_1+n_2)$ matrix H defined by

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \tag{52}$$

is skew-symmetric.

(A5) The matrix H has an inverse, which is denoted by P. The matrix P is "partitioned" as follows:

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}, \tag{53}$$

where P_{ij} is of the order $n_i \times n_j$.

(A6) The operator $G_2 + P_{22}: L_{2e}^{n_2} \to L_{2e}^{n_2}$ satisfies

$$(G_2 + P_{22}) e_2 = 0 \Rightarrow e_2 = 0. \tag{54}$$

The assumptions (A1)–(A6) are much less restrictive than they appear to be at a first glance. This is brought out in § 5, where some applications are studied.

The main result of this section is given next:

THEOREM 4. *Under assumptions* (A1)–(A6), *the system* (48) *is unstable. In particular, whenever* u_1 *and* u_2 *are of the form* $u_1 = H_{11} v_1$, $u_2 = H_{21} v_1$, *for some* $v_1 \in M_1^{\perp}/\{0\}$, *we have that either* $y_1 \notin L_2^{n_1}$ *or* $y_2 \notin L_2^{n_2}$.

Proof. Let $u_1 = H_{11} v_1$, $u_2 = H_{21} v_1$, for some $v_1 \in M_1^{\perp}/\{0\}$. Since $P = H^{-1}$, it is easily shown that

$$P_{11} u_1 + P_{12} u_2 = v_1 \in M_1^{\perp}/\{0\}, \tag{55}$$

$$P_{21} u_1 + P_{22} u_2 = 0. \tag{56}$$

Now, suppose by way of contradiction that $y_1 \in L_2^{n_1}$ and $y_2 \in L_2^{n_2}$. Then, clearly, $e_1 \in L_2^{n_1}$, $e_2 \in L_2^{n_2}$, from (48a) and (48b). Now, $e_1 \in L_2^{n_1}$, $y_1 \in L_2^{n_1}$ implies that

$e_1 \in M_1$. By routine manipulation, we get

$$\begin{aligned}\langle y_1, u_1\rangle + \langle y_2, u_2\rangle &= \langle y_1, e_1\rangle + \langle y_2, e_2\rangle \\ &\quad - [\langle y_1, H_{11}y_1 + H_{12}y_2\rangle + \langle y_2, H_{21}y_1 + H_{22}y_2\rangle] \\ &= \langle y_1, e_1\rangle + \langle y_2, e_2\rangle \\ &\geqq \delta_1 \|e_1\|^2,\end{aligned} \tag{57}$$

where we have used (A4), (A2) and (A3), in succession. Also, one can "solve" (48a) and (48b) for y_1 and Py_2, and obtain

$$y_1 = P_{11}(u_1 - e_1) + P_{12}(u_2 - e_2), \tag{58}$$

$$y_2 = P_{21}(u_1 - e_1) + P_{22}(u_2 - e_2), \tag{59}$$

Hence, from (58), (59), we get

$$\begin{aligned}\langle y_1, e_1\rangle + \langle y_2, e_2\rangle &= \langle P_{11}(u_1 - e_1) + P_{12}(u_2 - e_2), e_1\rangle \\ &\quad + \langle P_{21}(u_1 - e_1) + P_{22}(u_2 - e_2), e_2\rangle \\ &= -[\langle P_{11}e_1 + P_{12}e_2, e_1\rangle + \langle P_{21}e_1 + P_{22}e_2, e_2\rangle] \\ &\quad + \langle P_{11}u_1 + P_{12}u_2, e_1\rangle + \langle P_{21}u_1 + P_{22}u_2, e_2\rangle \\ &= 0,\end{aligned} \tag{60}$$

where we have used the skew-symmetry of P, and (55), (56), respectively. Combining (57) and (60), we see that $e_1 = 0$, whence $y_1 = 0$. Therefore (48a) and (48b) reduce to

$$0 = u_1 - H_{12}y_2, \tag{61a}$$

$$e_2 = u_2 - H_{22}y_2. \tag{61b}$$

Multiplying (61a) by P_{21}, (61b) by P_{22}, and adding gives

$$\begin{aligned}P_{22}e_2 &= P_{21}u_1 + P_{22}u_2 - (P_{21}H_{12} + P_{22}H_{22})y_2 \\ &= -y_2 = -G_2e_2,\end{aligned} \tag{62}$$

where we have used the facts that (i) $P_{21}u_1 + P_{22}u_2 = 0$, (ii) $P_{21}H_{12} + P_{22}H_{22}$ equals the identity operator on $L_2^{n_2}$, and (iii) (48d). Hence, (62) implies that

$$(G_2 + P_{22})e_2 = 0. \tag{63}$$

However, by (A6), this implies that $e_2 = 0$. By linearity of P_{22}, it follows that $P_{22}e_2 = 0$, whence from (63), we have $y_2 = G_2e_2 = 0$. Hence (61) implies that $u_1 = 0$, $u_2 = 0$. However, this contradicts the hypothesis that $P_{11}u_1 + P_{12}u_2 \neq 0$. Hence either $y_1 \notin L_2^{n_1}$ or $y_2 \notin L_2^{n_2}$. □

COROLLARY 4.1. *Let* (A7) *and* (A8) *be defined as follows*:

(A7) *The operator G_2 is strongly passive; i.e., there exists a positive constant δ_2 such that*

$$\langle e_2, G_2e_2\rangle_T \geqq \delta_2 \|e_2\|_T^2 \quad \forall T, \quad \forall e_2 \in L_{2e}^{n_2}. \tag{64}$$

(A8) *The operator G_2 is unbiased*; *i.e.*,

$$e_2 = 0 \Rightarrow G_2 e_2 = 0. \tag{65}$$

Then the conclusions of Theorem 4 *hold with* (A3) *and* (A6) *replaced by* (A7) *and* (A8).

Proof. With (A7) in place, (57) is modified to

$$\langle y_1, u_1\rangle + \langle y_2, u_2\rangle \geqq \delta_1 \|e_1\|^2 + \delta_2 \|e_2\|^2, \tag{66}$$

and hence from (60) and (66), it follows that $e_1 = 0$, $e_2 = 0$; since G_1 is linear, $e_1 = 0$ implies $y_1 = 0$, while the unbiasedness of G_2 implies that $y_2 = 0$. The desired contradiction is obtained immediately, showing that either $y_1 \notin L_2^{n_1}$ or $y_2 \notin L_2^{n_2}$. □

As Theorem 4 is stated, one can only conclude that, corresponding to a certain class of inputs, *either* $y_1 \notin L_2^{n_1}$ *or* $y_2 \in L_2^{n_2}$. By adding two extra assumptions, it is possible to draw a stronger conclusion, namely that, for a certain class of inputs, $y_1 \notin L_2^{n_1}$. This is demonstrated next:

COROLLARY 4.2. *Let* (A9) *and* (A10) *be defined as follows*:

(A9) *The operator $I + H_{22}G_2$: $L_{2e}^{n_2} \times L_{2e}^{n_2}$ satisfies the condition*

$$(I + H_{22}G_2)e_2 \in L_2^{n_2} \Rightarrow e_2 \in L_2^{n_2}. \tag{67}$$

(A10) *G_2 maps $L_2^{n_2}$ into itself.*

With these definitions, if (A1)–(A6) *and* (A9)–(A10) *hold, then $y_1 \notin L_2^{n_1}$ whenever u_1 and u_2 are of the form $u_1 = H_{11}v_1$, $u_2 = H_{21}v_1$ for some $v_1 = M_1^{\perp}/\{0\}$. The same conclusion is valid if* (A1)–(A4) *and* (A7)–(A10) *hold.*

Proof. Suppose u_1 and u_2 are of the indicated form, and assume by way of contradiction that $y_1 \in L_2^{n_1}$. Then $u_2 - H_{21}y_1 \in L_2^{n_2}$. Thus (48b) assumes the form

$$(I + H_{22}G_2)e_2 = u_2 - H_{21}y_1; \tag{68}$$

since the right side of (67) belongs to $L_2^{n_2}$, it follows by (A9) that $e_2 \in L_2^{n_2}$, and then by (A10) that $y_2 \in L_2^{n_2}$. Thus, from (48a), we have that $e_1 \in L_2^{n_1}$. From this point onwards, the proof is as in Theorem 4. □

5. Application of direct criteria. In this section, we illustrate the applications of the criteria obtained in § 4. Specifically, we discuss the implications of assumptions (A1)–(A10), and we show how the earlier results of Takeda and Bergen [7] can be obtained as special cases of Theorem 4. We also discuss the considerations involved in introducing "multipliers", and close out with an illustrative example.

Discussion of (A1)–(A10). (A1) states that the first subsystem is unstable in the particular manner specified, while (A2) requires the unstable subsystem to be "conditionally strongly passive." If G_1 is represented by a convolution integral, these two conditions are easy to check, as detailed in [7], [8]. (A3) requires the subsystem G_2, *which need not be stable or linear*, to be passive. (A4) requires the interactions to be "nondissipative", while (A5) requires the interaction matrix to be invertible. This last condition has an interesting implication: namely, bounded-input–bounded-output stability is equivalent to bounded-input–bounded-error

stability if H is invertible. On the other hand, if H is singular, then bounded-input–bounded-output stability implies bounded-input–bounded-error stability, but the reverse implication may not be true. (A6) and (A9) pertain to the behavior of the second subsystem, in the case where only the "self-interaction" terms H_{11} and H_{22} are present—(A6) states, roughly, that a zero output can only be due to a zero input, while (A9) states that, with only self-interaction terms, the second subsystem is stable. Note that (A9) is trivially satisfied if $H_{22} = 0$. (A7), (A8) and (A10) are self-explanatory. Note that (A6) implies (A10) if $P_{22} = 0$.

Application 3.[1] Consider a system described by (29) and (30), which is the one studied by Takeda and Bergen [7]. As demonstrated in § 3, this system can be put in the framework of (48) by defining

$$H = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Hence (A4) and (A5) are satisfied. Also, since

$$P = H^{-1} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

(A6) reduces to the requirement that $G_2 e_2 = 0 \Rightarrow e_2 = 0$. Now, if (A1)–(A3) hold, then Theorem 4 states that either $y_1 \notin L_2$ or $y_2 \notin L_2$ whenever $u_2 \in M_1/\{0\}$ and $u_1 = 0$. Moreover, Corollary 4.2 is applicable to this situation, because (i) (A9) is satisfied trivially because $H_{22} = 0$, and (ii) since $P_{22} = 0$, (A6) implies (A10). Hence, by Corollary 4.2, we can conclude that whenever $u_1 = 0$ and $u_2 \in M_1^{\perp}/\{0\}$, it is in fact y_1 that does not belong to L_2. This is the same as the result obtained in [7].

More generally, if we let $n_1 = n_2$ (so that $L_2^{n_1}$ and $L_2^{n_2}$ are the same space), and let H be of the form

$$H = \begin{bmatrix} 0 & A \\ -A^* & 0 \end{bmatrix}$$

(where A^* denotes the adjoint of A) such that A^*A and AA^* are both invertible, then we obtain the results of Sundereshan [11].

However, in comparing Theorem 4 to earlier results in [7] and [11], it is important to recognize two facts: 1. There is no previous analogue to Theorem 4. In fact, earlier results are really special cases of Corollary 4.2, where stronger conclusions are drawn than in Theorem 4, but at the expense of added conditions. Thus, Theorem 4 is an original contribution, whereby weakened conditions for instability are given. 2. Theorem 4 removed the very stringent requirement, laid down in [7] and [11], to the effect that *exactly* half of the subsystems are unstable. Here, stable and unstable subsystems can occur in any combination. However, a mathematically essential assumption (whose physical significance is not too clear), is that the *total* number of subsystems is even, because otherwise H would be a skew-symmetric matrix of odd order which is always singular.

Application 4. In any stability or instability criteria derived using passivity, the scope of application of the criteria is vastly enlarged by the introduction of

[1] Recall that Applications 1 and 2 are contained in § 3.

multipliers. It is possible to introduce multipliers into Theorem 4, but some care must be taken, as illustrated in the sequel.

Given the original system description (48), suppose we want to introduce a multiplier Z. Then first of all, we must have $n_1 = n_2$ (i.e. $L_2^{n_1}$ and $L_2^{n_2}$ are the same space) in order for Z^{-1} to exist. Hence, we are immediately restricted to having exactly half unstable and half stable subsystems. (This assumption is *not* made in Theorem 4—indeed, it is possible to have stable and unstable subsystems in any mixture therein.) Suppose $n_1 = n_2$ and that $Z: L_2^{n_1} \to L_2^{n_1}$ has the property that Z^{-1} exists. We introduce the multiplier Z into the system (48) by defining

$$y_1' = Zy_1, \qquad e_2' = Ze_2.$$

Then the system description (48) is modified to

$$e_1 = u_1 - H_{11}Z^{-1}y_1' - H_{12}y_2, \tag{69a}$$

$$e_2' = Ze_2 - ZH_{21}Z^{-1}y_1' - ZH_{22}y_2, \tag{69b}$$

$$y_1' = ZG_1e_1, \tag{69c}$$

$$y_2 = G_2Z^{-1}e_2'. \tag{69d}$$

Thus G_1 and G_2 have been replaced by ZG_1 and G_2Z^{-1}, respectively, as desired. However, in the process the operator H has been modified to

$$H' = \begin{bmatrix} H_{11}Z^{-1} & H_{12} \\ ZH_{21}Z^{-1} & ZH_{22} \end{bmatrix}.$$

The main difficulty with introducing multipliers is that, even though the original operator H satisfies (A4), the new operator H' need not do so. This difficulty is not encountered in the "single-loop" feedback stability theory, because if

$$H = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

then $H' = H$, regardless of what Z is. However, in the more general case under study here, one can use a multiplier Z only if the new "interconnection operator" H' satisfies (A4) and (A5).

Application 5. In general, it is not always easy to rearrange the system in such a way that the interconnection operator H satisfies (A3) and (A4). To illustrate this, consider the system shown in Fig. 1. If we number the various subsystems as shown, the interconnection matrix becomes

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

which is not even skew-symmetric. With a little thought, one can rearrange the given system as shown in Fig. 2, in which case H is skew-symmetric. However,

since there are five subsystems, H is singular. In order to make H nonsingular as well as skew-symmetric, one is obliged to "pad" the system by adding a sixth subsystem, as in Fig. 3. The operator F_5' is so chosen that $F_5'(I+F_6F_5')^{-1}=F_5$; i.e., $F_5'=F_5(I-F_6F_5)^{-1}$. It can be shown that, if F_5 is strongly passive, then F_5' is also strongly passive provided ε is chosen sufficiently small. In this case, the operators G and H are given by

$$G=\begin{bmatrix} F_1 & & & & & \\ & F_2-F_5 & & & 0 & \\ & & F_3 & & & \\ & & & F_4-F_5 & & \\ & 0 & & & F_5' & \\ & & & & & F_6 \end{bmatrix},$$

$$H=\begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}.$$

Depending on which are the stable subsystems and which are the unstable subsystems in Fig. 3, one can get a variety of instability criteria. Let us assume, for the sake of definiteness, that F_1 is an unstable operator belonging to the class U and satisfies (A1)–(A2). Under these conditions, the system depicted in Fig. 3 is unstable provided (i) the operators F_2-F_5, F_5, F_4-F_5 are all passive, and (ii) (A6) holds. The full expansion of this latter condition involves only routine algebra, and is therefore omitted in the interests of brevity.

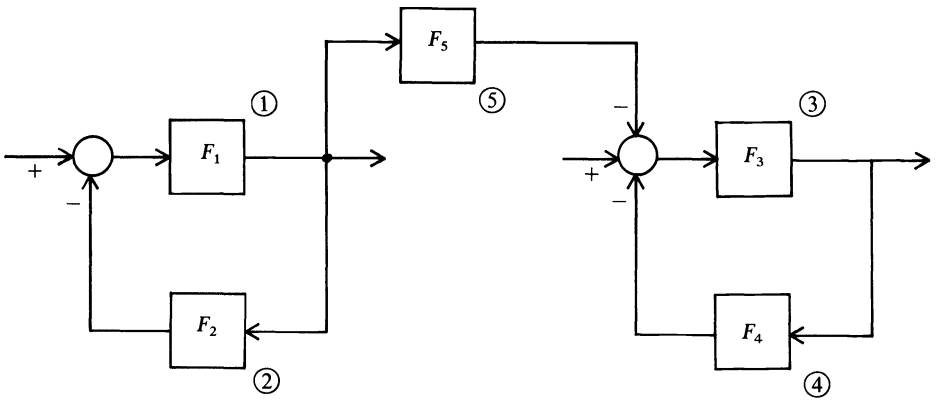


FIG. 1

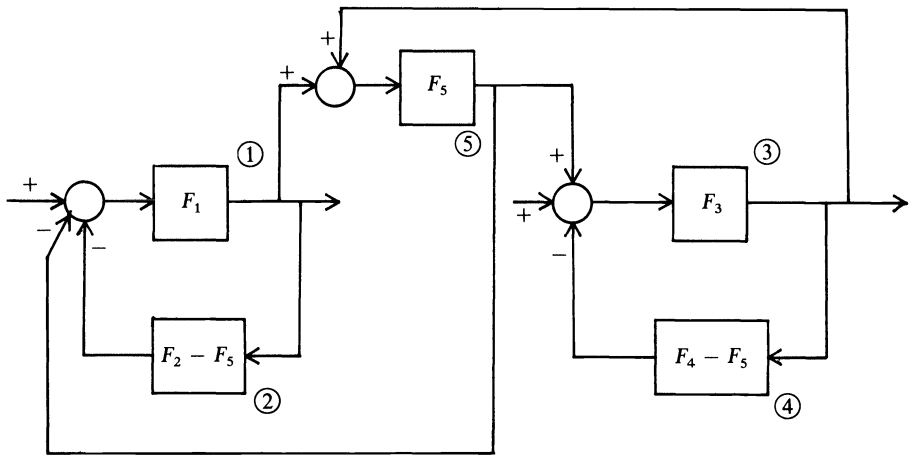


FIG. 2

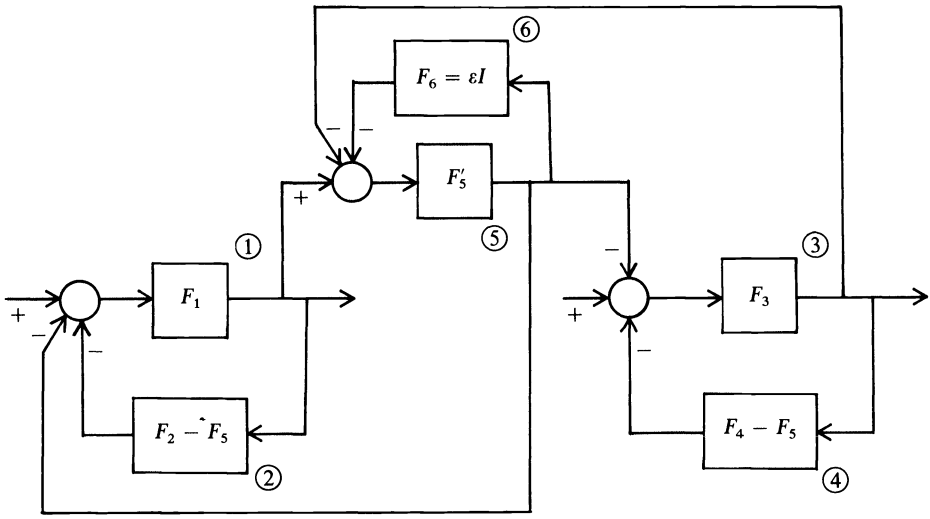


FIG. 3

6. Conclusions. In this paper, several results have been presented pertaining to the L_2-instability of interconnected feedback systems. These results fall naturally into two categories: (i) criteria that involve the construction of an auxiliary test matrix, and (ii) criteria that can be applied directly to the various subsystem and interconnection operators. It so happens that the test matrix criteria are of a "small gain" type, while the direct criteria are of a "passivity" type. However, this need not always be so. Recent research indicates that "passivity type" criteria involving test matrices can also be derived. These results will be reported elsewhere.

All of the theorems presented here generalize, in a very natural way, the corresponding results for the "single-loop case". But, more importantly, by

utilizing the results derived here, it is possible to study the stability status of interconnected systems containing some stable and some unstable subsystems. In an earlier work [3], instability results are derived that can be applied only to the very limited case of all subsystems being strongly unstable, whereas in [11], one is obliged to assume that exactly half of the subsystems are unstable while the other half are stable. No such unnatural assumptions are made in this paper. Furthermore, the results given here are natural instability counterparts to recent results in the stability of interconnected systems [4], [10].

An as yet unsolved problem, which assumes a great deal of importance in light of the results of this paper (as well as related work [10]), is the following: Given an interconnected system of the form (3), when is it possible to rearrange the system and redefine the subsystem operators in such a way that the interconnection matrix is skew-symmetric?

Acknowledgment. The author gratefully acknowledges several enlightening discussions with Dr. M. K. Sundareshan.

REFERENCES

[1] F. N. BAILEY, *The application of Lyapunov's second method to interconnected systems*, this Journal, 3 (1966), pp. 443–462.

[2] D. D. SILJAK, *Stability of large-scale systems under structural perturbations*, IEEE Trans. Systems, Man, and Cybernetics, SMC-2 (1972), pp. 657–663.

[3] L. T. GRUJIC AND D. D. SILJAK, *Asymptotic stability and instability of large-scale systems*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 643–645.

[4] D. W. PORTER AND A. N. MICHEL, *Input-output stability of time-varying non-linear multiloop feedback systems*, Ibid., AC-19 (1974), pp. 422–427.

[5] E. LASLEY AND A. N. MICHEL, *Input-output stability of interconnected systems*, Ibid., AC-21 (1976), pp. 84–89.

[6] B. D. O. ANDERSON, *The small gain theorem, the passivity theorem, and their equivalence*, J. Franklin Inst., 293 (1972), pp. 105–115.

[7] S. TAKEDA AND A. R. BERGEN, *Instability of feedback systems by orthogonal decomposition of* L_2, IEEE Trans. Automatic Control, AC-18 (1973), pp. 631–636.

[8] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[9] F. R. GANTMACHER, *Matrix Theory*, vol. II, Chelsea, New York, 1959.

[10] M. K. SUNDARESHAN AND M. VIDYASAGAR, L_2*-stability of large-scale dynamical systems—criteria via positive operator theory*, IEEE Trans. Automatic Control, to appear.

[11] M. K. SUNDARESHAN, *Large-scale dynamical systems—*L_2*-instability criteria*, submitted for publication.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 2, February

AN EXTENSION OF DUALITY-STABILITY RELATIONS TO NONCONVEX OPTIMIZATION PROBLEMS*

E. J. BALDER†

Abstract. By an effective extension of the conjugate function concept a general framework for duality-stability relations in nonconvex optimization problems can be studied. The results obtained show strong correspondences with the duality theory for convex minimization problems. In specializations to mathematical programming problems the canonical Lagrangian of the model appears as the extended Lagrangian considered in exterior penalty function methods.

1. Introduction. Given an initial nonconvex optimization problem we consider it embedded in a family of perturbed optimization problems (the initial problem corresponding to the zero perturbation). It is well-known that the main results on duality in convex minimization problems are intimately connected with stability properties (such as continuity and lower semi-continuity) of the canonical perturbation function of the family of perturbed problems at the zero perturbation [1], [13], [16], [18], [19], [28], [29], [30]. One can distinguish between asymmetric and symmetric duality descriptions. The former makes sense, for instance, in optimization problems in mathematical programming [1] and allows then some slightly more general statements than the latter. In turn the symmetric approach allows the introduction of a family of perturbed dual problems whose corresponding dual problem ("the dual's dual") is the original problem, e.g., [13], [18], [30]. Such symmetry is obtained in the description of convex minimization problems by assuming ab initio lower semi-continuity of the convex functions that are minimized.

We shall proceed now as follows. After the introduction of some notions of our framework we shall be in a position to give a heuristic motivation for the extension of the conjugate concept and related results which appear in a formal fashion in § 2. In § 3 the results obtained are then applied in our framework to obtain asymmetric duality-stability relations for nonconvex optimization problems. Such relations were already obtained by Rockafellar [31] in a particular case. The notion of Lagrangian is introduced in the framework and we show that in specializations to mathematical programming problems this Lagrangian appears as the well-known extended Lagrangians one encounters in exterior penalty function methods. The interpretation of such Lagrangians via the conjugate function concept with its appealing geometric intuition may be helpful to practitioners. The so-called exact multiplier methods, for instance, will have a simple geometric interpretation in terms of a stability property—an extended form of subdifferentiability—of the perturbation function of the problem.

The very general approach taken in our treatment seems fully justified in view of existing interest in problems with infinitely many constraints, such as in

* Received by the editors June 16, 1975, and in revised form May 21, 1976.

† Department of Statistics, University of California, Berkeley, California. Now at Mathematical Institute, University of Utrecht, De Uithof, Utrecht, the Netherlands. This research was supported by the Commonwealth Fund of New York (Harkness Fellowships) and by the National Science Foundation under Grant GP 3109-X2.

optimal control (e.g., [17]). We also give results, again in a very general setting, on the properties which the multipliers in the unconstrained optimization method must satisfy in order to ensure that the (almost-) optimizers of the extended Lagrangian yield an optimal solution of the original problem. Finally, we shall discuss some conditions on the original problem that guarantee certain forms of subdifferentiability of the perturbation function.

We shall now introduce a part of the framework which we shall use for our duality-stability descriptions. It is an adaptation (also notationally) to the nonconvex case of the model in the interesting paper [18] by Joly and Laurent which continued the impressive work by Rockafellar (e.g., [28], [29]).

In what follows $\mathbb{R}$ will denote the real numbers, $\mathbb{R}_+$ the nonnegative real numbers, $\bar{\mathbb{R}}$ the extended real numbers with the convention $(+\infty)+(-\infty)=(-\infty)+(+\infty)=+\infty$. The m-dimensional Euclidean space will be denoted by $\mathbb{R}^m$, its nonnegative (nonpositive) part by $\mathbb{R}^m_+(\mathbb{R}^m_-)$. Whenever this is relevant $\mathbb{R}^m$ will be supposed to have the topology induced by the Euclidean metric.

Let X be a nonempty set, let f be an extended real-valued functional on X, we write then $f \in \bar{\mathbb{R}}^X$. To avoid trivialities we will assume $f \not\equiv +\infty$, i.e., f is not identically equal to $+\infty$ on X.

Consider the *initial minimization problem*:

$$\text{(P)} \qquad \inf_{x \in X} f(x),$$

and call its value α.

It is useful to observe that one can always extend f to some suitable universe $\tilde{X}$ by setting f equal to $+\infty$ on $\tilde{X} \setminus X$.

Let U be a set in which we fix a certain element and denote it by σ. Let the functional $\phi \in \bar{\mathbb{R}}^{X \times U}$ be such that $\phi(x, \sigma) = f(x)$ for all $x \in X$.

(P) can be thought of as embedded in the family of *perturbed minimization problems*

$$(\text{P}_u) \qquad \inf_{x \in X} \phi(x, u), \qquad u \in U.$$

For $u \in U$ denote the value of (P_u) by $h(u)$; $h(\sigma) = \alpha$.

We now give a brief sketch of duality-stability relations in convex minimization problems and motivate heuristically what is to be formalized for nonconvex optimization problems in subsequent sections.

In convex minimization problems X and U are taken to be vector spaces; the fixed element in U is taken to be the (algebraic) zero element of U. A vector space V is introduced, the dual space, supposed to be in bilinear correspondence with U (i.e., such that there exists a bilinear form $\langle \cdot, \cdot \rangle$ on $U \times V$; in this outline we omit the topological aspects of the matter). In case U is the m-dimensional Euclidean space $\mathbb{R}^m$ (mathematical programming) one takes V to be $\mathbb{R}^m$; the usual inner product provides the bilinear form. The dual functional $h_* \in \bar{\mathbb{R}}^V$ is defined as $h_*(v) = \inf_{u \in U}(h(u) - \langle u, v \rangle)$, $v \in V$. Obviously, for $v \in V$, $h_*(v) = \sup\{\eta \mid \eta \in \mathbb{R}, \eta + \langle \cdot, v \rangle \leqq h\}$. The dual optimization problem (Q_*), defined by $\sup_{v \in V} h_*(v)$, consists therefore of a consideration of the supremum of all those

η's for which there exists $v \in V$ such that $\eta + \langle \cdot, v \rangle \leqq h$. If the original convex minimization problem (P) has been embedded appropriately (i.e., by a suitable choice of U and ϕ) the perturbation function h is convex. From any picture one forms of the situation it will be obvious that there is ample reason to investigate the possibility of having $\sup_{v \in V} h_*(v) = h(\sigma) = \inf_{x \in X} f(x)$, and that the behavior of h at σ will play some role.

In the nonconvex case the perturbation function h will be nonconvex in general. The picture of the convex case leads us to look for a set V and a nonbilinear form $c \in \mathbb{R}^{U \times V}$ such that the graphs of the "affine" functionals $\eta + c(\cdot, v)$, $\eta \in \mathbb{R}$, $v \in V$—hypersurfaces—lying below the graph of h have a capacity to become pointed near σ. This would then create a possibility of having the supremum of all those η's for which there exists $v \in V$ such that $\eta + c(\cdot, v) \leqq h$ equal to $h(\sigma)$, provided that h behaves reasonably at σ. In other words, we can then expect duality-stability relations to hold.

2. Conjugate functions. Let V be a set, $c \in \mathbb{R}^{U \times V}$ a functional which we shall refer to as the *coupling functional* of U and V [8], [11], [24]. The coupling functional will play a role analogous to the one the bilinear duality on a couple of paired topological vector spaces plays in convex minimization problems [19].

Following a suggestion made by Moreau [23], [24], who observed that the bilinear form which appears in the original conjugate function apparatus can be replaced by a nonbilinear form without invalidating many of the essential tautologies, we can use c to describe the extended conjugate function apparatus [11], [23], [24], [32], [33].

By an *elementary c-functional* on U we shall indicate a functional of the form $c(\cdot, v) + \eta$ for some $v \in V$, $\eta \in \bar{\mathbb{R}}$; in case $\eta \in \mathbb{R}$ in the previous form the corresponding functional will be called a *finite elementary c-functional.*

Let us denote by $\Gamma^c(U)$ the set of all functionals $a \in \bar{\mathbb{R}}^U$ that are the supremum of a family of elementary c-functionals on U:

$$a = \sup_{i \in I} (c(\cdot, v_i) + \eta_i),$$

for some nonempty index set I, $\{v_i\} \subset V$, $\{\eta_i\} \subset \bar{\mathbb{R}}$. Mutatis mutandis we define the elementary c-functionals on V and the set $\Gamma^c(V)$, reversing the roles of U and V.

The *c-conjugate functional* $a^c \in \Gamma^c(V)$ of a functional $a \in \bar{\mathbb{R}}^U$ is defined by

$$a^c(v) = \sup_{u \in U} (c(u, v) - a(u)), \qquad v \in V,$$

and mutatis mutandis we define the c-conjugate of a functional on V.

The *second c-conjugate functional* $a^{cc} \in \Gamma^c(U)$ of $a \in \bar{\mathbb{R}}^U$ is defined by repetition: $a^{cc} = (a^c)^c$.

A trivial consequence of the definition is that for $a \in \bar{\mathbb{R}}^U$ for all $u \in U, v \in V$

$$a(u) + a^c(v) \geqq c(u, v) \qquad \text{(Young's inequality)},$$

hence always for $a \in \bar{\mathbb{R}}^U$

$$a \geqq a^{cc}. \tag{2.1}$$

The following statement follows directly from the definition [11, Satz 3.6]:

$$a = a^{cc} \quad \text{iff} \quad a \in \Gamma^c(U),$$

or, what is to say the same, the greatest minorant of a which belongs to $\Gamma^c(U)$ is a^{cc}.

Let $\varepsilon \geqq 0$. The functional $a \in \bar{\mathbb{R}}^U$ is said to be *ε-c-subdifferentiable at $u_0 \in U$* if $a(u_0)$ is finite and if there exists $v_0 \in V$ such that for all $u \in U$

$$a(u) - a(u_0) \geqq c(u, v_0) - c(u_0, v_0) - \varepsilon$$

(in other words, if there exists a finite elementary c-functional which is a minorant of a whose value at u_0 differs ε from $a(u_0)$). Such a $v_0 \in V$ is called an *ε-c-subgradient of a at u_0*.

The (possibly empty) set of all ε-c-subgradients of a at u_0 is called the *ε-c-subdifferential of a at u_0*, denoted by $c - \partial_\varepsilon a(u_0)$.

In case $\varepsilon = 0$ the reference to the prefix 0 is omitted entirely. Thus we have *c-subdifferentiable*, *c-subgradient*, *c-subdifferential*, $c - \partial a(u_0)$, etc.

Combining the above it is simple to observe that for $a \in \bar{\mathbb{R}}^U$, $\varepsilon \geqq 0$, $u_0 \in U$, $v_0 \in V$

$$v_0 \in c - \partial_\varepsilon a(u_0) \quad \text{iff} \quad c(u_0, v_0) \leqq a(u_0) + a^c(v_0) \leqq c(u_0, v_0) + \varepsilon,$$

hence

$$c - \partial_\varepsilon a(u_0) \neq \varnothing \quad \text{implies} \quad a^{cc}(u_0) \leqq a(u_0) \leqq a^{cc}(u_0) + \varepsilon, \tag{2.2}$$

also, because a and a^{cc} have the same elementary c-functionals as minorants

$$a(u_0) = a^{cc}(u_0) \quad \text{implies} \quad c - \partial a(u_0) = c - \partial a^{cc}(u_0). \tag{2.3}$$

Moreover, for $a \in \Gamma^c(U)$, $u_0 \in U$, $v_0 \in V$

$$\begin{aligned} v_0 \in c - \partial a(u_0) \quad &\text{iff} \quad a(u_0) + a^c(v_0) = c(u_0, v_0) \\ &\text{iff} \quad u_0 \in c - \partial a^c(v_0). \end{aligned} \tag{2.4}$$

The class $\Gamma^c(U)$ is easy to identify in the case of convex minimization problems, where U and V are topological vector spaces having locally convex topologies, compatible with the bilinear duality c. In that case the class $\Gamma^c(U)$, with the exclusion of its two elements which are identically equal to $+\infty$ and $-\infty$ respectively, is exactly the set of all lower semi-continuous, convex, proper functionals on U [19]. This is an immediate consequence of the Hahn–Banach theorem.

As we are about to see, another situation can occur in which, for a suitable coupling functional c, the class $\Gamma^c(U)$ is easy to identify. Of course this is very important since the class $\Gamma^c(U)$ is extremely intractable by its definition. In fact, we shall see that once such an identification has been made several duality-stability relations follow immediately.

Let U be equipped with a topology. The coupling functional $c \in \bar{\mathbb{R}}^{U \times V}$ is said to be of *needle type at $u_0 \in U$* if for every neighborhood N of u_0 and every $v \in V$, $\eta \in \mathbb{R}$ there exist a $v' \in V$ and a neighborhood N' of u_0, $N' \subset N$, such that for all $u \notin N'$

$$c(u, v') - c(u_0, v') \leqq c(u, v) + \eta,$$

and such that for all $u \in N'$

$$c(u, v') - c(u_0, v') \leqq 0.$$

The coupling functional c is said to be of *needle type on* U if c is of needle type at all points of U.

The following lemma will be of help to identify coupling functionals of needle type in case U is metrizable.

LEMMA 1. *Suppose the topology on U is generated by a metric d. Let $u_0 \in U$, let s be a monotonically increasing function $s: \mathbb{R}_+ \to \mathbb{R}_+$ with $s(0) = 0$ and a constant K such that for all $x \in \mathbb{R}_+$, $s(2x) \leqq Ks(x)$. Suppose the coupling functional $c \in \mathbb{R}^{U \times V}$ has the following property: for every $v \in V$ there exist $u_1 \in U$, $\eta_1 \in \mathbb{R}$, $\rho_1 \in \mathbb{R}$ such that*

$$c(\cdot, v) \geqq \rho_1 s(d(\cdot, u_1)) + \eta_1.$$

Suppose also that the collection of finite elementary functionals includes all functionals $-\rho s(d(\cdot, u_0))$, $\rho \geqq 0$ (or functionals that minorize these and that are equal to zero at u_0). Then c is of needle type at u_0.

Proof. Let N be a neighborhood of u_0. Let $\delta > 0$ be such that the open ball N' around u_0 with radius δ is contained in N. For given $v \in V$, $\eta \in \mathbb{R}$ there exist $u_1 \in U$, $\eta_1, \rho_1 \in \mathbb{R}$—we may suppose without loss of generality that $\rho_1 > 0$—such that

$$c(u, v) + \eta \geqq -\rho_1 s(d(u, u_1)) + \eta_1$$

for all $u \in U$. We claim that for ρ large enough $-\rho s(d(u, u_0))$ is smaller than the right-hand side of the inequality above for all $u \notin N'$. Indeed, for $u \in U$ with $d(u, u_0) \geqq \delta$ and $d(u, u_0) \leqq d(u_0, u_1)$ we have by the triangle inequality and the monotonicity of s that our claim holds. And for $u \in U$ with $d(u, u_0) \geqq \delta$ and $d(u, u_0) > d(u_0, u_1)$ we have by the triangle inequality and the properties of s that

$$\rho s(d(u, u_0)) - \rho_1 s(d(u, u_1)) \geqq (\rho - \rho_1 K) s(\delta),$$

so our claim also holds for those $u \notin N'$. Now all that remains to be checked is the second inequality in the definition of the needle type of a coupling functional at u_0, and this is trivial to verify.

We observe that monomials on $\mathbb{R}_+$ satisfy the conditions imposed upon s in the lemma.

Example 1a. $U = \mathbb{R}^m$, $V = \mathbb{R}^m_+$. Define c by

$$c(u, v) = -\sum_{i=1}^{m} v_i |u_i|, \qquad u \in \mathbb{R}^m, \quad v \in \mathbb{R}^m_+.$$

Then c is of needle type at the origin.

Example 2a. $U = H$, a Hilbert space, $V = \mathbb{R}_+ \times H$,

$$c(u, v) = -v_0 |u|^2 - \langle w, u \rangle, \quad v = (v_0, w) \in \mathbb{R}_+ \times H, \qquad u \in H,$$

where we use standard notation for the norm and inner product on H. Then c is of needle type on U.

Example 3a. $U = L$, a normed vector space, $V = \mathbb{R}_+$, $\gamma > 0$,

$$c(u, v) = -v \|u\|^\gamma, \qquad v \in \mathbb{R}_+, \quad u \in L,$$

where we denote the norm on L by $\|\cdot\|$. Then c is of needle type at the zero element of L.

A functional on U which is minorized by a finite elementary c-functional will be called *c-tempered*. For $a \in \bar{\mathbb{R}}^U$ we have

$$\begin{aligned} &a \text{ is } c\text{-tempered} \quad \text{iff} \quad a^{cc} \not\equiv -\infty \\ &\qquad\qquad\quad \text{iff} \quad \text{for all } u \in U, \quad a^{cc}(u) \neq -\infty. \end{aligned} \tag{2.5}$$

The *l.s.c. (lower semi-continuous) hull* of a functional $a \in \bar{\mathbb{R}}^U$ is defined by

$$\bar{a}(u) = \liminf_{u' \to u} a(u').$$

THEOREM 1. *If $c \in \mathbb{R}^{U \times V}$ is of needle type at $u_0 \in U$, then for every c-tempered functional $a \in \bar{\mathbb{R}}^U$ which is l.s.c. at u_0 we have*

$$a(u_0) = a^{cc}(u_0).$$

Moreover, if for all $v \in V$, $c(\cdot, v)$ is l.s.c. at u_0, then for every c-tempered $a \in \bar{\mathbb{R}}^U$

$$\bar{a}(u_0) = a^{cc}(u_0).$$

Proof. First suppose $a(u_0)$ is finite; let $\varepsilon > 0$ be arbitrary. Since a is supposed to be l.s.c. at u_0 there exists a neighborhood N of u_0 on which a takes values $\geqq a(u_0) - \varepsilon$. Also, we know that a is c-tempered, i.e., that for certain $v \in V$, $\eta \in \mathbb{R}$, $a \geqq c(\cdot, v) + \eta$. Since c is of needle type at u_0 there exists $v' \in V$ and a neighborhood N' of u_0, $N' \subset N$, such that for all $u \notin N'$, $c(u, v') - c(u_0, v') \leqq c(u, v) + \eta - a(u_0) + \varepsilon$ and such that for all $u \in N'$, $c(u, v') \leqq c(u_0, v')$. Now the elementary c-functional $c(\cdot, v') - c(u_0, v') + a(u_0) - \varepsilon$ minorizes a, hence $a^{cc}(u_0) \geqq a(u_0) - \varepsilon$. It follows that $a(u_0) = a^{cc}(u_0)$. In case $a(u_0)$ is not finite, i.e., equal to $+\infty$, the argument is repeated by taking increasingly large numbers instead of $a(u_0) - \varepsilon$.

In case $c(\cdot, v)$ is l.s.c. at u_0 for all $v \in V$, then for $a \in \bar{\mathbb{R}}^U$, c-tempered, its l.s.c. hull $\bar{a}$ is also c-tempered. By the previous part of the theorem $\bar{a}(u_0) = a^{cc}(u_0)$.

THEOREM 2. *If $c \in \mathbb{R}^{U \times V}$ is of needle type on U, then every l.s.c. c-tempered functional belongs to $\Gamma^c(U)$.*

Moreover, if for all $v \in V$, $c(\cdot, v)$ is l.s.c. on U, then for every $a \in \bar{\mathbb{R}}^U$, $a \not\equiv -\infty$,

$$a \in \Gamma^c(U) \quad \textit{iff} \quad a \textit{ is l.s.c. and } c\textit{-tempered.}$$

3. Duality-stability relations. The results of the previous section will be applied in the framework for nonconvex optimization problems.

Suppose that the coupling functional of U and V, introduced in § 2, is normalized in the sense that for all $v \in V$, $c(\sigma, v) = 0$ (where σ denotes the fixed element of U).

Define $g = h^c \in \Gamma^c(V)$. We define then the dual minimization problem corresponding to the family $\{(P_u) | u \in U\}$ to be

$$\inf_{v \in V} g(v), \tag{Q}$$

and call its value β.

Remarks. Usually the dual of a minimization problem is formulated as a maximization problem (such as in the heuristics of § 1), but of course the present formulation, following [18] and [19], constitutes no real difference.

It seems worthwhile to observe that (Q) may well be a convex minimization problem. This is the case, for instance, when V is a vector space and $c(u, \cdot)$ is convex on V for all $u \in U$ [11, Satz 4.3]. The coupling functionals in the examples of § 2 all satisfy this requirement.

It is clear that we have, by (1.1),

$$-\alpha = -h(\sigma) \leqq -h^{cc}(\sigma) = \inf_{v \in V} g(v) = \beta. \tag{3.1}$$

Due to our assumption $f \not\equiv +\infty$, we have in addition

$$\alpha < +\infty, \qquad \beta > -\infty. \tag{3.2}$$

Because of this and (2.5) we can distinguish between two cases.

Case 1. $\beta = +\infty$. In this case $g \equiv +\infty$, h is not c-tempered and a duality gap exists when $\alpha \neq -\infty$.

Case 2. β finite. Now h is c-tempered and because of Theorem 1 we have immediately

THEOREM 3. *If $\beta < +\infty$ and c is of needle type at σ, then*

$$h \text{ is l.s.c. at } \sigma \quad \text{implies} \quad -\alpha = \beta.$$

Moreover, if also for all $v \in V$, $c(\cdot, v)$ is l.s.c. at σ, then

$$-\beta = \bar{h}(\sigma).$$

Conditions in terms of the original problem that ensure lower semicontinuity of the perturbation function are well-known for convex minimization problems [18], [19], [30], also for certain types of nonconvex problems [7], [19]. Since most of our interest will lie in optimization problems of the kind one encounters in mathematical programming, it seems appropriate to single out results for this class here. We shall say that a minimization problem $\inf_{x \in X} f(x)$ is of *mathematical programming type* (or is a *mathematical program*) if its objective function has the following structure. There exist a functional $f_0 \in \mathbb{R}^X$, a map $G: X \to Z$, where Z is a topological vector space ordered by some closed convex cone C in Z ($z' \leqq z$ iff $z' - z \in C$, $z, z' \in Z$) such that $f(x) = f_0(x)$ for $x \in X$ if $G(x) \leqq 0$ (0 denotes the algebraic zero of Z) and $f(x) = +\infty$ for $x \in X$ if $G(x) \not\leqq 0$. In this setup it is customary to take $U = Z$, $\sigma = 0$ and to consider the perturbed problems

$$(\mathrm{P}_u) \qquad \inf_{x \in X} (f_0(x) | G(x) \leqq u), \qquad u \in U.$$

That is, we define $\phi(x, u) = f_0(x)$ if $G(x) \leqq u$, $\phi(x, u) = +\infty$ otherwise, $x \in X$, $u \in U$.

For such a mathematical program the following condition ensures lower semi-continuity of the perturbation function at σ: f_0 l.s.c. on X, G order closed

(i.e., $u_\nu \to u_0$, $x_\nu \to x_0$, $u_\nu \geqq G(x_\nu)$ implies $u_0 \geqq G(x_0)$ for any net $\{u_\nu\}$, u_0 in U and any net $\{x_\nu\}$, x_0 in X) and there exist $\bar{u} \in \operatorname{int} C$, $\bar{\alpha} > \alpha$, such that $\{f_0 \leqq \bar{\alpha}\} \cap \{G \leqq \bar{u}\}$ is compact (if int C is empty it suffices to require the existence of $\bar{\alpha} > \alpha$ such that $\{f_0 \leqq \bar{\alpha}\}$ is compact). The finite dimensional transcription is well-known.

Since $g \in \Gamma^c(V)$ we have by (2.4) that if $\beta < +\infty$ the set B of solutions of (Q), $\{\bar{v} \in V | g(\bar{v}) = \beta\}$ is characterized by

$$B = c - \partial h^{cc}(\sigma).$$

Thus, by (2.2) and (2.3), c-subdifferentiability of h at σ is equivalent to the existence of a finite dual solution and the absence of a duality gap; that is,

$$c - \partial h(\sigma) \neq \varnothing \quad \text{iff} \quad B \neq \varnothing \quad \text{and} \quad -\alpha = \beta < +\infty. \tag{3.3}$$

We shall denote the set of primal solutions, $\{\bar{x} \in X | f(\bar{x}) = \alpha\}$, by A.

The asymmetry, mentioned in the introduction, in our approach is further underlined by a study of the *Lagrangian* ℓ of (P)—or rather $\{(P_u) | u \in U\}$—which is defined by

$$\ell(x, v) = \sup_{u \in U} (c(u, v) - \phi(x, u)), \qquad x \in X, \quad v \in V.$$

The Lagrangian of the mathematical program is of the form

$$\ell(x, v) = -f_0(x) + \bar{c}(G(x), v), \qquad x \in X, \quad v \in V,$$

where we define

$$\bar{c}(u, v) = \sup (c(u', v) | u' \in U, u' \geqq u), \qquad u \in U, \quad v \in V.$$

Of course, for $v \in V$, $\bar{c}(\cdot, v)$ is monotonically nonincreasing. In specializations ℓ appears as the extended Lagrangian one encounters in exterior penalty function methods [12], [31], [34]:

Example 1b. Mathematical program; c, U, V as in Example 1a. Let $G = (g_1, \cdots, g_p, g_{p+1}, \cdots, g_m)$, where $g_i : X \to \mathbb{R}$, $i = 1, \cdots, m$, $1 \leqq p \leqq m$, *let* $C = \mathbb{R}^p_- \times (0, \cdots, 0)$ be the negative cone in U; then

$$\bar{c}(G(x), v) = -\sum_{i=1}^{p} v_i \max(g_i(x), 0) - \sum_{i=p+1}^{m} v_i |g_i(x)|, \qquad x \in X, \quad v \in V.$$

Example 2b. Mathematical program; c, U, V as in Example 2a with $H = \mathbb{R}^m$. Let G, C be as in Example 1b; then

$$\bar{c}(G(x), v) = -\sum_{i=1}^{p} [v_0 \max^2 (g_i(x), -w_i/2v_0) + w_i \max (g_i(x), -w_i/2v_0)]$$

$$- \sum_{i=p+1}^{m} [v_0 (g_i(x))^2 + w_i g_i(x)], \qquad x \in X, \quad v = (v_0, w) \in V.$$

Example 3b. Mathematical program; c, U, V as in Example 3a. Let $G : X \to U$, $C = \{\sigma\}$; then

$$\bar{c}(G(x), v) = -v \|G(x)\|^\gamma, \qquad x \in X, \quad v \in V.$$

As in [19], one finds

$$\beta = \inf_{v \in V} \sup_{x \in X} \ell(x, v), \tag{3.4}$$

since in fact $\sup_{x \in X} \ell(x, v) = g(v)$, $v \in V$. On the other hand, however, $-f(x) \leqq \inf_{v \in V} \ell(x, v)$, $x \in X$, and in general there is no equality, as one can verify by a simple counterexample. We shall observe soon that in optimization problems of mathematical programming type "symmetry" for the Lagrangian (i.e., equality instead of the inequality above) is inherently present in case c is of needle type at σ. A similar remark can be made for convex minimization problems.

THEOREM 4. *If for $\bar{x} \in X$ and $\bar{v} \in V$ for all $x \in X$, $v \in V$*

$$\ell(x, \bar{v}) \leqq \ell(\bar{x}, \bar{v}) \leqq \ell(\bar{x}, v),$$

then $\bar{v} \in B$. In case $\beta < +\infty$, c is of needle type at σ and $\phi(\bar{x}, \cdot)$ is l.s.c. at σ we have in addition that $\bar{x} \in A$. If, moreover, for all $v \in V$, $c(\cdot, v)$ is l.s.c. at σ and $-\alpha = \beta$, then

$$\bar{x} \in A \quad \textit{iff} \quad \phi(\bar{x}, \cdot) \textit{ is l.s.c. at } \sigma.$$

Proof. From the above and (3.4) it will be clear that $g(\bar{v}) = \sup_{x \in X} \ell(x, \bar{v}) = \beta = \inf_{v \in V} \ell(\bar{x}, v) \leqq -\alpha$. In case $\beta < +\infty$ and c is of needle type at σ we have, by Theorem 1, that $\phi(\bar{x}, \cdot)$'s lower semi-continuity at σ implies $-f(\bar{x}) = -\phi_{\bar{x}}^{cc}(\sigma) = \inf_{v \in V} \ell(\bar{x}, v)$, where we denote $\phi(\bar{x}, \cdot)$ by $\phi_{\bar{x}}$. By the same theorem the last statement is proved.

Remark. It is easy to observe that for an optimization problem of mathematical programming type, as introduced above, one always has that for all $x \in X$, $\phi(x, \cdot)$ is l.s.c. at σ.

The following theorem is a consequence of the definition of ℓ and (3.3).

THEOREM 5. *If h is c-subdifferentiable at σ, then for $\bar{x} \in A$, $\bar{v} \in B$ we have that for all $x \in X$, $v \in V$*

$$\ell(x, \bar{v}) \leqq \ell(\bar{x}, \bar{v}) \leqq \ell(\bar{x}, v);$$

in particular for $\bar{v} \in c - \partial h(\sigma)$

$$\sup_{x \in X} \ell(x, \bar{v}) = -\alpha.$$

The last statement of Theorem 5 conveys some of the essence of what is known under the term exact multipliers for the extended Lagrangian. In a single (if exact) maximization the value of the original problem is attained and for the optimization problem of mathematical programming type we can then speak about a single unconstrained optimization. This notion has received some attention in recent years [5], [15], [25], [34], due to the well-known disadvantages of sequential minimization where a dual sequence $\{v_k\}$ is generated such that $\lim_{k \to \infty} \sup_{x \in X} \ell(x, v_k) = \lim_{k \to \infty} g(v_k) = -\alpha$ (see Theorem 3); this implies that the v_k are almost-subdifferentials of h. The essential problem here is that if the v_k generated show an extremal behavior (i.e., one or more of the components—real numbers—of the v_k tend to infinity for growing k) this invariably causes the Hessian matrices of the extended Lagrangians—and cases where these do not exist are computationally even worse—to become ill-conditioned. And this of

course badly affects the search at each step for an (almost-) maximizer x_k of $\ell(\cdot, v_k)$.

Another problem is, naturally, that any cluster point of $\{x_k\}$ need not necessarily be an optimal solution of the original problem, and usually one is more interested in optimal solutions than in the optimal value of a problem. For an exact multiplier we also wish to obtain an optimal solution via single exact optimization.

We shall address ourselves to the questions raised. In this matter we shall need some structure which is present if we suppose the optimization problem (P) to be of the very general mathematical programming type defined above. This will be assumed henceforth. We also need an additional structural property of the coupling functional.

The coupling functional $c \in \mathbb{R}^{U \times V}$ is defined to be *flexible at* $\sigma \in U$ if for every neighborhood N of σ and every v', $v'' \in V$, $\eta \in \mathbb{R}$ there exist a $v \in V$ and a neighborhood N' of σ, $N' \subset N$, such that for all $u \notin N'$

$$c(u, v) \leqq c(u, v') + \eta, \tag{3.5}$$

and such that for all $u \in N'$

$$c(u, v) \leqq c(u, v''). \tag{3.6}$$

Obviously if c is a coupling functional which is flexible at σ and for which for some $v \in V$ $c(\cdot, v) \leqq 0$, then c is of needle type at σ.

The coupling functionals of Examples 1, 2, 3 are all flexible at σ. In the vein of Lemma 1 more of such functionals can be determined. Suppose that a sequential procedure produces in its course a sequence of dual parameters $\{v_k\}$. Any sensible sequential procedure will adjust its parameters in such a way that $g(v_k) \to \beta$ (supposed finite) and be based on the presumption that $\beta = -\alpha$ (see Theorem 3). Such adjustment can take place either by jacking up some component(s) of the parameter indiscriminately [12], [26], with its ensuing numerical difficulties, or by a more sophisticated method, such as Hestenes' multiplier method [4], [27] in which careful shifting of the "parabolic needle" (Example 2) of a certain sharpness, which is held fixed during each phase of shifting, takes place before the sharpness is increased. For the goal should be—with as little jacking up of the parameters as possible—the construction of almost-subdifferentials of h, as is evident by defining $\varepsilon_k = g(v_k) + \alpha$, $k \in \mathbb{N}$, which shows that $v_k \in c - \partial_{\varepsilon k} h(\sigma)$, $k \in \mathbb{N}$, with $\varepsilon_k \to 0$.

Suppose further that each optimization step of $\ell(\cdot, v_k)$, $k \in \mathbb{N}$, yields an almost-optimizer x_k within a certain precision λ_k. We are interested in conditions guaranteeing that any cluster point of $\{x_k\}$ will be an optimal solution of the original problem.

THEOREM 6. *Suppose the topology on U is generated by a norm. Suppose that the coupling functional $c \in \mathbb{R}^{U \times V}$ is flexible at σ. Let $\{N_k\}$ be a sequence of open balls around σ whose diameters decrease monotonically to zero. Suppose that in order to solve the original problem a sequence $\{v'_k\}$ in V is generated, $v'_k \in c - \partial_{\varepsilon_k} h(\sigma)$, $k \in \mathbb{N}$, where $\{\varepsilon_k\}$ is a sequence of nonnegative numbers converging to zero. Let $\{\lambda_k\}$ be a sequence of nonnegative numbers. Let $\{\eta_k\}$ be a sequence of real numbers such that*

$\eta_k < -\varepsilon_k - \lambda_k$, $k \in \mathbb{N}$. *For* $k \in \mathbb{N}$ *let* v_k *be an element in* V *satisfying* (3.5), (3.6) *with* $N = N_k$, $\eta = \eta_k$, $v' = v'' = v'_k$. *Let* $x_k \in X$ *be such that* $\ell(x_k, v_k) \geqq \sup_x \ell(x, v_k) - \lambda_k$, $k \in \mathbb{N}$. *Then if* G *is order closed we have that every cluster point of* $\{x_k\}$ *is feasible.*

If in addition f_0 *is l.s.c.,* $\{\eta_k\}$ *and* $\{\lambda_k\}$ *converge to zero, and the family* $\{c(\cdot, v_k)\}$ *is equi-upper-semicontinuous at* σ, *then every cluster point of* $\{x_k\}$ *is a solution of the original program.*

Proof. Observe that we have implicitly assumed that α is finite and that β equals $-\alpha$. The values $g(v_k)$, $k \in \mathbb{N}$, are finite, therefore also the values $-f_0(x_k) + \bar{c}(G(x_k), v_k)$, $k \in \mathbb{N}$. Let $u_k \in U$ satisfy $c(u_k, v_k) \geqq \bar{c}(G(x_k), v_k) - \delta_k$, $u_k \geqq G(x_k)$, where $\delta_k = \frac{1}{2}(-\eta_k - \varepsilon_k - \lambda_k)$, $k \in \mathbb{N}$. Now, for $k \in \mathbb{N}$,

$$-\alpha - \lambda_k \leqq -f_0(x_k) + c(u_k, v_k) + \delta_k \leqq -h(u_k) + c(u_k, v_k) + \delta_k. \tag{3.7}$$

We must have $u_k \in N_k$, $k \in \mathbb{N}$, since assumption of the contrary leads to $-\alpha - \lambda_k \leqq g(v'_k) + \delta_k + \eta_k$, hence to a contradiction. In other words $u_k \to \sigma$ and by order closedness of G the proof of the first part is completed.

If the additional assumptions are satisfied we obviously have

$$\limsup_k c(u_k, v_k) \leqq 0.$$

Thus if x^* is a cluster point of $\{x_k\}$ we find by (3.7) that $-\alpha \leqq -f_0(x^*)$ and the first part of the theorem guarantees feasibility of x^*.

Of course equi-upper-semicontinuity is present trivially if, for all $k \in \mathbb{N}$, $c(\cdot, v_k) \leqq 0$, but also the following criterion is immediate if, for all $k \in \mathbb{N}$, $c(\cdot, v_k)$ happens to be concave (Examples 1, 2, 3).

LEMMA 2. *Let* $\{a_i\}_{i \in I}$ *be a family of concave functionals on the vector space* U *equipped with a norm* $\|\cdot\|$. *Let* $a'_i(\sigma; u)$ *denote the Gateaux derivative of* a_i *at* σ *in the direction* u, $u \in U$, $i \in I$. *Then* $\{a_i\}$ *is equi-upper-semicontinuous at* σ *if there exists a constant* K *such that for every* $i \in I$, $u \in U$, $\|u\| = 1$, $a'_i(\sigma; u) \leqq K$.

Some comment on the meaning of Theorem 6 ought to be given. Certainly the theorem does not specify how to adjust the dual parameters $\{v'_k\}$, nor does it explain how to find from these the parameters $\{v_k\}$ whose existence is postulated by the flexibility assumption. Rather do we intend to provide practitioners with a general structure which, it is hoped, will facilitate the analysis of concrete methods and will provide geometric intuition. Thus it should be obvious, for instance, that the choice of λ_k's (η_k's) and N_k's will be a compromise between computational accuracy in the maximization steps, the amount of increase in some component(s) of the parameters and the rate at which the points x_k approach the feasible region.

In the case of Example 2 (finite dimensions) we can refer to [31, Thm. 3] for a result similar to Theorem 6 (via Lemma 2) with a certain elegant choice of λ_k's, ε_k's and N_k's. The counterexample following that result shows that the equi-upper-semicontinuity condition for the v_k's produced is indispensible.

Consider now the case of an exact multiplier with exact optimization of the Lagrangian (e.g., [5], [15], [25], [34]). Suppose that somehow—we shall discuss this aspect of the matter below—we have been able to determine an exact multiplier $\bar{v} \in c - \partial h(\sigma)$. (Again implicit in this statement is the assumption $\alpha = -\beta$, a finite number). Suppose $\tilde{x} \in X$ maximizes the corresponding Lagrangian exactly. The question arises whether $\tilde{x}$ is an optimal solution of the original problem.

THEOREM 7. *Let $c \in \mathbb{R}^{U \times V}$ be flexible at σ. Suppose $\bar{v} \in c - \partial h(\sigma)$, and suppose that the $\bar{v}$ corresponding to* (3.5), (3.6) *for some neighborhood N of σ, some $\eta < 0$, $v' = v'' = \bar{v}$ has the additional property that $c(u, \bar{v}) < c(u, \bar{v})$ for all $u \in N \setminus \{\sigma\}$. Let $\tilde{x} \in X$ maximize $\ell(\cdot, \bar{v})$ and suppose that* $\sup (c(u, \bar{v}) | u \in U, u \geqq G(\tilde{x}))$ *is attained. Then $\tilde{x}$ is a solution of the original problem.*

Proof. Let $u_1 \in U$ attain the supremum mentioned in the statement of the theorem. Then $-\alpha = -f_0(\tilde{x}) + \bar{c}(G(\tilde{x}), \bar{v}) \leqq -h(u_1) + c(u_1, \bar{v}) = -h(u_1) + c(u_1, \bar{v}) + c(u_1, \bar{v}) - c(u_1, \bar{v}) \leqq -\alpha + c(u_1, \bar{v}) - c(u_1, \bar{v})$. This implies that $c(u_1, \bar{v}) \leqq c(u_1, \bar{v})$, hence $u_1 = \sigma$. So $\tilde{x}$ is feasible. It also follows that $f_0(\tilde{x}) = \alpha$.

Remarks. In the case of inexact optimization a statement similar to Theorem 7 can be made which requires a slightly stronger additional property of $c(\cdot, \bar{v}) - c(\cdot, \bar{v})$, namely that the collection of its level sets forms a neighborhood base at σ.

It will be clear that the assumption that $\bar{c}(G(\tilde{x}), \bar{v}) = c(u, \bar{v})$ for some $u \in U$, $u \geqq G(\tilde{x})$ is not a heavy one. If, for instance, $c(\cdot, \bar{v})$ is u.s.c., concave and $\{u | c(u, \bar{v}) \geqq 0\}$ is compact, the assumption is already satisfied, because then all level sets of $c(\cdot, \bar{v})$ will be compact [19].

In the case of Example 2 (finite dimension) the possibility of solving the original problem by exact (or inexact) maximization of the Lagrangian corresponding to an exact multiplier has already been pointed out by Rockafellar [31]. Work of Zangwill [34], Pietrzykowski [25] and Howe [15] on such exactness in the case of Example 1 seems of less practical value since the Lagrangian is nondifferentiable (however, see [9]).

The conclusion that nondifferentiabilities are a necessary evil in exact methods, reached in [5], seems premature in view of both the case dealt with by Rockafellar and our framework. The conclusion reached in [5] is based upon a formulation of the Lagrangian which, in our model, could correspond to a coupling functional c that satisfies the additional restriction $c(-u, v) = c(u, v)$, $u \in U$, $v \in V$. There seems to be no a priori justification for such an approach. On the contrary, any picture of the situation shows us that if one requires in such a case smoothness of $c(\cdot, v)$ at σ to ensure differentiability of $\ell(\cdot, v)$, $v \in V$, thereby forcing the derivatives of all c-elementary functionals to be "zero" at σ, one forfeits immediately the possibility to have a c-subgradient of h at σ in all but a few very special cases.

Problems that are formulated as local problems, i.e., the search for a local minimum, still fit into our framework by an adequate specialization of X.

Let us finally consider the question of the existence of c-subgradients of the perturbation function at the zero perturbation; we have seen these subgradients constitute the exact multipliers, so knowledge about them could be very useful indeed (Theorem 7). Unfortunately, this will turn out to be a very difficult matter.

To begin with we shall restrict ourselves to the case where the perturbation space U is a topological vector space. For $a \in \bar{\mathbb{R}}^U$, $u_0 \in U$, $u \in U$ we define the *upper* and *lower Dini derivatives of a at u_0 in the direction u by*

$$\bar{D}a(u_0; u) = \limsup_{\lambda \downarrow 0} (a(u_0 + \lambda u) - a(u_0))/\lambda,$$

$$\underline{D}a(u_0; u) = \liminf_{\lambda \downarrow 0} (a(u_0 + \lambda u) - a(u_0))/\lambda.$$

Remark. Let $a \in \bar{\mathbb{R}}^U$. A necessary condition for c-subdifferentiability of a at σ is that there exists a $v \in V$ such that for all $u \in U$

$$\underline{D}a(\sigma; u) \geqq \bar{D}c_v(\sigma; u),$$

where c_v denotes $c(\cdot, v)$.

We shall say that $a \in \bar{\mathbb{R}}^U$ is *locally convex at* $u_0 \in U$ [3] if there exists a neighborhood N of u_0 such that for all $u \in N$, $0 \leqq \lambda \leqq 1$,

$$a(\lambda u + (1-\lambda)u_0) \leqq \lambda a(u) + (1-\lambda)a(u_0),$$

and we shall say that $a \in \mathbb{R}^U$ is *locally concave at* u_0 if $-a$ is locally convex at u_0.

The following lemma provides sufficient conditions for c-subdifferentiability in a situation where first order properties alone give enough information.

LEMMA 3. *Suppose that the coupling functional c is flexible at σ. If $a \in \bar{\mathbb{R}}^U$ is locally convex and finite at σ and c-tempered, and if there exist a $v \in V$ and a neighborhood N of σ such that c_v is locally concave at σ and such that for all $u \in N$*

$$\underline{D}a(\sigma; u) \geqq \bar{D}c_v(\sigma; u),$$

then $c - \partial a(\sigma) \neq \varnothing$.

Proof. Observe that there exists a neighborhood N' of σ in which $a(u) - a(\sigma) \geqq \underline{D}a(\sigma; u) \geqq \bar{D}c_v(\sigma; u) \geqq c_v(u)$, $u \in N'$. Then use the flexibility property.

The next result concerns a conclusion that can be drawn by studying the behavior of the directional derivatives in a neighborhood of σ.

LEMMA 4. *Suppose that the coupling functional c is flexible at σ. If $a \in \bar{\mathbb{R}}^U$ is c-tempered and if there exist a $v \in V$ and a (circled) neighborhood N of σ such that*

(a) *for all $u \in N$, $a - c_v$ is continuous and finite-valued on $[\sigma, u]$ (the line segment $\{\lambda u | 0 \leqq \lambda \leqq 1\}$),*

(b) *for all $u \in N$ and for all u' in $[\sigma, u]$, except at most a countable number, $\underline{D}a(u'; u) \geqq \bar{D}c_v(u'; u)$,*

then $c - \partial a(\sigma) \neq \varnothing$.

Proof. Apply [22, 34.1] and use flexibility.

As is well-known, second order conditions for a functional can guarantee its convexity in a small neighborhood. This motivates the following lemma, where we denote the topological dual (adjoint space) of U by U^*.

LEMMA 5. *Suppose that the coupling functional c is flexible at σ. If c has a linear component, that is, if for all $v \in V$ and $u^* \in U^*$ $c(\cdot, v) + \langle \cdot, u^* \rangle$ is a c-elementary functional, if a is c-tempered and if there exist a $v \in V$ and a neighborhood N of σ such that $a - c_v$ is convex in N, finite and continuous at σ, then $c - \partial a(\sigma) \neq \varnothing$.*

Lemmas 3, 4 and 5 can be useful in cases where the behavior of the perturbation function h near σ is known, for instance in the convex case. It will be clear that in the nonconvex case usually both first and second order properties of the perturbation function of the embedded problem must be known in order to prove (extended) subdifferentiability. This knowledge, of course, has to be formulated in terms of the original ingredients of the optimization problem, since actual computation of any value of h will usually be as difficult as solving the original problem.

Such sensitivity analysis has been conducted recently in [6], [10] and [20]. Although some interesting results have been obtained, the analysis made up until

now seems far from complete. We also refer to [14] for additional references.

Second order differentiability properties of the perturbation function have been discussed in [10]; here the perturbation parameter is supposed to be one-dimensional. Another way to guarantee second order differentiability of the perturbation function is to impose the standard second order sufficiency conditions of mathematical programming [12], [21], [31].

Taking a closer look at the matter, we remark that a result in [34] for the coupling functional of Example 1 can be derived directly from Theorem 7, Lemma 3 and the observation that the Slater condition imposed in [34] guarantees the existence of a subgradient—in the sense of convex analysis—of the (convex) perturbation function.

Another result on exact multipliers with the same coupling functional, a nonconvex local problem in [25], follows from Theorem 7, using Theorem 3 in [20].

Acknowledgment. The author wishes to thank the referee for stimulating criticism and for some helpful references.

REFERENCES

[1] K. J. ARROW, L. HURWICZ AND H. UZAWA, *Studies in Linear and Nonlinear Programming*, Stanford University Press, Stanford, Calif., 1958.

[2] M. S. BAZARAA, *Geometry and resolution of duality gaps*, Naval Res. Logist. Quart., 20 (1973), pp. 357–366.

[3] M. S. BAZARAA, J. J. GOODE AND M. Z. NASHED, *On the cones of tangents with applications to mathematical programming*, J. Optimization Theory Appl., 13 (1974), pp. 389–426.

[4] D. P. BERTSEKAS, *Multiplier methods: A survey*, Proc. 1975 IFAC 6th Triennial World Congress, Boston, Mass., to appear.

[5] ———, *Necessary and sufficient conditions for a penalty method to be exact*, Math. Prog., 9 (1975), pp. 87–99.

[6] V. V. BERESNEV AND B. N. PSHENICHNYI, *The differential properties of minimum functions*, Ž. Vyčisl. Mat. i Mat. Fiz., 14 (1974), pp. 639–651 = U.S.S.R. Computational Math. and Math. Phys., 14 (1974), pp. 101–113.

[7] C. BERGE, *Espaces Topologiques et Fonctions Multivoques*, Dunod, Paris, 1959.

[8] A. BRØNDSTED, *Convexification of conjugate functions*, Math. Scand., 36 (1975), pp. 131–136.

[9] A. R. CONN, *Constrained optimization using a nondifferentiable penalty function*, SIAM J. Numer. Anal., 10 (1973), pp. 760–784.

[10] V. F. DEMYANOV AND A. B. PEVNYĬ, *First and second marginal values of mathematical programming problems*, Dokl. Akad. Nauk SSSR, 207 (1972), pp. 277–280 = Soviet Math. Dokl., 13 (1972), pp. 1502–1506.

[11] K. H. ELSTER AND R. NEHSE, *Zur Theorie der Polarfunktionale*, Math. Operationsforsch. Statist., 5 (1974), pp. 3–21.

[12] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[13] A. M. GEOFFRION, *Duality in nonlinear programming: A simplified applications-oriented development*, SIAM Rev., 3 (1971), pp. 1–37.

[14] W. HOGAN, *Directional derivatives for extremal-value functions with applications to the completely convex case*, Operations Res., 21 (1973), pp. 188–209.

[15] S. HOWE, *New conditions for exactness of a simple penalty function*, this Journal, 11 (1973), pp. 378–381.

[16] A. D. IOFFE AND V. M. TIKHOMIROV, *Duality of convex functions and extremum problems*, Uspehi Mat. Nauk, 23 (1968), no. 6, pp. 51–116 = Russian Math. Surveys, 23 (1968), no. 6, pp. 53–124.

[17] R. JANIN, *Sur la dualité en programmation dynamique*, C.R. Acad. Sci. Paris Sér. A, 277 (1973), pp. 1195–1197.

[18] J. L. JOLY AND P. J. LAURENT, *Stability and duality in convex minimization problems*, Rev. Française Automat. Informat. Recherche Opérationelle Sér. Rouge, 5 (1971), no. 2, pp. 3–42.

[19] P. J. LAURENT, *Approximation et Optimisation*, Hermann, Paris, 1972.

[20] E. S. LEVITIN, *On differential properties of the optimum value of parametric problems of mathematical programming*, Dokl. Akad. Nauk SSSR, 215 (1974), pp. 792–795 = Soviet Math. Dokl., 15 (1974), pp. 603-608.

[21] D. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass., 1973.

[22] E. J. MCSHANE, *Integration*, Princeton University Press, Princeton, N.J., 1947.

[23] J. J. MOREAU, *Fonctionelles Convexes*, Séminaire sur les Équations aux Dérivées Partielles, Collège de France, Paris, 1966.

[24] ———, *Inf-convolution, sous-addivité, convexité des fonctions numériques*, J. Math. Pures Appl., 49 (1970), pp. 109–154.

[25] T. PIETRZYKOWSKI, *An exact potential method for constrained maxima*, SIAM J. Numer. Anal., 6 (1972), pp. 299–304.

[26] B. T. POLYAK, *The convergence rate of the penalty function method*, Ž. Vyčisl. Mat. i Mat. Fiz., 11 (1971), pp. 3–11 = U.S.S.R. Computational Math. and Math. Phys., 11 (1971), pp. 1–12.

[27] B. T. POLYAK AND N. V. TRET'YAKOV, *The method of penalty estimates for conditional extremum problems*, Ž. Vyčisl. Mat. i Mat. Fiz., 13 (1973), pp. 34–46 = U.S.S.R. Computational Math. and Math. Phys., 13 (1973), pp. 42–58.

[28] R. T. ROCKAFELLAR, *Duality and stability in extremum problems involving convex functions*, Pacific J. Math., 21 (1967), pp. 167–187.

[29] ———, *Convex functions and duality in optimization problems and dynamics*, Lecture Notes in Operations Research and Mathematical Economics, vol. 11, Springer-Verlag, New York, pp. 117–141.

[30] ———, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[31] ———*Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.

[32] E. A. WEISS, *Konjugierte Funktionen*, Arch. Math., 20 (1969), pp. 538–545.

[33] ——, *Verallgemeinerte konvexe Funktionen*, Math. Scand., 35 (1974), pp. 129–144.

[34] W. I. ZANGWILL, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

ON OPTIMAL CONTROL PROBLEMS WITH BOUNDED STATE VARIABLES AND CONTROL APPEARING LINEARLY*

H. MAURER†

Abstract. Necessary conditions for the switching function, holding at junction points of optimal interior and boundary arcs or at contact points with the boundary, are given. These conditions are used to derive necessary conditions for the optimality of junctions between interior and boundary arcs. The junction theorems obtained are similar to those developed for singular control problems in [1] and establish a duality between singular control problems and control problems with bounded state variables and control appearing linearly. The transition from unconstrained to constrained extremals is discussed with respect to the order p of the state constraint. A numerical example is given where the adjoint variables are not unique but form a convex set which is determined numerically.

1. Introduction. Jacobson, Lele and Speyer [2] and Hamilton [3] have studied the necessary conditions for junctions between optimal interior and boundary arcs in control problems with bounded state variables. Under the assumption of a *regular* Hamiltonian—this assumption usually holds if the control variable appears *nonlinearly*—they obtain a certain smoothness of the control at junction points. Depending on the order p of the state inequality constraint, it is then possible to predict whether the constrained extremal will contain a boundary arc or will only touch the boundary.

Control problems with bounded state variables and control variable appearing *linearly* have received little theoretical and practical attention. Here the Hamiltonian fails to be regular. Thus one cannot establish the smoothness of the control variable at junction points. In these problems the switching function plays a fundamental role. Instead of smoothness properties of the control variable, smoothness conditions for the switching function at junction or contact points are derived in this paper. These conditions are used to obtain necessary conditions for the optimality of junctions between interior and boundary arcs which are similar to junction conditions for singular control problems developed by McDanell and Powers [1]. The results were suggested by the observation that a boundary arc is a singular arc in the sense of the minimum principle.

2. Statement of the problem. We consider the following control problem with control appearing linearly: determine the scalar, *piecewise continuous* control $u(t)$, $t \in [0, T]$, which minimizes the functional

$$J(u) = G(x(T)) \tag{2.1}$$

subject to

$$\dot{x} = f(x, u) = f_1(x) + f_2(x)u, \tag{2.2}$$

$$x(0) = x_0, \qquad \psi(x(T)) = 0, \tag{2.3}$$

$$|u(t)| \leqq K(t), \quad K(t) > 0, \quad 0 \leqq t \leqq T, \tag{2.4}$$

* Received by the editors August 27, 1974, and in revised form June 10, 1976.

† Mathematisches Institut der Universität Würzburg, 87 Würzburg, Am Hubland, West Germany. The author was supported by a Postdoctoral Fellowship of the Canada Council at the University of British Columbia, Vancouver, Canada.

and the scalar state inequality constraint of order p

$$S(x) \leqq \alpha, \qquad \alpha \in \mathbb{R}. \tag{2.5}$$

The state x is an n-vector. Henceforth all vectors are column-vectors. The functions $G: \mathbb{R}^n \to \mathbb{R}$, $\psi: \mathbb{R}^n \to \mathbb{R}^k$, $k < n$, are supposed to be differentiable. The functions $f_1, f_2: \mathbb{R}^n \to \mathbb{R}^n$, $S: \mathbb{R}^n \to \mathbb{R}$ are assumed to be analytic in a suitable domain; $K(t)$ is assumed to be analytic in $[0, T]$. These assumptions guarantee the validity of the expressions developed in the sequel, although less stringent differentiability properties would suffice which are, however, lengthy to state in each different case. The terminal time T is fixed.

Nonautonomous control problems and control problems with an integral cost criterion in (2.1) or with free end-time can be reduced to the above form by introducing additional state variables.

An extremal arc of (2.1)–(2.4) is called an *unconstrained* extremal whereas an extremal arc of (2.1)–(2.5) is called a *constrained* extremal. For given $\alpha \in \mathbb{R}$ the state constraint (2.5) is called *active* if the optimal unconstrained trajectory $x^0(t)$ violates (2.5).

Along a trajectory $x(t)$ of (2.2) the ith time derivative of $S(x(t))$ will be denoted by S^i, $i \geqq 0$, where $S^0 = S$. By definition of the order p of the inequality constraint (2.5), S^p is the first derivative containing the control u explicitly. S^p contains u linearly because of (2.2) and hence

$$\begin{aligned} S^i &= S^i(x), \qquad i = 0, \cdots, p-1, \\ S^p &= S^p(x, u) = a(x) + b(x)u. \end{aligned} \tag{2.6}$$

A subarc of $x(t)$ for which $S(x(t)) < \alpha$ is called an *interior arc*; a *boundary arc* is a subarc of $x(t)$ where $S(x(t)) \equiv \alpha$ for $t_1 \leqq t \leqq t_2$ with $0 \leqq t_1 < t_2 \leqq T$. Here t_1 and t_2 are called the *entry*- and *exit-time* of the boundary arc; t_1 and t_2 are also termed *junction points*. An arc $x(t)$ is said to have a *contact point* with the boundary at $t_1 \in (0, T)$ if $S(x(t_1)) = \alpha$ and $S(x(t)) < \alpha$ for $t \neq t_1$ in a neighborhood of t_1. If $x(t)$ has a contact point with the boundary at t_1 for $p \geqq 2$, then the relation $S^1(x(t_1)) = 0$ follows and hence $x(t)$ *touches* the boundary at t_1.

By differentiation of $S(x(t)) \equiv \alpha$, $t_1 \leqq t \leqq t_2$, one gets the *entry conditions* of a boundary arc

$$S(x(t_1)) = \alpha, \quad S^i(x(t_1)) = 0, \qquad i = 1, \cdots, p-1. \tag{2.7}$$

The *boundary control* is determined by $S^p(x, u) = 0$ which gives in view of (2.6) the feedback expression

$$u_b = u(x) = -a(x)/b(x). \tag{2.8}$$

Along $x(t)$ we consider from now on the functions S^i, a, b, u_b as functions of time.

General Assumption 2.1. The following conditions hold on a boundary arc in $[t_1, t_2]$: $b(t) \neq 0$ for $t \in [t_1, t_2]$ and $|u_b(t)| < K(t)$, i.e., $|a(t)| < |b(t)|K(t)$, for $t \in (t_1, t_2)$.

By virtue of $b(t) \neq 0$ and the analyticity of the functions f_1, f_2, S, the boundary control $u_b(t)$ is analytic in (t_1, t_2).

3. Necessary conditions. The necessary conditions are developed in [2], [4], [5]. It has been shown in [6] that the function η^* of bounded variation in [4, Thm. 3.2], [5, Thm. 14.2] has a continuous derivative η on the interior of a boundary arc for pth order state inequality constraints provided that Assumption 2.1 holds. This result follows by rigorizing the formal arguments in [2].

Define the Hamiltonian by

$$H(x, u, \lambda, \eta) = \lambda^T f_1(x) + \lambda^T f_2(x) u + \eta S(x) \tag{3.1}$$

where $\lambda \in \mathbb{R}^n$, $\eta \in \mathbb{R}$ and where the superscript denotes the transpose. The necessary conditions of the minimum principle then are the following:

1. There exists a scalar function $\eta(t) \geqq 0$, a real number $\eta_0 \geqq 0$ and $\sigma \in \mathbb{R}^k$ such that the adjoint variable $\lambda(t) \in \mathbb{R}^n$ satisfies

$$\dot{\lambda}^T = -\lambda^T f_x - \eta S_x \quad \text{a.e.}, \qquad \lambda^T(T) = \eta_0 G_x(x(T)) + \sigma^T \psi_x(x(T)). \tag{3.2}$$

2. The function $\eta(t)$ satisfies $\eta(t)(S(x(t)) - \alpha) \equiv 0$, $t \in [0, T]$, and is continuous on the interior of a boundary arc.
3. The jump condition at a contact point or junction point t_1 is

$$\lambda^T(t_1^+) = \lambda^T(t_1^-) - \nu_1 S_x(x(t_1)), \qquad \nu_1 \geqq 0. \tag{3.3}$$

4. The optimal control $u(t)$ minimizes the Hamiltonian, i.e.,

$$H(x(t), u(t), \lambda(t), \eta(t)) = \min_{|u| \leqq K(t)} H(x(t), u, \lambda(t), \eta(t)). \tag{3.4}$$

The coefficient of u in (3.1) is called the *switching function* and is denoted by

$$\phi(t) = \lambda^T(t) f_2(x(t)). \tag{3.5}$$

The switching function determines the optimal control in (3.4) as follows.

Optimal control for interior arcs: Nonsingular case. Let $\phi(t)$ have only isolated zeros on a subinterval $I \subset [0, T]$. Then the optimal control is *nonsingular* on I and is given by

$$u(t) = -K(t) \operatorname{sgn} \phi(t), \qquad t \in I. \tag{3.6}$$

Singular case. If $\phi(t) \equiv 0$ on a subinterval $I \subset [0, T]$ then $u(t)$ is *singular* on I. Let $q \geqq 1$ be the order of the singular arc; cf. [7]. $\phi^{(2q)}$ is the lowest order time derivative of ϕ which contains u explicitly and we have

$$\phi^{(2q)} = A(x, \lambda) + B(x, \lambda) u. \tag{3.7}$$

We note that $\phi^{(2q)} = \lambda^T \varphi_{2q}$ in the terminology of § 4. Let $A(t) = A(x(t), \lambda(t))$, $B(t) = B(x(t), \lambda(t))$. It is assumed throughout this paper that the strengthened generalized Legendre–Clebsch-condition (GLC-condition)

$$(-1)^q B(t) > 0, \qquad t \in I, \tag{3.8}$$

holds and that the singular control $u_s(t) = -A(t)/B(t)$ determined by $\phi^{(2q)}(t) \equiv 0$ satisfies

$$|u_s(t)| < K(t), \quad \text{i.e.,} \quad |A(t)| < |B(t)| K(t) \quad \text{for } t \in \mathring{I}. \tag{3.9}$$

Optimal control for boundary arcs. The optimal control is the boundary control u_b in (2.8). The minimization in (3.4) then implies in view of Assumption 2.1 that

$$H_u = \phi(t) = 0, \qquad t_1^+ \leqq t \leqq t_2^-. \tag{3.10}$$

Thus the boundary control is a singular control in the sense of the minimum principle. One can expect therefore a duality between singular control problems (without state constraints) and control problems with bounded state variables and control appearing linearly. Such a duality has already been conjectured in the literature; cf. [1], [2]. The junction theorems in § 5 will clarify this duality.

4. Relations for the switching function at contact points or junction points. Let $u(t)$ be a control which is analytic on a subinterval $I \subset [0, T]$ and let $x(t)$ be an analytic solution of (2.2) on I. Define the functions $\varphi_i : I \to \mathbb{R}^n$ recursively as in [3, (17)] by

$$\varphi_0 = f_u = f_2, \quad \varphi_{i+1} = \dot{\varphi}_i - f_x \varphi_i, \quad i \geqq 0. \tag{4.1}$$

Then the following relations hold on I:

$$S_x^{i-1-k} \varphi_k = (-1)^k S_u^i, \quad i = 1, \cdots, p, \quad k = 0, \cdots, i-1. \tag{4.2}$$

The formulas are proved by induction over i and k. For $i = 1, \cdots, p$ and $k = 0$ (4.2) follows from $S^i = S_x^{i-1} f$ and $\varphi_0 = f_u$. We outline the induction step $i \to i+1$ and $k \to k+1$. Suppose that (4.2) is true for $i+1$ and k and also for i and k. Then we get $S_x^{i-(k+1)} \varphi_k = (-1)^k S_u^i = 0$ as $i < p$. Differentiating the last equation yields

$$0 = S_{xx}^{i-(k+1)} f \varphi_k + S_x^{i-(k+1)} \dot{\varphi}_k.$$

Using this relation we obtain in view of $S_x^{i-k} = (S_x^{i-(k+1)} f)_x$ and (4.1):

$$S_x^{i-k} \varphi_k = (S_{xx}^{i-(k+1)} f + S_x^{i-(k+1)} f_x) \varphi_k = -S_x^{i-(k+1)} \varphi_{k+1}.$$

Thus, (4.2) is shown to be true for $i+1$ and $k+1$.

Setting $k = i-1$ in (4.2) we find the following formulas given without proof in [3, (18)]:

$$S_x \varphi_i = \begin{cases} 0, & i < p-1, \\ (-1)^{p-1} S_u^p = (-1)^{p-1} b, & i = p-1. \end{cases} \tag{4.3}$$

On a subinterval I of an *interior arc*, where the control is analytic, the ith time derivative $\phi^{(i)}$ is given by

$$\phi^{(i)}(t) = \lambda^T(t) \varphi_i(t), \qquad t \in I, \quad i \geqq 0. \tag{4.4}$$

By using the adjoint equation (3.2) and (4.3) we obtain on a *boundary arc* in $[t_1, t_2]$ the relations

$$\phi^{(i)}(t) = \lambda^T(t) \varphi_i(t) = 0, \qquad t_1^+ \leqq t \leqq t_2^-, \quad i = 0, \cdots, p-1, \tag{4.5}$$

$$\phi^{(p)}(t) = \lambda^T(t) \varphi_p(t) - (-1)^{p-1} \eta(t) b(t) = 0, \qquad t_1^+ \leqq t \leqq t_2^-. \tag{4.6}$$

Let k be the lowest integer such that φ_k contains the control u explicitly. It follows from the theory of singular control problems (cf. [7]) that k is *even*, i.e., $k = 2q$ with $q \geqq 1$; cf. (3.7).

LEMMA 4.1. *Let $t_1 \in (0, T)$ be a contact point or junction point of an optimal control u and let u be piecewise continuous in a neighborhood of t_1. Let $u^{(r)}$, $r \geqq 0$, be the lowest order time derivative of u which is discontinuous at t_1. Suppose that $p \leqq 2q + r$. If $\nu_1 \geqq 0$ is the jump in* (3.3), *then the following relations hold*:

$$\phi^{(i)}(t_1^+) = \phi^{(i)}(t_1^-), \qquad i = 0, \cdots, p-2, \tag{4.7}$$

$$\phi^{(p-1)}(t_1^+) = \phi^{(p-1)}(t_1^-) - \nu_1(-1)^{p-1}b(t_1). \tag{4.8}$$

Proof. As the control u is piecewise continuous and hence piecewise analytic in a neighborhood of t_1, the derivatives $\phi^{(i)}(t_1^{\pm})$ exist. The assumption $p \leqq 2q + r$ (which is always satisfied for $p \leqq 2$) implies that φ_i is continuous at t_1, $i = 0, \cdots, p-1$. Substituting the jump condition (3.3) we get for $i = 0, \cdots, p-1$:

$$\phi^{(i)}(t_1^+) = \lambda^T(t_1^+)\varphi_i(t_1) = \phi^{(i)}(t_1^-) - \nu_1 S_x(x(t_1))\varphi_i(t_1).$$

Combining this with (4.3) gives the equations (4.7), (4.8).

Now let t_1 be the *entry-time* of a boundary arc. Then $\phi^{(i)}(t_1^+) = 0$ for $i \geqq 0$ by virtue of (3.10) and (4.7), (4.8) imply

$$\phi^{(i)}(t_1^-) = 0 \quad \text{for } i = 0, \cdots, p-2, \tag{4.9}$$

$$\nu_1 = (-1)^{p-1}\phi^{(p-1)}(t_1^-)/b(t_1) \geqq 0. \tag{4.10}$$

The relations (4.9), (4.10) remain valid at the exit-time t_2 with t_1^- (resp. ν_1) replaced by t_2^+ (resp. $-\nu_2$). Equation (4.9) constitutes $p-1$ additional relations at t_1 besides (2.7). Also, (4.9) should be viewed as an analogy to the fact that for a regular Hamiltonian (this is mostly the case if the control appears nonlinearly) the control $u(t)$ and its first $p-2$ time derivatives are continuous at t_1; cf. [2]. Thus (4.10) corresponds to [2, (81)].

5. Junction theorems. This section is concerned with the study of junctions between interior arcs and boundary arcs. As a boundary arc is a singular arc one can expect junction theorems similar to those in McDanell and Powers [1]. In this paper the admissible controls were assumed to be piecewise continuous and hence all controls are *piecewise analytic* in the sense of [1] as the functions f_1, f_2, S, K were assumed to be analytic. The following proofs are carried through for the entry-time t_1 but are also valid for the exit-time t_2 with minor modifications.

5.1. Junctions between interior nonsingular arcs and boundary arcs.

THEOREM 5.1. *Let t_1 be the time at which an interior nonsingular arc and a boundary arc of an optimal control u are joined. Let $u^{(r)}$, $r \geqq 0$, be the lowest order derivative of u which is discontinuous at t_1 and let $p \leqq 2q + r$. If $\nu_1 > 0$, then $p + r$ is an even integer.*

Proof. This proof is closely modeled after the proof for Theorem 1 in [1]. Let $\varepsilon > 0$ be an arbitrary number such that $t_1 - \varepsilon$ is a point on the interior arc and $t_1 + \varepsilon$ is a point on the boundary arc. By $u_n^{(i)}(t_1)$ and $u_b^{(i)}(t_1)$ we mean the limit as $\varepsilon \to 0$ of $u^{(i)}(t_1 - \varepsilon)$ and $u^{(i)}(t_1 + \varepsilon)$ respectively. Expanding $S(t_1 - \varepsilon)$ in a Taylor series about t_1 one finds that

$$S^{(p+r)}(t_1^-) = \frac{d^r}{dt^r}(a + bu)(t_1^-) \tag{5.1}$$

is the first nonzero term of the Taylor series. Since $a+bu_b \equiv 0$ and hence $a=-bu_b$ one obtains

$$S(t_1-\varepsilon)=(-1)^{p+r}\frac{\varepsilon^{p+r}}{(p+r)!}b(t_1)[u_n^{(r)}(t_1)-u_b^{(r)}(t_1)]+o(\varepsilon^{p+r}). \tag{5.2}$$

Let $\sigma=-\operatorname{sgn}\phi(t_1-\varepsilon)$. Then $u_n(t)=\sigma K(t)$ and thus $u_n^{(i)}(t_1)=\sigma K^{(i)}(t_1)$ for $i\geqq 0$. A Taylor expansion on the boundary arc gives

$$\sigma K(t_1+\varepsilon)-u(t_1+\varepsilon)=\frac{\varepsilon^r}{r!}[u_n^{(r)}(t_1)-u_b^{(r)}(t_1)]+o(\varepsilon^r). \tag{5.3}$$

Substituting this in (5.2) yields

$$S(t_1-\varepsilon)=(-1)^{p+r}\varepsilon^p\frac{r!}{(p+r)!}b(t_1)[\sigma K(t_1+\varepsilon)-u(t_1+\varepsilon)]+o(\varepsilon^{p+r}). \tag{5.4}$$

Since $S(t_1-\varepsilon)<0$ on the interior arc, (5.4) implies (as $\varepsilon>0$)

$$-1=(-1)^{p+r}\operatorname{sgn} b(t_1)\operatorname{sgn}[\sigma K(t_1+\varepsilon)-u(t_1+\varepsilon)]. \tag{5.5}$$

By assumption $\nu_1>0$ and hence $\phi^{(p-1)}(t_1^-)\neq 0$ from (4.10). Expanding $\phi(t_1-\varepsilon)$ in Taylor series and using (4.9), (4.10), we find

$$\sigma=-\operatorname{sgn}\phi(t_1-\varepsilon)=(-1)^p\operatorname{sgn}\phi^{(p-1)}(t_1^-)=-\operatorname{sgn} b(t_1). \tag{5.6}$$

Combining this with Assumption 2.1 (that $|u(t_1+\varepsilon)|<K(t_1+\varepsilon)$) yields $\operatorname{sgn} b(t_1)\operatorname{sgn}[\sigma K(t_1+\varepsilon)-u(t_1+\varepsilon)]=\operatorname{sgn} b(t_1)\operatorname{sgn}\sigma=-1$. Hence it follows from (5.5) that $(-1)^{p+r}=1$ and thus $p+r$ is even.

COROLLARY 5.2. *Let the point t_1 be given and let $u^{(r)}$, $r\geqq 0$, be the lowest order derivative of an optimal control u which is discontinuous at t_1 when t_1 is considered as a junction point between an interior nonsingular arc and a boundary arc. Let $p\leqq 2q+r$.*

(i) *If $p+r$ is odd and $\nu_1>0$, i.e., $\phi^{(p-1)}(t_1^-)\neq 0$, then t_1 can only be a contact point with the boundary.*

(ii) *Let $p+r$ be odd. Then $\nu_1=0$ if t_1 is a junction point between an interior nonsingular arc and a boundary arc.*

To our knowledge only the case $r=0$ occurs in the numerical examples treated in the literature. A sufficient condition for $r=0$ is $|a(t_1)|<|b(t_1)|K(t_1)$. In particular we have $r=0$ for $a(t_1)=0$.

Let us investigate now the case $\nu_1=0$ in (4.10), i.e., $\phi^{(p-1)}(t_1^-)=0$, as it has been done for a regular Hamiltonian in [3]. The next result is *dual* to a result for singular control problems [1, Thm. 2].

THEOREM 5.3. *Let t_1 be a point at which a nonsingular interior arc and a boundary arc of an optimal control u are joined. Let $\phi^{(p+m)}(t_1^-)$, $m\geqq 0$, be the lowest order nonvanishing derivative of ϕ and let $u^{(r)}$, $r\geqq 0$, be the lowest order derivative of u which is discontinuous at t_1.*

(i) *If $p+m<2q+r$, then $p+r+m$ is an odd integer.*

(ii) *Let $p\leqq 2q+r\leqq p+m$ and let j be the lowest integer such that $\lambda^T(t_1)\varphi_j(t_1^+)\neq 0$. Then $2q+r\leqq j$ and $-\operatorname{sgn}(\lambda^T(t_1)\varphi_{p+m}(t_1^-)\lambda^T(t_1)\cdot\varphi_j(t_1^+))=(-1)^{p+r+m}$.*

Proof. The proof is based on the proof of Theorem 5.1 and it suffices to modify (5.6). Let $j = \min\{i | \lambda^T(t_1)\varphi_i(t_1^+) \neq 0\}$. Then by definition of the integer m we have $j \gtrless p$ as $p \lesseqgtr 2q+r$. Differentiating the expression (4.6) on the boundary then yields $\eta^{(i)}(t_1^+) = 0$ for $i = 0, \cdots, j-p-1$ and

$$\lambda^T(t_1)\varphi_j(t_1^+) = (-1)^{p-1}\eta^{(j-p)}(t_1^+)b(t_1) \neq 0. \tag{5.7}$$

We have $\eta^{(j-p)}(t_1^+) > 0$ as $\eta(t) \gtrless 0$. Now define $s = \operatorname{sgn}(\lambda^T(t_1)\varphi_{p+m}(t_1^-)\lambda^T(t_1) \cdot \varphi_j(t_1^+))$ where $\phi^{(p+m)}(t_1^-) = \lambda^T(t_1)\varphi_{p+m}(t_1^-) \neq 0$. Then (5.6) is replaced by

$$\sigma = -\operatorname{sgn}\phi(t_1-\varepsilon) = (-1)^{p+m+1}\operatorname{sgn}\phi^{(p+m)}(t_1^-) = (-1)^m s \cdot \operatorname{sgn} b(t_1). \tag{5.8}$$

In this equation (5.7) and $\eta^{(j-p)}(t_1^+) > 0$ are used. We then have $\operatorname{sgn} b(t_1) \operatorname{sgn}[\sigma K(t_1+\varepsilon) - u(t_1+\varepsilon)] = \operatorname{sgn} b(t_1) \operatorname{sgn}\sigma = (-1)^m s$. Hence (5.5) yields

$$-1 = (-1)^{p+r+m}s. \tag{5.9}$$

If $p+m < 2q+r$, then $\varphi_i(t)$ is continuous at t_1 for $i = 0, \cdots, p+m$, in which case $j = p+m$ and $s = 1$. Then (5.9) implies that $p+r+m$ is odd. If $p \lesseqgtr 2q+r \lesseqgtr p+m$, then $j \gtrless 2q+r$ and part (ii) of the theorem follows from (5.9).

The proofs of Theorems 5.1 and 5.3 motivate the interesting fact that the conditions $\nu_1 \gtrless 0$ and $\eta(t) \gtrless 0$ play the dual role to the GLC-condition $(-1)^q B(t) \gtrless 0$ in singular control problems.

Summing up we find that for the normal case $r = 0$ a rough *classification* of the constrained extremals with respect to the order p is as follows: for $p = 1$ boundary arcs and contact points are possible and $\nu_1 = 0$ holds at every junction point or contact point provided that $|a(t_1)| < |b(t_1)|K(t_1)$. The last statement anticipates Theorem 5.6(i). Corollary 5.2(i) then states that the constrained extremal *touches* the boundary only for *p odd*, $p \gtrless 3$ and $\nu_1 > 0$. The same result holds for a regular Hamiltonian; cf. [2]. Finally for $\nu_1 > 0$ and *p even* contact points and boundary arcs are possible.

Remarks. 1. One should use care in applying the junction theorems of this section in the presence of several control variables. Further work on the exact differentiability properties of the control components at junction points is required to obtain similar junction theorems in this case.

2. It is possible to derive formally conditions for *nonanalytic* junctions as in [1] where the control $u(t)$ has an infinite number of switches in the neighborhood of a junction point t_1 on the interior nonsingular arc. We shall not follow this up due to a complete lack of examples.

5.2. Junctions between interior singular arcs and boundary arcs. The next theorem is nearly identical to [1, Thm. 1] and is remarkable insofar as it does not involve the order p of the state inequality constraint.

THEOREM 5.4. *Let t_1 be a point where an interior singular arc and a boundary arc of an optimal control u are joined. Let q be the order of the singular arc and assume that the strengthened GLC-condition $(-1)^q B(t_1) > 0$ holds. Let $u^{(r)}$, $r \gtrless 0$, be the lowest order derivative of u which is discontinuous at t_1 and let $p \lesseqgtr 2q+r$. Then $\nu_1 = 0$, i.e., the multiplier $\lambda(t)$ is continuous at t_1, and $q+r$ is an odd integer.*

Proof. The proof is similar to those for Theorems 5.1 and 5.3. The singular control u_s and the boundary control u_b are determined by $A+Bu_s=0$ and $a+bu_b=0$. The Taylor expansion (5.2) becomes here

$$0>S(t_1-\varepsilon)=(-1)^{p+r}\frac{\varepsilon^{p+r}}{(p+r)!}b(t_1)[u_s^{(r)}(t_1)-u_b^{(r)}(t_1)]+o(\varepsilon^{p+r}). \tag{5.10}$$

On the interior singular arc the relation $\phi^{(i)}(t)=\lambda^T(t)\varphi_i(t)=0$ holds for all $i\geqq 0$. This gives $\nu_1=0$ by virtue of (4.10) and $p\leqq 2q+r$. Hence $\lambda(t)$ is continuous at t_1. Furthermore $\varphi_i(t)$ is continuous at t_1 for $i<2q+r$. Then we obtain

$$\begin{aligned}&\lambda^T(t_1)\varphi_i(t_1^+)=0, \qquad i<2q+r,\\ &\lambda^T(t_1)\varphi_{2q+r}(t_1^+)\neq 0\end{aligned} \tag{5.11}$$

as $u^{(r)}$ is discontinuous at t_1. Define $j=2q+r-p$ and differentiate the relation (4.6) on the boundary j times, whence $\eta^{(i)}(t_1^+)=0$ for $i=0,\cdots,j-1$ and

$$(-1)^{p-1}\eta^{(j)}(t_1^+)b(t_1)=(A+Bu_b)^{(r)}(t_1). \tag{5.12}$$

Subtracting the identity $(A+Bu_s)^{(r)}(t_1)=0$ from (5.12) results in

$$(-1)^{p-1}\eta^{(j)}(t_1^+)b(t_1)=B(t_1)(u_b^{(r)}(t_1)-u_s^{(r)}(t_1)). \tag{5.13}$$

Substituting this in (5.10) and taking the sign on both sides gives

$$-1=(-1)^r \operatorname{sgn}\eta^{(j)}(t_1^+)\operatorname{sgn}B(t_1)=(-1)^{r+q} \tag{5.14}$$

as $\eta^{(j)}(t_1^+)>0$ and $(-1)^qB(t_1)>0$. The conclusion that $q+r$ is odd then follows from (5.14).

In the normal case $r=0$, Theorem 5.4 implies that boundary arcs are possible if q is odd whereas the constrained extremal touches only the boundary if q is even. A necessary and sufficient condition for $r=0$ is contained in the next lemma, whose proof is trivial.

LEMMA 5.5. *Let t_1 be a junction point of an interior singular arc and a boundary arc. Then the control u is discontinuous at t_1 iff $(aB-bA)(t_1)\neq 0$.*

In the presence of several control variables Theorem 5.4 will remain valid only in special cases. A numerical example with two variables and $p=1$, $q=1$ is discussed in [8, Example 7.1].

5.3. The case $p=1$. In numerical examples the constrained extremals contain in general only boundary arcs. It will be shown in Theorem 5.6(ii) that the constrained extremal has at least one boundary arc. Let t_1 and t_2 be the entry-time and exit-time of a boundary arc and assume $r=0$. If the interior arc is nonsingular, then $\nu_1=0$ and $\nu_2=0$ by Corollary 5.2(ii) and (4.10) yields the following relations for the switching function:

$$\phi(t_1^-)=0, \qquad \phi(t_2^+)=0. \tag{5.15}$$

These relations are important for the numerical calculation of the multipliers $\lambda(t)$. Moreover the integer m in Theorem 5.3 must be even for boundary arcs if $1+m<2q$. To our knowledge only examples with $m=0$, i.e., $\dot\phi(t_1^-)\neq 0$ and $\dot\phi(t_2^+)\neq 0$, are known so far; cf. Example 2, § 7 below. If the interior arc is singular, then $\nu_1=\nu_2=0$ follows from Theorem 5.4.

THEOREM 5.6. *Let $p=1$ and assume that $|a(t_1)|<|b(t_1)|K(t_1)$ holds at every contact point t_1.*

(i) *Let t_1 be a contact point with the boundary. Then $\nu_1=0$ and $\phi(t_1)=0$. The control u is discontinuous at t_1 if t_1 is interior to a nonsingular arc or if t_1 is a junction point between a nonsingular and a singular arc. If $(aB-bA)\cdot(t_1)\neq 0$, then t_1 cannot be interior to a singular arc.*

(ii) *Let the unconstrained extremal be uniquely defined by the minimum principle. Then a constrained extremal corresponding to an active constraint* (2.5) *cannot have only contact points but contains at least one boundary arc.*

Proof. At a contact point t_1 one has $S(t_1)=0$ and $S(t_1-\varepsilon)<0$, $S(t_1+\varepsilon)<0$ for $\varepsilon>0$ small. Hence

$$S^1(t_1^-)=(a+bu)(t_1^-)\geqq 0, \qquad S^1(t_1^+)=(a+bu)(t_1^+)\leqq 0. \tag{5.16}$$

Case (a). Let t_1 be interior to a *nonsingular* arc. If we assume that the control u is continuous at t_1, then (5.16) implies $(a+bu)(t_1)=0$, which contradicts $u(t_1)=\pm K(t_1)$ and the assumption $|a(t_1)|<|b(t_1)|K(t_1)$. Therefore u is discontinuous at t_1 and we get by subtracting the two inequalities in (5.16)

$$b(t_1)(u(t_1^-)-u(t_1^+))>0. \tag{5.17}$$

We have by (4.8)

$$\phi(t_1^+)=\phi(t_1^-)-\nu_1 b(t_1), \qquad \nu_1\geqq 0. \tag{5.18}$$

Now $b(t_1)>0$ implies $u(t_1^-)>0$, $u(t_1^+)<0$ and therefore $\phi(t_1^-)\leqq 0$, $\phi(t_1^+)\geqq 0$ by the minimum principle. Hence $\nu_1 b(t_1)\leqq 0$ by (5.18) resulting in $\nu_1=0$. Similarly, $b(t_1)<0$ implies $\nu_1 b(t_1)\geqq 0$ and thus $\nu_1=0$.

Case (b). Let the point t_1 lie on a *singular* arc. If t_1 is interior to this singular arc, then clearly $\phi(t_1^-)=\phi(t_1^+)=0$ and $\nu_1=0$ follows from (5.18). The singular control u_s is continuous at t_1 and (5.16) implies $(a+bu_s)(t_1)=0$, which is equivalent to $(aB-bA)(t_1)=0$. Then it follows from $(aB-bA)(t_1)\neq 0$ that t_1 must be a junction point between a singular and a nonsingular arc. In this case the assumption $|a(t_1)|<|b(t_1)|K(t_1)$ also yields that u is discontinuous at t_1. Let $u(t_1^-)$ be the singular control. Then $\phi(t_1^-)=0$ in (5.18) and arguing as in Case (a) gives the desired result $\nu_1=0$. This proves part (i).

Suppose now that the constrained extremal has only contact points. Then $\nu_1=0$ holds at every contact point by (i) and thus $\lambda(t)$ is continuous on $[0, T]$. So, the constrained extremal satisfies also the necessary conditions of the minimum principle for the unconstrained extremal. This contradicts the uniqueness assumption as the equality of the unconstrained and constrained extremal is ruled out by the assumption that the constraint (2.5) is active. This proves part (ii).

6. Transition from unconstrained to constrained extremals. An attempt is made to outline the *qualitative behavior* of the constrained extremals depending on the order p and the parameter α in (2.5). In particular, we want to discuss for what parameters α boundary arcs may occur. We restrict the discussion of boundary arcs to the orders $p=1$ and *p even* for which the normal cases $r=m=0$ in Theorem 5.3 and $r=0$, $\nu_1>0$ in Theorem 5.1 are assumed. Let $x^0(t)$, $u^0(t)$,

$\lambda^0(t)$ be the unconstrained extremal and let $\phi^0(t)$ be the unconstrained switching function. Define

$$\alpha_0 = \max\{S(x^0(t)) \mid t \in [0, T]\} = S(x^0(t_1)). \tag{6.1}$$

The following assumption is made throughout this section.

Assumption 6.1. There are only finitely many points t_1 satisfying (6.1) and $t_1 = 0$ and $t_1 = T$ do not satisfy (6.1).

This assumption means that the unconstrained extremal has only finitely many contact points and no boundary arcs for the *trivial* constraint $S(x) \leqq \alpha_0$. The constraint (2.5) is only of interest for $\alpha \leqq \alpha_0$ and is active for $\alpha < \alpha_0$. We denote the problem (2.1)–(2.5) by (P_α) and an extremal of (P_α) by $x(t; \alpha), u(t; \alpha), \lambda(t; \alpha)$. The following concept of stability is useful in order to describe the qualitative behavior of the extremals.

DEFINITION 6.2. (P_α) is called *strongly stable at* $\bar{\alpha}$ if there exists a neighborhood V of $\bar{\alpha}$ such that (i) (P_α) has a unique extremal $x(t; \alpha), u(t; \alpha), \lambda(t; \alpha)$ with $\eta_0 = 1$ in (3.2) for $\alpha \in V$, and (ii) the function $\alpha \to (x(\cdot\,; \alpha), u(\cdot\,; \alpha), \lambda(\cdot\,; \alpha))$ is continuous at $\bar{\alpha}$.

LEMMA 6.3. *Let* $p = 1$ *and assume that* $|a(t_1)| < |b(t_1)|K(t_1)$ *holds at every point* t_1 *in* (6.1).

(i) *The unconstrained extremal satisfies the entry-conditions* (2.7), (5.15) *for* $\alpha = \alpha_0$ *at every point* t_1 *in* (6.1).

(ii) *Let* (P_α) *be strongly stable at* α_0. *Then there exists* $\alpha_1 < \alpha_0$ *such that the constrained extremal for* $\alpha \in [\alpha_1, \alpha_0)$ *contains at least one boundary arc. Any such boundary arc evolves continuously with respect to* α *from a point* t_1 *in* (6.1).

Proof. Part (i) is a consequence of Theorem 5.6(i). If (P_α) is strongly stable at α_0 then the hypothesis of this lemma implies that the hypothesis of Theorem 5.6 is satisfied for $\alpha \in [\alpha_1, \alpha_0)$ with suitable $\alpha_1 < \alpha_0$. Hence (ii) follows from Theorem 5.6(ii).

LEMMA 6.4. *Let* p *be even and let* $p \leqq 2q$. *Assume that at every point* t_1 *in* (6.1) *the following condition holds: either* $|a(t_1)| < |b(t_1)|K(t_1)|$ *if* u^0 *is discontinuous at* t_1 *or* ϕ^0 *does not have a zero of order* p *at* t_1 *if* μ^0 *is continuous at* t_1.

(i) *The unconstrained extremal does not satisfy the entry-conditions* (2.7), (4.9) *for* $\alpha = \alpha_0$ *at every point* $t_1 \in [0, T]$.

(ii) *Let* (P_α) *be strongly stable at* α_0. *Then there exists* $\alpha_1 < \alpha_0$ *such that the constrained extremal for* $\alpha \in [\alpha_1, \alpha_0)$ *touches the boundary only. Moreover, if* ϕ^0 *has only simple zeros in* $(0, T)$ *and if* $\phi^0(0) \neq 0, \phi^0(T) \neq 0, \phi^0(t_1) \neq 0$ *for all points* t_1 *in* (6.1), *then* $\alpha_1 < \alpha_0$ *can be chosen in such a way that the constrained extremal for* $\alpha \in [\alpha_1, \alpha_0)$ *has the same number of switching points as the unconstrained extremal.*

Proof. (i). Assume that (2.7), (4.9) hold at a point $t_1 \in [0, T]$. Then this point t_1 must also satisfy (6.1). Since the function $S(t) = S(x(t))$ has a maximum at t_1, the inequalities

$$S^p(t_1^{\pm}) = (a + bu^0)(t_1^{\pm}) \leqq 0 \tag{6.2}$$

follow from $S^i(t_1) = 0, i = 1, \cdots, p-1$, and p even. If u^0 is discontinuous at t_1 we obtain $(a \pm bK)(t_1) \leqq 0$ from (6.2), which contradicts the assumption $|a(t_1)| <$

$|b(t_1)|K(t_1)$. If u^0 is continuous at t_1 then ϕ^0 does not change sign in a neighborhood of t_1. As $\phi^{0(i)}(t_1)=0$, $i=0,\cdots,p-2$, and p is even, we find $\phi^{0(p-1)}(t_1)=0$ in this case, which contradicts the assumption that ϕ^0 does not have a zero of order p at t_1.

(ii). If (P_α) is strongly stable at α_0, then (i) implies that (2.7), (4.9) cannot be fulfilled at any point $t_1\in[0, T]$ for $\alpha\in[\alpha_1, \alpha_0)$ with suitable $\alpha_1<\alpha_0$. This proves the first statement in (ii), and the second statement is a direct consequence of the assumed shape of ϕ^0.

The meaning of Lemma 6.4 is the following: if p is *even* and if (P_α) is strongly stable at α_0, then boundary arcs can only occur after a *transitional phase* in $[\alpha_1, \alpha_0)$ where the constrained extremal has only contact points. However, there are examples for p even where this transitional phase does not occur; compare Example 2, § 7. In this case, (P_α) cannot be strongly stable at α_0. The next lemma is concerned with such a situation.

LEMMA 6.5. *Let p be even, $p\leqq 2q$, and let the unconstrained extremal be unique. Suppose that ϕ^0 has simple zeros in* (0, T) *and that $\phi^0(0)\neq 0$, $\phi^0(T)\neq 0$, $\phi^0(t_1)\neq 0$ for every point t_1 in* (6.1). *Further assume that the switching points of the nonsingular control u^0 are already uniquely determined by the end-conditions* (2.3). *Then (P_α) is not strongly stable at α_0.*

Proof. Suppose that (P_α) is strongly stable at α_0. Then Lemma 6.4(ii) shows that the constrained extremal for $\alpha\in[\alpha_1, \alpha_0)$, $\alpha_1<\alpha_0$ suitable, has only contact points and has the same number of switching points as the unconstrained extremal. But then the constrained trajectory for $\alpha\in[\alpha_1, \alpha_0)$ must coincide with the unconstrained trajectory as we have assumed that the switching points of u^0 are already uniquely determined by (2.3). This is a contradiction and hence (P_α) cannot be strongly stable at α_0.

An application of Lemma 6.5 is provided by Example 2, § 7, for $p=2$. In this example the multipliers $\lambda(t;\alpha_0)$ corresponding to the trivial constraint $S(x)\leqq\alpha_0$ are *not* unique whereas the unconstrained multiplier $\lambda^0(t)$ is unique.

7. Numerical examples.

Example 1. In this example the inequality constraint $x_1\geqq\alpha$ is used instead of $x_1\leqq\alpha$. This explains the different sign for α_0. The problem is to minimize the end-time

$$J(u)=T \tag{7.1}$$

subject to

$$\begin{aligned} &\dot{x}_1=x_2, && x_1(0)=3, \quad x_1(T)=0,\\ &\dot{x}_2=-x_1+u, && x_2(0)=1, \quad x_2(T)=0,\\ &|u(t)|\leqq 1, && 0\leqq t\leqq T, \end{aligned} \tag{7.2}$$

and the state inequality constraint of order $p=2$

$$S(x)=-x_1\leqq-\alpha, \quad \text{i.e.,} \quad x_1\geqq\alpha. \tag{7.3}$$

The unconstrained optimal trajectory $x^0(t)$ is the well-known solution of the undamped oscillator problem (see [9, p. 111]) and consists of the arc x_0AC0

shown in Fig. 1 where $A = (3, -1)$ and $C = (-1, 1)$ are the switching points of the unconstrained optimal control $u^0(t)$.

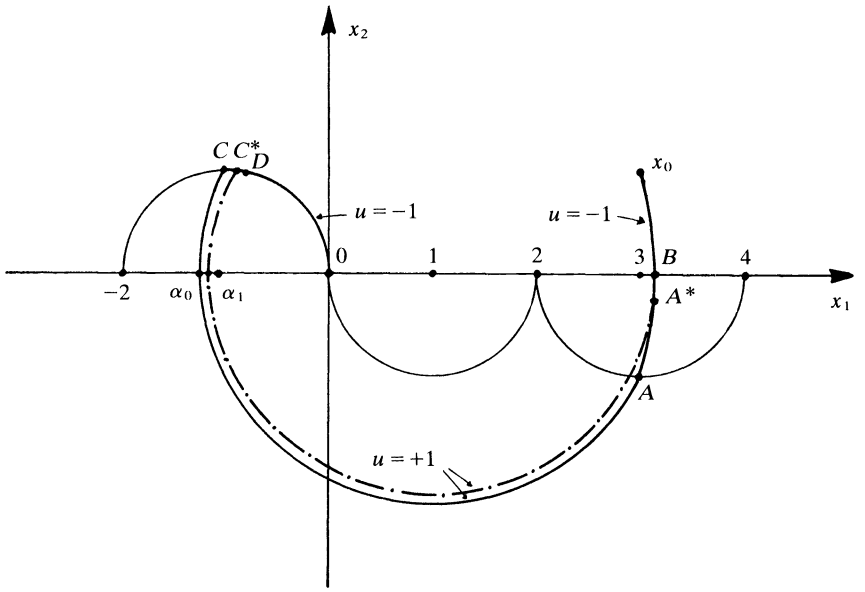


FIG. 1. *Geometrical construction of extremals touching the boundary*

One gets $\alpha_0 = \min\{x_1^0(t) \mid t \in [0, T]\} = x_1^0(t_1) = 1 - \sqrt{5} = -1.2361$. The control $u^0(t)$ is continuous at t_1 and $\phi^0(t_1) \neq 0$. The multiplier $\lambda(t; \alpha_0)$ with $\eta_0 = 1$ in (3.2) is unique and is equal to the unconstrained multiplier $\lambda^0(t)$. Since boundary arcs for $\alpha_0 < \alpha$ would be restricted to the stationary point $(\alpha, 0)$ they cannot form part of a constrained extremal because of the functional (7.1). Therefore, any constrained extremal has only contact points. It is not difficult to compute the constrained extremals and to verify that (P_α) is strongly stable at α_0. Hence Lemma 6.4(ii) applies. Figure 1 indicates how the constrained extremals touching the boundary are constructed with the same number of switching points as the unconstrained extremals: one chooses first the switching point A^* lying between $A = (3, -1)$ and $B = (\sqrt{17} - 1, 0)$ on the circle through x_0 with center at $(-1, 0)$. This results in a second switching point C^* lying between $C = (-1, 1)$ and $D = (\sqrt{17} - 5, 1 - (\sqrt{17} - 4)^2)$ on the circle through $(0, 0)$ with center at $(-1, 0)$. If A^* moves from A to B, then the corresponding contact point $(\alpha, 0)$ moves from $(\alpha_0, 0)$ to $(\alpha_1, 0)$ with $\alpha_0 = 1 - \sqrt{5} = -1.2361$ and $\alpha_1 = 3 - \sqrt{17} = -1.1231$. This geometrical construction shows that the constrained extremal exists for parameters $\alpha \in (\alpha_0, \alpha_1]$. Moreover, (P_α) is strongly stable at $\alpha \in [\alpha_0, \alpha_1)$.

It is easily shown by elementary calculations that the relations (4.6), (4.7) for the switching function at the contact point t_1 hold in this example with $\nu_1 > 0$ for $\alpha_0 < \alpha < \alpha_1$. In the limit one obtains formally $\phi(t_1^-) = 0$ for $\alpha = \alpha_1$. Thus for $\alpha = \alpha_1$ the entry condition (4.9) for a boundary arc is satisfied. But no constrained extremal exists for $\alpha_1 < \alpha$ as mentioned before.

Example 2. The following time-optimal control problem of a nuclear reactor is taken from Hassan et al. [10] where some numerical results are given. A detailed numerical solution was obtained by Heidemann [11] and the author by using numerical techniques similar to those in Maurer and Gillessen [12]. The encountered boundary value problems were solved with the method of multiple shooting developed by Bulirsch, Stoer and Deuflhard [13]. The numerical calculations were performed on the computer terminal of the Mathematisches Institut der Universität Köln. The following presentation stresses the structure of the solution in relation to the junction theorems of § 5 more than the numerical aspect which is elaborated in [11].

The problem is to minimize the end-time

$$J(u) = T \tag{7.4}$$

subject to

$$\begin{array}{lll} \dot{x}_1 = k_1(x_3-1)x_1 + k_2 x_2, & x_1(0) = n_0, & x_1(T) = n_T, \\ \dot{x}_2 = k_1 x_1 - k_2 x_2, & x_2(0) = n_0 k_1/k_2, & x_2(T) = n_T k_1/k_2, \\ \dot{x}_3 = u, & x_3(0) = 0, & x_3(T) = 0, \\ |u(t)| \leqq 0.2, \quad 0 \leqq t \leqq T, & & \end{array} \tag{7.5}$$

where x_1 is neutron density, x_2 is delayed neutron concentration, x_3 is reactivity, $k_1 = 5.0$, $k_2 = 0.1$, $n_0 = 0.04$, $n_T = 0.06$. The process is subject to either

(I) the state inequality constraint of order $p = 1$

$$S(x) = x_3 \leqq \rho \tag{7.6}$$

or

(II) the state inequality constraint of order $p = 2$

$$S(x) = x_1 \leqq \alpha. \tag{7.7}$$

In the sequel, the adjoint variables are calculated for $\eta_0 = 1$ in (3.2).

Unconstrained extremal. The optimal control is

$$u^0(t) = \begin{cases} 0.2, & t \in [0, t^*), \\ -0.2, & t \in (t^*, t^{**}), \\ 0.2, & t \in (t^{**}, T], \end{cases} \tag{7.8}$$

with $T = 7.047806$, $t^* = 0.4798784 \cdot T$, $t^{**} = 0.9798784 \cdot T$. The multiplier $\lambda^0(t)$ is determined by the initial value $\lambda^0(0) = -(2.970144, 2.845469, 5.0)^T$. The three parameters t^*, t^{**}, T in (7.8) are uniquely determined by the three conditions for the final state $x(T)$. The switching function $\phi^0(t) = \lambda_3^0(t)$ is shown in Fig. 2 (its negative has been chosen because of (3.6)) with respect to normalized time t/T.

(I) *Inequality constraint* (7.6) *of order* $p = 1$. For the unconstrained extremal $x^0(t)$ one obtains $\rho_0 = \max\{x_3^0(t) \mid t \in [0, T]\} = x_3^0(t^*) = 0.2 \cdot t^* = 0.6764179$. The point t^* is a switching point of the control $u^0(t)$ and therefore $\phi^0(t^*) = 0$. The numerical results in [11] suggest that the problem (P_ρ) is strongly stable at $\rho \in (0, \rho_0]$. Since $S^1(x, u) = u$, i.e., $a(t) = 0$ and $b(t) = 1$, the boundary control is

$u_b = 0$. The constrained extremal contains one boundary arc and no contact points for $\rho \in (0, \rho_0)$ (cf. Lemma 6.3(ii)) and the optimal control is

$$u(t) = \begin{cases} 0.2, & t \in [0, 5\rho), \\ 0.0, & t \in [5\rho, t_2], \\ -0.2, & t \in (t_2, t^{**}), \\ 0.2, & t \in (t^{**}, T]. \end{cases} \tag{7.9}$$

The three parameters t_2, t^{**}, T are uniquely determined by the final state $x(T)$. The numerical results also demonstrate that Theorem 5.3(i) holds with $p = 1, r = 0$ and $m = 0$ and thus $\dot{\phi}(t_1^-) \neq 0$, $t_1 = 5\rho$, $\dot{\phi}(t_2^+) \neq 0$ for $\rho \in (0, \rho_0)$.

(II) *Inequality constraint* (7.7) *of order* $p = 2$. We get $\alpha_0 = \max\{x_1^0(t) \mid t \in [0, T]\} = x_1^0(t_1) = 0.1213660$, $t_1 = 0.5429932 \cdot T$ where $T = 7.047806$. The point t_1 is not a switching point of $u^0(t)$ and $\phi^0(t_1) \neq 0$; see Fig. 2. The assumptions of Lemma 6.5 are satisfied here and hence (P_α) is not strongly stable at α_0.

It will turn out below that the multipliers $\lambda(t; \alpha_0)$ are not unique whereas $\lambda(t; \alpha)$ is unique for $\eta_T < \alpha < \alpha_0$.

The constrained extremal contains one boundary arc for $n_T < \alpha < \alpha_0$. Here the contact point t_1 for $\alpha = \alpha_0$ splits up into the boundary arc for $\alpha < \alpha_0$. The optimal control is

$$u(t) = \begin{cases} 0.2, & t \in [0, t^*), \\ -0.2, & t \in (t^*, t_1), \\ -k_2 x_3, & t \in [t_1, t_2], \\ -0.2, & t \in (t_2, t^{**}), \\ 0.2, & t \in (t^{**}, T). \end{cases} \tag{7.10}$$

The boundary control $u = -k_2 x_3$ is obtained from $S^2(x, u) = 0$. The five parameters t^*, t_1, t_2, t^{**}, T are uniquely determined by the three conditions for $x(T)$ and the two entry conditions $S(t_1) = \alpha$, $S^1(t_1) = 0$. The multipliers $\lambda(t; \alpha)$ are uniquely determined by the necessary conditions and are continuous in α. Hence (P_α) is strongly stable at $\alpha \in (n_T, \alpha_0)$. One finds $\nu_1 = \lambda_1(t_1^-) > 0$ and $\eta(t) = -k_1 \lambda_2(t) > 0$, $\lambda_1(t) = \lambda_3(t) \equiv 0$ for $t \in [t_1, t_2]$. The numerical values for $-k_2 x_3$ are such that the control $u(t)$ in (7.10) is always discontinuous at t_1 and t_2. Thus the necessary conditions of Theorem 5.1 for a boundary arc are satisfied with $\nu_1 > 0, p = 2, r = 0$; i.e., $p + r$ is even. For reasons of numerical comparison the numerical values are given for $\alpha = 0.105$: $T = 7.344445$, $t^* = 0.4253868 \cdot T$, $t_1 = 0.4793580 \cdot T$, $t_2 = 0.6285059 \cdot T$, $t^{**} = 0.9806927 \cdot T$, $\lambda(0) = -(4.052238, 3.882192, 5.0)^T$, $\nu_1 = 3.83874$.

Performing the limit $\alpha \uparrow \alpha_0$, i.e., $t_1 - t_2 \to 0$, in (7.10) we get the dashed switching function shown in Fig. 2 with respect to normalized time t/T. This switching function is admissible for the unconstrained control (7.8) in the sense of the minimum principle for the trivial constraint $x_1 \leqq \alpha_0$. However, this switching function is *not* an admissible switching function for the unconstrained problem (2.1)–(2.4).

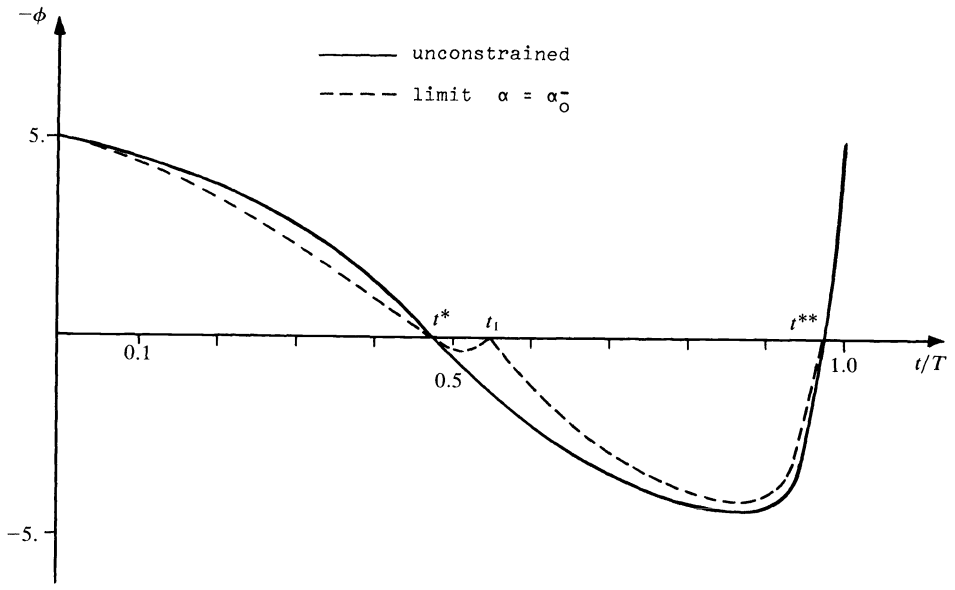


FIG. 2. *Unconstrained and the limiting constrained switching function for* $\alpha = \alpha_0$

The numerical values for the multiplier $\bar{\lambda}(t)$ corresponding to the dashed switching function are fixed by the initial value $\bar{\lambda}(0) = -(3.376317, 3.234608, 5.0)^T$ and the jump $\bar{\nu}_1 = 9.056300$ of $\bar{\lambda}(t)$ at $t_1 = 0.5429932 \cdot T$. Now we want to characterize *all* multipliers $\lambda(t; \alpha_0)$ satisfying the necessary conditions with $\eta_0 = 1$ in (3.2). The multipliers can be parametrized by the interval $[0, \bar{\nu}_1]$ in the following way: for given $\nu_1 \in [0, \bar{\nu}_1]$ there exists $\lambda(t; \alpha_0)$ satisfying the necessary conditions with prescribed jump ν_1 at t_1. Thus the range of $\lambda(t; \alpha_0)$ sweeps from the unconstrained multiplier $\lambda^0(t)$, for which $\nu_1 = 0$ and $\phi^0(t_1) \neq 0$, to the special constrained multiplier $\bar{\lambda}(t)$ with jump $\bar{\nu}_1$ at t_1 for which the entry condition $\phi(t_1) = 0$ for a boundary arc is satisfied. Then for $\alpha < \alpha_0$ the entry condition $\phi(t_1^-) = 0$ can be met continuously with multipliers $\lambda(t; \alpha)$ which are continuous in α. In particular we have $\nu_1(\alpha) + \nu_2(\alpha) \to \bar{\nu}_1$ for $\alpha \uparrow \alpha_0$ where $\nu_i(\alpha)$ are the jumps of $\lambda(t; \alpha)$ at the points t_i, $i = 1, 2$, in (7.10).

The limiting case corresponding to $\bar{\lambda}(t)$ and $\bar{\nu}_1$ was not computed by performing the limit $t_1 - t_2 \to 0$ in (7.10)—this leads to a numerically unstable boundary value problem—but was computed by translating the above considerations numerically. Regard ν_1 as a *homotopy-parameter* and calculate the multiplier satisfying the necessary conditions with given jump ν_1 at t_1 starting the homotopy at $\nu_1 = 0$. For instance one obtains $\lambda(0) = -(3.104701, 2.974376, 5.0)^T$ for $\nu_1 = 3$. Then $\bar{\nu}_1$ is the homotopy-parameter (treated numerically as a *free* parameter) which produces the entry condition $\phi(t_1) = 0$.

Let us denote the value of the functional (7.4) subject to the state constraint $x_1 \leqq \alpha$ by $T(\alpha)$. Then the numerical results suggest that $T(\cdot)$ is not differentiable at α_0 but is convex in a neighborhood of α_0. Furthermore, it turns out that the

right-sided derivative is $T'(\alpha_0^+)=0$ and the left-sided derivative is $T'(\alpha_0^-)=-\bar{\nu}_1=-9.056300$. Thus we have

$$-\partial T(\alpha_0)=-[T'(\alpha_0^-),\ T'(\alpha_0^+)]=[0,\ \bar{\nu}_1] \tag{7.11}$$

where $\partial T(\alpha_0)$ is the subgradient of $T(\alpha)$ at α_0. This is a special result of the perturbation theory developed in [14].

Example 3. Consider the problem of minimizing

$$J(u)=\frac{1}{2}\int_0^3 x^2\,dt \tag{7.12}$$

subject to

$$\dot{x}=u,\quad |u|\leqq 1,\quad x(0)=x(3)=1 \tag{7.13}$$

and the state inequality constraint of order $p=1$

$$S(t,x)=\tfrac{1}{2}t-x\leqq\alpha,\qquad \alpha\in\mathbb{R}. \tag{7.14}$$

The *unconstrained* extremal is

$$(x^0(t),u^0(t),\lambda^0(t))=\begin{cases}(1-t,-1,\tfrac{1}{2}(t-1)^2), & t\in[0,1),\\ (0,0,0), & t\in[1,2],\\ (t-2,1,-\tfrac{1}{2}(t-2)^2), & t\in(2,3],\end{cases} \tag{7.15}$$

and has a singular arc of order $q=1$ in $[1,2]$. Here (2.6) and (3.7) give $S^1=\frac{1}{2}-u$, $\phi^{(2)}=-u$. Thus $a=\frac{1}{2}$, $b=-1$, $A=0$, $B=-1$ and $aB-bA=-\frac{1}{2}$. We have

$$\alpha_0=\max\{S(t,x^0(t))\,|\,t\in[0,3]\}=S(2,x^0(2))=1$$

where the point $t_1=2$ is a junction point between a singular and a nonsingular arc; cf. Theorem 5.6(i).

The *constrained* extremal exists for $\alpha\in[\frac{1}{2},1]$. It is clear from (7.15) that the constrained extremal is

$$(x(t),u(t))=\begin{cases}(1-t,-1), & t\in[0,1),\\ (0,0), & t\in[1,2\alpha),\\ (\frac{1}{2}t-\alpha,\frac{1}{2}), & t\in[2\alpha,4-2\alpha],\\ (t-2,1), & t\in(4-2\alpha,3].\end{cases} \tag{7.16}$$

The optimal trajectory has a singular arc in $[1,2\alpha]$ and a boundary arc in $[2\alpha,4-2\alpha]$. The multiplier $\lambda(t)$ satisfies $\dot{\lambda}=-x+\eta$ with $\eta(t)=0$ for $t\notin[2\alpha,4-2\alpha]$ and $\eta(t)\geqq 0$ for $t\in[2\alpha,4-2\alpha]$. The switching function is $\phi(t)=\lambda(t)$ and hence $\lambda(t)=0$ for $t\in[1,4-2\alpha]$. These conditions and the minimum principle (3.6) determine uniquely $\lambda(t)$, and we get

$$\lambda(t;\alpha)=\begin{cases}\frac{1}{2}(t-1)^2, & t\in[0,1],\\ 0, & t\in[1,4-2\alpha],\\ \frac{1}{2}\{-(t-2)^2+(2-2\alpha)^2\}, & t\in[4-2\alpha,3],\end{cases} \tag{7.17}$$

and $\eta(t)=x(t)=\frac{1}{2}t-\alpha\geqq 0$ for $t\in[2\alpha,4-2\alpha]$. The multiplier $\lambda(t;\alpha)$ depends continuously on α and hence (P_α) is strongly stable at $\alpha\in(\frac{1}{2},1]$. For $\alpha=\frac{1}{2}$ the

constrained extremal is formally contained in (7.16), (7.17) where the singular arc in $[1, 2\alpha]$ and the nonsingular arc in $[4-2\alpha, 3]$ degenerates.

Let us check the various junction theorems for $\alpha \in (\frac{1}{2}, 1)$ first. At the junction $t_1 = 1$ of the nonsingular and singular arc Theorem 1 of [1] holds with $q = 1$, $r = 0$; i.e., $q+r$ is odd. At the junction $t_2 = 2\alpha$ of the singular interior arc and the boundary arc we have $q = 1$, $r = 0$ and $\eta(t_2) = 0$, $\dot{\eta}(t_2^+) = \frac{1}{2}$. Thus Theorem 5.4 is verified with $q+r$ odd and also (5.13) holds with $j = 2q+r-p = 1$. Finally, at the junction $t_3 = 4-2\alpha$ of the boundary arc and the nonsingular interior arc we get $\phi(t_3) = 0$, $\dot{\phi}(t_3^+) = t_3 - 2 = 2-2\alpha \neq 0$. Then Theorem 5.3(i) applies with $p = 1$, $m = 0$, $r = 0$, $q = 1$; i.e., $p+m < 2q+r$ and $p+r+m = 1$ is odd.

Now we consider $\alpha = \frac{1}{2}$. Here $t_1 = 1$ is a junction of a nonsingular interior arc and a boundary arc. We obtain $\phi(t_1) = \dot{\phi}(t_1) = 0$, $\phi^{(2)}(t_1^-) = 1$ and hence $m = 1$ in Theorem 5.3. Since we have $p = 1 < 2q+r = p+m = 2$, part (ii) of Theorem 5.3 applies. There the integer j becomes $j = 2$ as $\lambda(t_1)\varphi_i(t_1) = 0$, $i = 0, 1$, and $\lambda(t_1)\varphi_2(t_1^+) = -u(t_1^+) = -\frac{1}{2}$. This follows also from (5.7) by virtue of $\eta(t_1) = 0$, $\dot{\eta}(t_1^+) = \frac{1}{2}$. Hence $-\operatorname{sgn}(\lambda(t_1)\varphi_2(t_1^-)\lambda(t_1)\varphi_2(t_1^+)) = 1 = (-1)^{p+r+m}$ as $p+r+m = 2$ in Theorem 5.3(ii).

Acknowledgment. I would like to thank Mr. U. Heidemann for making available to me his numerical results for Example 2, § 7. Further, I am grateful to the referees whose comments helped to improve the paper.

REFERENCES

[1] J. P. McDanell and W. F. Powers, *Necessary conditions for joining singular and nonsingular subarcs*, this Journal, 9 (1971), pp. 161–173.

[2] D. H. Jacobson, M. M. Lele and J. L. Speyer, *New necessary conditions of optimality for control problems with state-variable inequality constraints*, J. Math. Anal. Appl., 35 (1971), pp. 255–284.

[3] W. E. Hamilton Jr., *On nonexistence of boundary arcs in control problems with bounded state variables*, IEEE Trans. Automatic Control, AC-17, (1972), pp. 338–343.

[4] D. O. Norris, *Nonlinear programming applied to state-constrained optimization problems*, J. Math. Anal. Appl., 43 (1973), pp. 261–272.

[5] I. V. Girsanov, *Lectures on mathematical theory of extremum problems*, Lecture Notes in Economics and Mathematical Systems, vol. 67, Springer-Verlag, Berlin-Heidelberg-New York, 1972.

[6] H. Maurer, *Optimale Steuerprozesse mit Zustandsbeschränkungen*, Habilitationsschrift, Mathematisches Institut der Universität Würzburg, Würzburg, W. Germany, March 1976.

[7] H. J. Kelley, R. E. Kopp and H. G. Moyer, *Singular extremals*, Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967, Chap. 3.

[8] H. Maurer, *Numerical solution of singular control problems using multiple shooting techniques*, J. Optimization Theory Appl., 18 (1976), pp. 235–257.

[9] E. G. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[10] M. A. Hassan, M. A. R. Ghonaimy and N. R. Abdel Malek, *Computational solution of the nuclear reactor minimum time start-up problem with state-space constraints*, 2nd IFAC Symp. on Multivariable Technical Control Systems, Düsseldorf, Oct. 11–13, North-Holland, Amsterdam, 1971.

[11] U. Heidemann, *Numerische Berechnung optimaler Steuerungen mit Hilfe der Mehrzielmethode bei Beschränkungen im Zustandsraum mit Beispielen aus der Physik, Chemie und Raumfahrt*, Diploma thesis, Mathematisches Institut der Universität Köln, Köln, W. Germany, 1974.

[12] H. MAURER AND W. GILLESSEN, *Application of multiple shooting to the numerical solution of optimal control problems with bounded state variables*, Computing, 15 (1975), pp. 105–126.

[13] R. BULIRSCH, J. STOER AND P. DEUFLHARD, *Numerical Solution of Nonlinear Two Point Boundary Value Problems I*, Numerische Mathematik, Handbook Series Approximation, to appear.

[14] H. MAURER, *A sensitivity result for nonlinear infinite programming problems II: Applications to optimal control problems*, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

CONTROLLABILITY AND OBSERVABILITY OF PARABOLIC SYSTEMS—AN ADDENDUM TO TWO RECENT PAPERS OF Y. SAKAWA*

JEAN-CLAUDE E. MARTIN†

Abstract. In two recent articles Sakawa has established necessary and sufficient conditions to insure the controllability (observability) of parabolic systems of finite multiplicity m. One of these conditions is that there must be at least m controllers (sensors). In this short note it is shown that the results of Sakawa can be extended to the case where only one discrete scanning controller (sensor) is used. In some special cases Sakawa's results can be utilized to establish the observability of a system when only one time-varying sensor is used.

Introduction. In two recent articles [1], [2], Sakawa has considered the controllability and observability problems for partial differential equations of parabolic type. For the controllability problem [1], Sakawa used controls of the form $\sum_{i=1}^{n} g_i(x)f_i(t)$ distributed over the system's spatial domain D or its boundary. In [2], the system's state $u(t, x)$ is observed with either n spatial averaging sensors whose outputs are $y_i(t)=\int_D w_i(x)u(t, x)\,dx$, $i=1, \cdots, n$, or n ideal pointwise sensors, i.e., $y_i(t)=u(t, x_i)$, $i=1, \cdots, n$. Sakawa established that for parabolic partial differential equations involving a time-invariant elliptic operator of finite multiplicity m there must be at least m suitable controls (sensors) to ensure the controllability (observability) of the system.

For economical reasons and physical simplicity it is often desirable in practice to use only one controller (sensor). The goal of this short note is to show that using the results of Sakawa it is possible to control (observe) the state of the system with only one controller (sensor). Two cases will be considered: (i) the discrete scanning control (observation) where the controller (sensor) can be moved discretely along a given trajectory in the interior of the system's spatial domain or on the boundary, (ii) for some special cases, the time-varying sensor.

Since the two problems—controllability and observability—have the same development, only the observability problem will be described in detail and, to facilitate reference to Sakawa [1] and [2], we shall use the same notations.

Problem statement. As in [2] let D be a bounded domain of an r-dimensional Euclidean space. The boundary of D is denoted by S and is supposed to be smooth enough.

We consider the linear parabolic partial differential equation

$$\partial u(t, x)/\partial t=\Delta u(t, x)-q(x)u(t, x) \quad \text{on } (0, T]\times D, \tag{1}$$

with the boundary condition

$$\begin{aligned} &\alpha(\xi)u(t, \xi)+(1-\alpha(\xi))\,\partial u(t, \xi)/\partial n=0,\\ &0\leqq\alpha(\xi)\leqq 1, \qquad \xi\in S, \end{aligned} \tag{2}$$

* Received by the editors June 16, 1975, and in revised form November 26, 1975.

† Department of Mathematics, Ecole Polytechnique Federale de Lausanne, CH-1007 Lausanne, Switzerland.

and the initial condition given by

$$\lim_{t\to 0} u(t, x) = u_0(x) \in L_2(D). \tag{3}$$

Under suitable conditions [2] there exists a unique solution to the initial-boundary value problem described by (1), (2) and (3) and the solution is given by

$$u(t, x) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \sum_{j=1}^{m_i} u_{ij}\Phi_{ij}(x),$$

where $\{\Phi_{ij}(x);\ j = 1, \cdots, m_i,\ i = 1, 2, \cdots\}$ is a complete orthonormal system in $L_2(D)$. Each Φ_{ij} satisfies

$$\Delta\Phi_{ij}(x) - q(x)\Phi_{ij}(x) = -\lambda_i\Phi_{ij}(x), \qquad x \in D,$$

with boundary conditions (2).

The coefficients u_{ij} are given by $u_{ij} = \int_D u_0(x)\Phi_{ij}(x)\,dx$ and the system is said to be of finite multiplicity m if $\sup(m_i) = m < \infty$.

Discrete scanning sensor. As in [2] we consider an observation of the solution over a finite time interval $(0, T]$ using either a spatial averaging sensor (measurement of type 1) or an ideal point-sensor (measurement of type 2).

Let χ_k; $k = 1, \cdots, N$, be subintervals of $(0, T]$ satisfying

$$\chi_k \neq \varnothing,$$
$$\chi_k \cap \chi_n \neq \varnothing, \quad k \neq n,$$
$$\bigcup_{k=1}^{N} \chi_k \subset (0, T].$$

The observation of the solution will be defined as

$$y_k(t) = \begin{cases} \displaystyle\int_D w(x - x_k)u(t, x)\,dx, & t \in \chi_k, \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

where $w(x - x_k)$ is the weighting function of the sensor "centered" at the point $x_k \in D$. $w(\cdot) \in L_2(D)$ in the case of a spatial averaging sensor and $w(x) = \delta(x)$—the Dirac measure—for ideal point-sensor.

Let

$$W_{ij}^k = \begin{cases} \displaystyle\int_D w(x - x_k)\Phi_{ij}(x)\,dx & \text{—measurement of type 1,} \\ \Phi_{ij}(x_k) & \text{—measurement of type 2,} \end{cases}$$

and as in [2] let us define the $N \times m_i$ matrices W_i by

$$W_i = \begin{bmatrix} w_{i1}^1 & \cdots & w_{im_i}^1 \\ \vdots & & \vdots \\ w_{i1}^N & \cdots & w_{im_i}^N \end{bmatrix}.$$

Then, using the same proof that is in [2, Thm. 1, p. 18] one can show that the system is observable in any finite time if and only if $N \geqq m$ and rank $W_i = m_i$, $i = 1, 2, \cdots$, which means that we must make at least m measurements at m different points. Restrictions concerning ideal point-sensors [2, p. 21] still hold in this case.

Time-varying sensor. In this section we consider elliptic operators whose eigenvalues $\{\lambda_i; i = 1, 2, \cdots\}$ satisfy the following "separation condition":

$$\begin{aligned} &\lambda_1 \geqq \rho > 0, \\ &\lambda_{i+1} - \lambda_i \geqq \rho, \qquad i = 1, 2, \cdots. \end{aligned} \tag{5}$$

Let the weighting function of the time-varying sensor $w(t, x) \in L_2(0, T \times D)$ be defined by

$$w(t, x) = \sum_{k=1}^{N} C_k(x)\, e^{-\rho_k t},$$

where $0 < \rho_k < \rho$, $k = 1, \cdots, N$.

The observation is defined by

$$y(t) = \int_D w(t, x) u(t, x)\, dx$$

and can be written as

$$y(t) = \sum_{i=1}^{\infty} \sum_{k=1}^{N} e^{-(\lambda_i + \rho_k)t} \sum_{j=1}^{m_i} u_{ij} w_{ij}^k,$$

where $w_{ij}^k = \int_D C_k(x) \Phi_{ij}(x)\, dx$.

Since $\rho_k < \rho$, $k = 1, \cdots, N$, and $\{\lambda_i\}$ satisfies the separation condition (5), we have

$$(\lambda_i + \rho_k) \neq (\lambda_j + \rho_n) \quad \text{for all } k, n \leqq N \quad \text{and all } i, j = 1, 2, \cdots. \tag{6}$$

Thus, Theorem 1 of [2, p. 18] holds and the system can be observed with only one static time-varying sensor.

Remark. In the case where the eigenvalues of the elliptic operator do not satisfy the separation condition (5), Theorem 1 of [2, p. 18] can still be applied to establish the observability of the system using only one time-varying sensor if N values ρ_k, $k = 1, \cdots, N$, can be found such that inequality (6) holds.

Controllability problem. Let us consider the parabolic partial differential equation

$$\partial u(t, x)/\partial t = \Delta u(t, x) - q(x) u(t, x) + g(x - x_k) f(t) \tag{7}$$

with boundary and initial conditions given by equations (2) and (3). Assume that the time interval $(0, t]$ can be covered by N nonvoid, nonoverlapping time subintervals χ_k;

$$(0, t] = \bigcup_{k=1}^{N} \chi_k.$$

The control's spatial function $g_k(x) \equiv g(x - x_k)$ is "centered" at the point x_k for $t \in \chi_k$ and, as in [1], $g_k(x)$; $k = 1, \cdots, N$, are Hölder continuous on the compact domain $\bar{D}$. It can also be assumed that for some values of k, $g_k \equiv 0$ which corresponds to the time needed to switch the controller from one location to another one.

The solution of the parabolic equation can be written as

$$u(t) = U_t u_0 + \sum_{k=1}^{N} \int_{\chi_k} U_{t-\tau} g_k f(\tau)\, d\tau,$$

where U_t is the fundamental solution at time t of the system defined by (7) and (2).

Using Theorem 3 of [1, p. 393] one can immediately see that system (7), (2) and (3) is null controllable if m locations x_k; $k = 1, \cdots, m$, can be found such that the matrices G_i (see [1, p. 393]) satisfy rank $G_i = m_i$ for all $i = 1, 2, \cdots$.

Conclusion. This short note gives two examples where the results of Sakawa [1] and [2] can be applied to prove the observability (and controllability) of parabolic partial differential equations of finite multiplicity m using only one sensor (controller).

The case of a discrete scanning sensor (controller) can be easily implemented in practice. The time-varying sensors which applied to more restrictive cases will also be much more difficult to realize technically.

REFERENCES

[1] Y. SAKAWA, *Controllability for partial differential equations of parabolic type*, this Journal, 12 (1974), pp. 389–400.

[2] ———, *Observability and related problems for partial differential equations of parabolic type*, this Journal, 13 (1975), pp. 14–27.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

THE SEPARATION PRINCIPLE FOR STOCHASTIC EVOLUTION EQUATIONS*

RUTH F. CURTAIN AND AKIRA ICHIKAWA†

Abstract. The separation principle is proved for a general class of linear infinite dimensional systems. The dynamical system is modeled as an abstract evolution equation, which includes linear ordinary equations, classes of linear partial differential equations and linear delay equations. The noise process in the control system is modeled using a stochastic integral with respect to a class of Hilbert space valued Gaussian stochastic processes, which includes the Wiener process as a special case. The observation process is finite dimensional and is corrupted by Gaussian type white noise, which is modeled using the Wiener integral. The cost functional to be minimized is quadratic.

Introduction. The separation principle is a classic theorem of finite dimensional stochastic control theory (see [14] and [1]) and it is natural to ask whether the principle holds for more general systems. In [5] Brooks gives an abstract version of the separation principle for a finite dimensional system with general disturbance and a weighted quadratic cost functional. As he uses a functional analytic approach he obtains the optimal control in a nonfeedback form, which is not too attractive from the applications point of view. Balakrishnan in [1] and [2] uses a similar approach, and so obtains the control in a similar form. His results are obtained for time invariant linear systems described by semigroups and so apply linear delay systems and distributed systems. It is worth noting that he uses a nonstandard model for the noise disturbances which differs from all the other authors. By imbedding the feedback stochastic control problem in a stochastic open loop problem, Lindquist in [13] develops a technique which embeds the separation principle to system disturbed by colored measurement noise and systems with time delays. In [4] Bensoussan and Viot extend the earlier work of [3] for systems described by parabolic partial differential equations of Lions' type. They define a fixed Hilbert space, which the class of feedback controls are dense in and then use a variational approach to obtain the separation principle for a general convex cost function and constrained controls.

In this paper we consider a very general class of systems which includes all the aforementioned except for an example in Lindquist [13] where he considers delay equations with delays in the control. With that exception, it is possible to model linear delay equations and distributed systems described by linear integro-differential equations in terms of an evolution operator (see Curtain and Pritchard [6]). The noise disturbance in the signal is modeled using a stochastic integral with respect to an orthogonal increment type process introduced by Curtain in [10], and which includes the usual Gaussian white noise, colored noise and also Poisson type noise. In § 2.4 the optimal control problem for complete observation is solved for this general noise disturbance. For the incomplete observation case a separation principle is obtained under the extra assumption that all noise disturbances are Gaussian. This stochastic evolution model has already been successfully used by Curtain in [8] and [9] to solve the filtering and smoothing problem,

* Received by the editors December 29, 1975, and in revised form July 28, 1976.

† Control Theory Centre, University of Warwick, Coventry, Warwickshire CV4 7AL, England.
The Control Theory Centre is supported by the Science Research Council.

and in fact this paper is a natural extension of this work. Using the results from [8] and [9], we are able to use an approach similar to that of Balakrishnan in [1] and [2] to prove the separation principle for the class of linear feedback controls. We are also able to embed it to the admissible class of Bensoussan and Viot in [4].

1. Mathematical preliminaries.

1.1. Evolution operators. We summarize the theory of evolution operators developed in [6] and [9] for the modeling of linear infinite dimensional systems appropriate for control 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 $\Delta(T)=\{(t, s): 0\leqq s\leqq t\leqq T\}$. 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 $[s, T]$.

Mild evolution operators are closed under perturbations by operators $D\in\mathcal{B}_\infty(T; \mathcal{L}(H))$ (the class of $\mathcal{L}(H)$-valued functions which are strongly measurable on T with $\operatorname{ess}_T \sup\|D(t)\|<\infty$) in the following sense:

$$\mathcal{U}_D(t, s)x=\mathcal{U}(t, s)x+\int_s^t \mathcal{U}(t, r)D(r)\mathcal{U}_D(r, s)x\, dr, \qquad x\in H. \tag{1.1}$$

If $D\in\mathcal{B}_\infty(T; \mathcal{L}(H))$, (1.1) has a unique solution which is also a mild evolution operator; $\mathcal{U}_D(t, s)$ is called the perturbation of $\mathcal{U}(t, s)$ corresponding to D.

1.2. Abstract probability theory. We recall the following definitions and results on Banach-space-valued random variables and conditional expectations from [9] and [4]. Let $(\Omega, \mathcal{P}, \mu)$ be a complete probability space, X, Y real separable Banach spaces, H, K real separable Hilbert spaces and $T=[0, T]$ a real finite time interval.

DEFINITION 1.2. An *X-valued random variable* is a map $u: \Omega\to X$ which is measurable with respect to μ-measure. If $u\in L_1(\Omega, \mu; X)$, we define its *expectation*

$$E\{u\}=\int_\Omega u\, d\mu.$$

If $u\in L_2(\Omega, \mu; H)$, we define its *covariance operator* by

$$\operatorname{Cov}(u)=E\{(u-E\{u\})\circ(u-E\{u\})\},$$

where $u\circ v\in\mathcal{L}(K, H)$ is defined for $u\in H, v\in K$ by

$$(u\circ v)h=u\langle v, h\rangle, \qquad h\in K.$$

Note that $u\circ u$ is a self-adjoint nuclear operator with trace $(u\circ u)=\|u\|^2$.

DEFINITION 1.3. An *X-valued stochastic process* is a map $u: T\times\Omega\to X$ which is measurable on $T\times\Omega$ using the Lebesgue measure on T.

DEFINITION 1.4. X- and Y-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 X and B in Y.

DEFINITION 1.5. An H-valued random variable $h \in L_2(\Omega, \mu; H)$ is *Gaussian* (*Poisson*) if $\langle h, e_i \rangle$ is a real Gaussian (Poisson) random variable for all i, where $\{e_i\}$ is a complete orthonormal basis for H.

DEFINITION 1.6. *Conditional expectation.* Let x, y be X- and Y-valued random variables respectively and let $\mathscr{X} = L_2(\Omega, \mu; X)$ and denote by σ the measure y induces on Y and $\mathscr{X}_y = \{x \in \mathscr{X}: x(\omega) = fy(\omega)\}$, where $f: Y \to X$ is measurable with respect to σ. Then the conditional expectation of x given y, $E_y\{x\}$ is the projection of x on $\mathscr{X}_y$.

Note that $\mathscr{X}_y$ is isometrically isomorphic to a closed subspace of $\mathscr{X}$ and so we write $\mathscr{X}_y = \tilde{L}_2(Y, \sigma; X)$. $E_y\{x\}$ also has the statistical interpretation as the best global estimate of x on y.

For estimation problems based on a Y-valued stochastic process $y(t) \in$ meas $(\Omega, \mu; C(T; Y))$, we define the random variable y_t to be the restriction of $y(s)$; $0 \leqq s \leqq T$ to $(0, t)$. Then y_t is a random variable with values in $C(0, t; Y)$ and $\mathscr{X}_{y_t}$ can be defined as above and $E_{y_t}\{x(t)\}$ is the conditional expectation of an X-valued stochastic process $x(t)$ with respect to y_t as t varies. $E_{y_t}\{x(t)\}$ is a well-defined X-valued stochastic process in $\int^{\oplus} \mathscr{X}_{y_t}\, dt$, where

$$\int^{\oplus} \mathscr{X}_{y_t}\, dt = \{x(t) \in L_2(T; \mathscr{X}) | x(t) \in \mathscr{X}_{y_t} \text{ a.e. } t\}.$$

In fact any $x \in \int^{\oplus} \mathscr{X}_{y_t}\, dt$ is an element of $\tilde{L}_2(C(0, t; Y), \sigma_t; L_2(0, t; \mathscr{X}))$, where σ_t is the probability measure induced by y_t on $C(0, t; Y)$.

1.3. Stochastic evolution equations on a Hilbert space. We summarize the main results from [9]. $(\Omega, \mathscr{P}, \mu)$ is a complete probability space and H, K are real separable Hilbert spaces.

DEFINITION 1.7. *Orthogonal increments process.* An *H-valued orthogonal increments process* $\{g(t), t \in T\}$ is such that

$$g(t) = \sum_{i=0}^{\infty} g_i(t) e_i, \tag{1.2}$$

where $\{e_i\}$ is a complete orthonormal basis for H, and $g_i(t)$ are real orthogonal increments processes satisfying

(a) $$E\{g_i(t)\} = \mu_i \rho(t),$$

where $\rho(t)$ is a monotonic nondecreasing real function and $\sum_{i=0}^{\infty} \mu_i < \infty$,

(b) $$E\{(\bar{g}_i(t_2) - \bar{g}_i(s_2))(\bar{g}_j(t_1) - \bar{g}_j(s_1))\} = 0, \qquad 0 \leqq s_1 < t_1 < s_2 < t_2 \leqq T,$$

(c) $$E\{(\bar{g}_i(t) - \bar{g}_i(s))(\bar{g}_j(t) - \bar{g}_j(s))\} = \lambda_{ij}(f(t) - f(s)), \qquad 0 \leqq s < t \leqq T,$$

where $\bar{g}_i(t) = g_i(t) - \mu_i \rho(t)$, f is a monotone nondecreasing function, $\sum_{i=0}^{\infty} \lambda_i < \infty$; $\lambda_{ii} = \lambda_i$ and $\lambda_{ij}^2 \leqq \lambda_i \lambda_j$. The *expectation function* is $r(t) = E\{g(t)\} = (\sum_{i=0}^{\infty} \mu_i e_i)\rho(t)$ and $\Lambda f(t)$ is the *incremental covariance function*, where Λ is given by

$$E\{[\bar{g}(t) - \bar{g}(s)] \circ [\bar{g}(t) - \bar{g}(s)]\} = \Lambda[f(t) - f(s)], \qquad 0 \leqq s \leqq t \leqq T,$$

where $\bar{g}(t) = g(t) - r(t)$. Λ is nuclear with trace $\Lambda = \sum_{i=0}^{\infty} \lambda_i$ and $\Lambda e_j = \sum_{i=0}^{\infty} \lambda_{ij} e_i$. If $r(t) = 0$, $g(t)$ is called a *centered orthogonal increments process.*

A special case of a centered process is the Wiener process,

$$w(t) = \sum_{i=0}^{\infty} \beta_i(t) e_i, \tag{1.3}$$

where β_i are mutually independent real Weiner processes with incremental covariance λ_i and $\sum_{i=0}^{\infty} \lambda_i < \infty$. $w(t)$ actually has independent increments and has continuous sample paths.

Another orthogonal increments process is the Poisson process,

$$p(t) = \sum_{i=0}^{\infty} \pi_i(t) e_i, \tag{1.4}$$

where π_i are mutually orthogonal real Poisson processes with parameter μ_i and $\sum_{i=0}^{\infty} \mu_i < \infty$.

For orthogonal increments processes we can define the stochastic integral $\int_T \Phi(s)\, dg(s)$ for $\Phi \in \mathscr{B}_2(T; \mathscr{L}(H, K))$ (the class of strongly measurable $\mathscr{L}(H, K$-valued functions with $\int_T \|\Phi(s)\|^2\, df(s) < \infty$).

$$\begin{aligned} &\int_0^t \Phi(s)\, dg(s) = \sum_{i=0}^{\infty} \int_0^t \Phi(s) e_i\, d\bar{g}_i(s) + \sum_{i=0}^{\infty} \mu_i \int_0^t \Phi(s) e_i\, d\rho(s), \\ &\int_0^t \Phi(s)\, dg(s) \in \mathscr{C}(T; L_2(\Omega, K)), \end{aligned} \tag{1.5}$$

and has the properties

$$E_{g_\alpha}\left\{\int_\alpha^t \Phi(s)\, dg(s)\right\} = \sum_{i=0}^{\infty} \mu_i \int_\alpha^t \Phi(s) e_i\, d\rho(s) = \int_\alpha^t \Phi(s)\, dr(s), \tag{1.6}$$

$$\begin{aligned} E\left\{\left\|\int_0^t \Phi(s)\, d\bar{g}(s)\right\|^2\right\} &= \int_0^t \operatorname{trace} \Phi^*(s)\Phi(s)\Lambda\, df(s) \\ &\leqq \int_0^t \|\Phi(s)\|^2 \operatorname{trace} \Lambda\, df(s). \end{aligned} \tag{1.7}$$

Using the definition of stochastic integration with respect to orthogonal increments, it is possible to consider stochastic evolution equations of the following type:

$$\begin{aligned} dx(t) &= A(t)x(t)\, dt + \Phi(t)\, dg(t) + h(t)\, dt, \\ x(0) &= x_0, \end{aligned} \tag{1.8}$$

where $A(t)$ is the generator of an evolution operator $\mathscr{U}(t, s)$, $\Phi \in \mathscr{B}_2(T; \mathscr{L}(K, H))$, $h \in L_2(T \times \Omega; H)$, $x_0 \in L_2(\Omega; H)$ and $g(t)$ satisfies Definition 1.7. First we define the mild solution to be

$$x(t) = \mathscr{U}(t, 0)x_0 + \int_0^t \mathscr{U}(t, s)\Phi(s)\, dg(s) + \int_0^t \mathscr{U}(t, s)h(s)\, ds. \tag{1.9}$$

Even if $\mathscr{U}(t, s)$ is only a mild evolution operator, (1.9) is a well-defined H-valued stochastic process and $\langle h, x(t)\rangle$ is continuous in mean square on T for all $h \in H$.

If $\mathcal{U}(t, s)$ is an almost strong evolution operator, under additional assumptions on Φ and g, (1.8) has a unique strong solution in the following sense (see [9]).

DEFINITION 1.8. Equation (1.8) *has a strong solution* $x(t)$ if $x \in \mathcal{C}(T; L_2(\Omega; H))$, $x(t) \in D(A(t))$ w.p. 1 and $x(t)$ satisfies (1.8) almost everywhere on $T \times \Omega$. We say that x *is unique* if whenever x_1 and x_2 are solutions,

$$\mu\{\omega: \sup_{t \in T} \|x_1(t) - x_2(t)\| = 0\} = 1.$$

2. The optimal control problem.

2.1. The model. Following the approach in [7], [8] we take the signal and observation models to be

$$x(t) = \mathcal{U}(t, 0)x_0 + \int_0^t \mathcal{U}(t, s)B(s)u(s)\, ds + \int_0^t \mathcal{U}(t, s)G(s)\, dg(s), \tag{2.1}$$

$$y(t) = \int_0^t C(s)x(s)\, ds + \int_0^t F(s)\, dw(s), \tag{2.2}$$

and the cost functional to be minimized

$$J(u) = E\left\{\int_0^T [\langle M(t)x(t), x(t)\rangle + \langle N(t)u(t), u(t)\rangle]\, dt\right\} + E\{\langle Rx(T), x(T)\rangle\}. \tag{2.3}$$

$(\Omega, \mathcal{P}, \mu)$ is a complete probability space, H, K, U are real separable Hilbert spaces, $T = [0, T]$ a real finite time interval, $\mathcal{U}(t, s)$ is a mild evolution operator on H, $B \in \mathcal{B}_\infty(T; \mathcal{L}(U, H))$, $G \in \mathcal{B}_2(T; \mathcal{L}(K, H))$, $x_0 \in L_2(\Omega, \mu; H)$; $E\{x_0\} = \bar{x}_0$, $\operatorname{Cov}\{x_0\} = P_0$ and g is K-valued orthogonal increments process. So if we take the control $u(t)$ to be in $L_2(T \times \Omega, U)$, (2.1) defines the signal $x(t)$ as an H-valued stochastic process. If w is an R^n-valued Wiener process with incremental covariance matrix W and if $F, F^{-1} \in L_2(T; \mathcal{L}(R^n))$, $C \in \mathcal{B}_\infty(T, \mathcal{L}(H, R^n))$, then the observation process $y(t)$ is a vector-valued stochastic process in

$$L_2(T \times \Omega; R^n) \cap \operatorname{meas}(\Omega, \mu; C(T; R^n)) \cap \mathcal{C}(T; L_2(\Omega, K^n)).$$

Finally we suppose that x_0, g, w are mutually independent and $M \in \mathcal{B}_\infty(T; \mathcal{L}(H))$, $N, N^{-1} \in \mathcal{B}_\infty(T; L(U))$, $R \in \mathcal{L}(H)$, $M(t) \geqq 0$, $N(t) > 0$, $R \geqq 0$.

Our problem is to find an admissible control $u \in \mathcal{U}_{\text{ad}}$ which minimizes $J(u)$. $\mathcal{U}_{\text{ad}}$ will be specified later.

2.2. Filtering results. We consider (2.1), (2.2) with no control, that is

$$\xi(t) = \mathcal{U}(t, 0)x_0 + \int_0^t \mathcal{U}(t, s)G(s)\, dg(s), \tag{2.4}$$

$$\eta(t) = \int_0^t C(s)\xi(s)\, ds + \int_0^t F(s)\, dw(s). \tag{2.5}$$

We quote the results from [8] for the filtering problem for the system (2.4), (2.5), that is, to find the best estimate $\hat{\xi}(t)$ of the form $\xi(t) = \int_0^t \mathcal{K}(t, s)\, d\eta(s) + \bar{\xi}(t)$ which minimizes $E\{\|\xi(t) - \hat{\xi}(t)\|^2\}$ for all $t \in T$.

There is a unique optimal linear least squares filter given by

$$\begin{aligned}\hat{\xi}(t) &= \mathcal{Y}(t,0)\bar{x}_0 + \int_0^t \mathcal{Y}(t,s)G(s)\,dr(s) + \int_0^t \mathcal{Y}(t,s)K(s)\,d\eta(s) \\ &= \mathcal{U}(t,0)\bar{x}_0 + \int_0^t \mathcal{U}(t,s)G(s)\,dr(s) + \int_0^t \mathcal{U}(t,s)K(s)\,d\nu(s),\end{aligned} \tag{2.6}$$

where $r(s) = E\{g(s)\}$, $\mathcal{Y}(t,s)$ is the perturbation of the mild evolution operator $\mathcal{U}(t,s)$ corresponding to $-K(t)C(t)$, $K(t) = P(t)C^*(t)(F(t)WF^*(t))^{-1}$ and $P(t)$, $\nu(t)$ are given by

$$\nu(t) = \eta(t) - \int_0^t C(s)\hat{\xi}(s)\,ds, \tag{2.7}$$

$$\begin{aligned}P(t)x = \mathcal{Y}(t,0)P_0\mathcal{Y}^*(t,0)x &+ \int_0^t \mathcal{Y}(t,s)[G(s)\Lambda G^*(s) \\ &+ P(s)C^*(s)(F(s)WF^*(s))^{-1}C(s)P(s)]Y^*(t,s)x\,ds.\end{aligned} \tag{2.8}$$

$P(t)$ is the unique solution of the Riccati equation (2.8) in the class of weakly continuous self-adjoint operator functions on H. $P(t)$ is also the covariance operator of the error process $\xi(t) - \hat{\xi}(t)$. The innovations process $\nu(t)$ defined by (2.7) also has the representation

$$\nu(t) = \int_0^t F(s)\,dv(s), \tag{2.9}$$

where $v(s)$ is an n-dimensional centered orthogonal increments process with incremental covariance matrix W. We also have

$$\mathcal{H}_{\eta_t} = \mathcal{H}_{\nu_t}, \tag{2.10}$$

where $\mathcal{H} = L_2(\Omega, \mu; X)$, X a Hilbert space and we are using the notation from § 1.3.

In [8] the special case of $g(t)$, a Wiener process and x_0, Gaussian, is considered, that is, (2.4) has a "Gaussian white noise" disturbance. Then $\hat{\xi}(t)$ is the best global estimate of $\xi(t)$ based on $\eta(s)$; $0 \leqq s \leqq t$, i.e.,

$$\hat{\xi}(t) = E_{\eta_t}\{\xi(t)\} = E_{\nu_t}\{\xi(t)\}. \tag{2.11}$$

This is also true whenever x_0 and $g(t)$ are Gaussian. Furthermore, $v(t)$ in (2.9) is now an n-dimensional Wiener process.

2.3. Admissible controls. Following Bensoussan and Viot [4], we take the class of admissible controls

$$\mathcal{U}_{\text{ad}} = \int^{\oplus} \mathcal{U}_{\eta_t}\,dt \cap \int^{\oplus} \mathcal{U}_{y_t}\,dt \subset L_2(T; \mathcal{U})$$

where $\mathcal{U} = L_2(\Omega, \mu; U)$. So for $u \in \mathcal{U}_{\text{ad}}$, (2.1) is a well-defined stochastic process and u is actually "feedback" in the sense that there exists a measurable map ψ

such that

$$u(t) = \psi(t, y_t).$$

Because of the problem involved in the existence and uniqueness of solutions of (2.1), (2.2) under arbitrary feedback controls, it is not clear how large is the class of feedback controls which yield controls in $\mathcal{U}_{\text{ad}}$. Although as in [4] we can show that feedback control laws $u(t) = \phi(t, y_t)$ where ϕ is measurable, nonanticipative and satisfies a uniform Lipschitz condition are admissible. We can also show that all linear feedback controls (see Lemma 2.1) and feedback controls of the observations with a small delay, $u(t) = \psi(t-\varepsilon, y_{t-\varepsilon})$ are admissible (see [4]).

First we define for each $u \in \mathcal{U}_{\text{ad}}$,

$$x_u(t) = \int_0^t \mathcal{U}(t, s)B(s)u(s)\, ds, \tag{2.12}$$

$$y_u(t) = \int_0^t C(s)x_u(s)\, ds, \tag{2.13}$$

and so we have

$$\begin{aligned} x(t) &= \xi(t) + x_u(t), \\ y(t) &= \eta(t) + y_u(t). \end{aligned} \tag{2.14}$$

Then define $\hat{x}(t)$, $e(t)$ by

$$\hat{x}(t) = \hat{\xi}(t) + x_u(t), \tag{2.15}$$

$$e(t) = x(t) - \hat{x}(t) = \xi(t) - \hat{\xi}(t). \tag{2.16}$$

Then from (2.6), (2.7) and (2.12), we obtain

$$\begin{aligned} \hat{x}(t) = \mathcal{U}(t, 0)\bar{x}_0 &+ \int_0^t \mathcal{U}(t, s)B(s)u(s)\, ds \\ &+ \int_0^t \mathcal{U}(t, s)K(s)\, d\nu(s) + \int_0^t \mathcal{U}(t, s)G(s)\, dr(s), \end{aligned} \tag{2.17}$$

$$\nu(t) = y(t) - \int_0^t C(s)\hat{x}(s)\, ds. \tag{2.18}$$

LEMMA 2.1 (cf. [12], [13]). *The following class of linear feedback controls are admissible*:

$$u(t) = \int_0^t L(t, s)\, dy(s) + \bar{u}(t),$$

where $\bar{u} \in L_2(T; U)$ *is nonrandom and* $L \in \mathcal{B}_2(\Delta(T); \mathcal{L}(R^n, U))$.

Proof. It is easy to see that (2.1) and (2.2) has a unique solution corresponding to $u(t)$. Hence $u \in \int^{\oplus} \mathcal{U}_{y_t}\, dt$. We show that $u \in \int^{\oplus} \mathcal{U}_{\nu_t}\, dt$ $(= \int^{\oplus} \mathcal{U}_{\eta_t}\, dt$ by (2.10)) by showing that there exists $R \in \mathcal{B}_2(\Delta(T); \mathcal{L}(R^n, U))$ such that $u(t) =$

$\int_0^t R(t, s)\, d\nu(s) + \bar{u}_1(t)$. Now

$$u(t) = \int_0^t L(t, s)\, dy(s) + \bar{u}(t)$$

$$= \int_0^t L(t, s)\, d\nu(s) + \int_0^t L(t, s)C(s)\Big[\mathcal{U}(s, 0)\bar{x}_0 + \int_0^s \mathcal{U}(s, \alpha)B(\alpha)u(\alpha)\, d\alpha$$

$$+ \int_0^s \mathcal{U}(s, \alpha)G(\alpha)\, dr(\alpha) + \int_0^s \mathcal{U}(s, \alpha)K(\alpha)\, d\nu(\alpha)\Big]\, ds + \bar{u}(t).$$

So

$$u(t) = \bar{u}_2(t) + \int_0^t L_1(t, s)\, d\nu(s) + \int_0^t L_2(t, s)u(s)\, ds, \tag{2.19}$$

where

$$\bar{u}_2(t) = \bar{u}(t) + \int_0^t L(t, s)C(s)\mathcal{U}(s, 0)\bar{x}_0\, ds$$

$$+ \int_0^t L(t, s)C(s) \int_0^s \mathcal{U}(s, \alpha)G(\alpha)\, dr(\alpha)\, ds$$

is deterministic and

$$L_1(t, s) = \int_s^t L(t, \alpha)C(\alpha)\mathcal{U}(\alpha, s)K(s)\, d\alpha + L(t, s),$$

$$L_2(t, s) = \int_s^t L(t, \alpha)C(\alpha)\mathcal{U}(\alpha, s)B(s)\, d\alpha.$$

But solving (2.19) for u is equivalent to solving the deterministic Volterra integral equation on $L_2(T, U)$,

$$f(t) = g_0(t) + \int_0^t L_2(t, s)f(s)\, ds,$$

and this always has a unique solution for $f \in L_2(T; U)$ of the form

$$f(t) = \int_0^t S(t, s)g_0(s)\, ds$$

for some kernel function $S \in \mathcal{B}_2(\Delta(T); \mathcal{L}(U))$ (see [11]). So

$$u(t) = \int_0^t S(t, s)\bar{u}_2(s)\, ds + \int_0^t S(t, s) \int_0^s L_1(s, \alpha)\, d\nu(\alpha)\, ds$$

$$= \int_0^t S(t, s)\bar{u}_2(s)\, ds + \int_0^t \int_\alpha^t S(t, s)L_1(s, \alpha)\, ds\, d\nu(\alpha)$$

$$= \bar{u}_1(t) + \int_0^t R(t, \alpha)\, d\nu(\alpha)$$

as required.

We note that $E_{\nu_t}\{\hat{x}(t)\} = \hat{x}(t)$ and if ξ is Gaussian $E_{\nu_t}\{\xi(t)\} = \hat{\xi}(t)$ and so it is easily verified that

$$E\{\langle M(t)x(t), x(t)\rangle\} = E\{\langle M(t)\hat{x}(t), \hat{x}(t)\rangle\} + E\{\langle M(t)e(t), e(t)\rangle\}. \tag{2.20}$$

So in the Gaussian case the problem of minimizing (2.3) is reduced to minimizing

$$J_0(u) = E\Big\{\int_0^T [\langle M(t)\hat{x}(t), \hat{x}(t)\rangle + \langle N(t)u(t), u(t)\rangle]\, dt\Big\} + E\{\langle R\hat{x}(T), \hat{x}(T)\rangle\Big\} \tag{2.21}$$

subject to (2.17).

2.4. Optimal control assuming complete observations. For complete observations, we do not need to assume x_0, $g(t)$ to be Gaussian. We suppose we can observe exactly the following signal process:

$$z(t) = \mathcal{U}(t, 0)x_0 + \int_0^t \mathcal{U}(t, s)B(s)u(s)\, ds + \int_0^t \mathcal{U}(t, s)G(s)\, dr(\alpha) + \int_0^t \mathcal{U}(t, s)G_0(s)\, dv(s), \tag{2.22}$$

where $G_0 \in \mathcal{B}_\infty(T; \mathcal{L}(K, H))$, v is a K-valued centered orthogonal increments process, and $x_0 \in L_2(\Omega; H)$. We wish to minimize the following cost functional over admissible controls in $\tilde{\mathcal{U}}_{\text{ad}} = \int^{\oplus} \mathcal{U}_{v_t}\, dt$:

$$J_1(u) = E\Big\{\int_0^T [\langle M(t)z(t), z(t)\rangle + \langle N(t)u(t), u(t)\rangle]\, dt\Big\} + E\{\langle Rz(T), z(T)\rangle\}. \tag{2.23}$$

For each $u \in \tilde{\mathcal{U}}_{\text{ad}}$, $z(t)$ is well-defined and can be written

$$z(t) = z_0(t) + (\Phi u)(t), \tag{2.24}$$

where

$$z_0(t) = \mathcal{U}(t, 0)x_0 + \int_0^t \mathcal{U}(t, \dot{s})N(s)\, dr(s) + \int_0^t \mathcal{U}(t, s)G_0(s)\, dv(s)$$

is independent of u and $\Phi \in \mathcal{L}(L_2(T, U), L_2(T, H))$ is given by

$$(\Phi u)(t) = \int_0^t \mathcal{U}(t, s)B(s)u(s)\, ds. \tag{2.25}$$

Using standard techniques similar to those used in [1], [2], [5], we show that $J_1(u)$ has a unique minimizing control.

LEMMA 2.2. *There exists a unique minimizing control in $\tilde{\mathcal{U}}_{\text{ad}}$ given by*

$$u_*(t) = -N(t)^{-1}B^*\Big[\mathcal{U}^*(T, t)RE_{\nu_t}\{z_*(T)\} + \int_t^T \mathcal{U}^*(s, t)M(s)E_{v_t}\{z_*(s)\}\, ds\Big], \tag{2.26}$$

where $z_*(t)$ is the signal (2.22) (*or* (2.24)) *corresponding to $u_*(t)$.*

Proof. Let $\mathcal{U} = L_2(\Omega, \mu; U)$; then it is easily verified that $\tilde{\mathcal{U}}_{ad}$ is a subspace of $L_2(T; \mathcal{U})$. Thus $\tilde{\mathcal{U}}_{ad}$ itself is a Hilbert space. Let Φ_1 be defined by $\Phi_1 u = (\Phi u)(T)$ and let $\mathcal{H} = L_2(\Omega_1 \mu; H)$. We denote by $(\cdot, \cdot)$ the inner product in $\mathcal{H}$ and by $\langle\langle \cdot, \cdot \rangle\rangle$ inner products in $L_2(T; \mathcal{H})$ (or $L_2(T; \mathcal{U})$). Then $J_1(u)$ can be written

$$J_1(u) = \langle\langle M(z_0 + \Phi u), z_0 + \Phi u \rangle\rangle + \langle\langle Nu, u \rangle\rangle + (R[z_0(T) + \Phi_1 u], z_0(T) + \Phi_1 u),$$

so it is strictly convex and lower semi-continuous. Hence there exists a unique minimizing element u_* given by

$$Nu_* + \Phi_1^* R[z_0(T) + \Phi_1 u_*] + \Phi^* M(z_0 + \Phi u_*) = 0.$$

Let $z_*(t) = z_0(t) + (\Phi u_*)(t)$; then interpreting adjoint operators Φ_1^*, Φ^* appropriately (see [1], [2]) we obtain

$$u_*(t) = -N(t)^{-1} B^*(t) \left[\mathcal{U}^*(T, t) R E_{v_t} \left\{ z_*(T) + \int_t^T \mathcal{U}^*(s, t) M(s) E_{v_t}\{z_*(s)\} \right\} \right] ds.$$

Define the adjoint state

$$p(t) = \int_t^T \mathcal{U}^*(s, t) M(s) z_*(s)\, ds + \mathcal{U}^*(T, t) R z_*(T); \tag{2.27}$$

then

$$u_*(t) = -N(t)^{-1} B^*(t) E_{v_t}\{p(t)\}.$$

From (2.22) we obtain

$$\begin{aligned} z_*(s) = \mathcal{U}(s, t) z_*(t) &- \int_t^s \mathcal{U}(s, \sigma) D(\sigma) \hat{p}(\sigma)\, d\sigma \\ &+ \int_t^s \mathcal{U}(s, \sigma) G(\sigma)\, dr(\sigma) + \int_t^s \mathcal{U}(s, \sigma) G_0(\sigma)\, dv(\sigma), \end{aligned} \tag{2.28}$$

where $D(t) = B(t) N(t)^{-1} B^*(t)$ and $\hat{p}(\sigma) = E_{v_\sigma}\{p(\sigma)\}$.

Let $Q(t)$ be the solution of the Riccati equation

$$\begin{aligned} Q(t)h = \mathcal{U}_Q^*(T, t) R \mathcal{U}_Q(T, t) h \\ + \int_t^T \mathcal{U}_Q^*(s, t)[M(s) + Q(s) D(s) Q(s)] U_Q(s, t) h\, ds, \qquad h \in H, \end{aligned} \tag{2.29}$$

where $\mathcal{U}_Q(s, t)$ is the perturbation of the evolution operator $\mathcal{U}(s, t)$ by $-D(s)Q(s)$. The existence and uniqueness of the solution of (2.29) in the class of weakly continuous self-adjoint operator-valued functions is shown in [6]. From Lemma 2.3 in [15] we also know that $Q(t)$ satisfies another equivalent Riccati equation:

$$\begin{aligned} Q(t)h = \mathcal{U}^*(T, t) R \mathcal{U}(T, t) h + \int_t^T \mathcal{U}^*(s, t) \\ \cdot [M(s) - Q(s) D(s) Q(s)] \mathcal{U}(s, t) h\, ds. \end{aligned} \tag{2.30}$$

Let $\tilde{p}(s) = E_{v_t}\{p(s)\}$, $s \geqq t$. The following lemma enables us to write u_* in the feedback form.

LEMMA 2.3.

$$\hat{p}(t)-Q(t)z_*(t)=\int_t^T \mathcal{U}_Q(s,t)Q(s)G(s)\,dr(s).$$

Proof. From (2.27) and (2.28) we have

$$\begin{aligned}\hat{p}(t)=\tilde{p}(t)=\int_t^T \mathcal{U}^*(s,t)M(s)\Big[\mathcal{U}(s,t)z_*(t)-\int_t^s \mathcal{U}(s,\sigma)D(\sigma)\tilde{p}(\sigma)\,d\sigma\\ +\int_t^s \mathcal{U}(s,\sigma)G(\sigma)\,dr(\sigma)\Big]\,ds\\ +\mathcal{U}^*(T,t)R\Big[\mathcal{U}(T,t)z_*(t)-\int_t^T \mathcal{U}(T,\sigma)D(\sigma)\tilde{p}(\sigma)\,d\sigma\\ +\int_t^T \mathcal{U}(t,\sigma)G(\sigma)\,dr(\sigma)\Big]\\ =\Big[\int_t^T \mathcal{U}^*(s,t)M(s)\mathcal{U}(s,t)\,ds+\mathcal{U}^*(T,t)R\mathcal{U}(T,t)\Big]z_*(t)\\ -\Big[\int_t^T \mathcal{U}^*(s,t)M(s)\int_t^s \mathcal{U}(s,\sigma)D(\sigma)\tilde{p}(\sigma)\,d\sigma\\ +\mathcal{U}^*(T,t)R\int_t^T \mathcal{U}(T,\sigma)D(\sigma)\tilde{p}(\sigma)\,d\sigma\Big]\\ +\Big[\int_t^T \mathcal{U}^*(s,t)M(s)\int_t^s \mathcal{U}(s,\sigma)G(\sigma)\,dr(\sigma)\\ +\mathcal{U}^*(T,t)R\int_t^T \mathcal{U}(T,\sigma)G(\sigma)\,dr(\sigma)\Big].\end{aligned}$$

Using (2.30), (2.28) and Fubini's theorem, we can show that

$$\begin{aligned}\tilde{p}(t)-Q(t)z_*(t)=-\int_t^T \mathcal{U}^*(s,t)Q(s)D(s)[\tilde{p}(s)-Q(s)\tilde{z}_*(s)]\,ds\\ +\int_t^T \mathcal{U}^*(s,t)Q(s)G(s)\,dr(s),\end{aligned}$$

where $\tilde{z}_*(s)=E_{v_t}\{z_*(s)\}$, $s\geqq t$. Since $\tilde{p}(T)-Q(T)\tilde{z}_*(T)=0$, using the adjoint version of (1.1) we have

$$\hat{p}(t)-Q(t)z_*(t)=\tilde{p}(t)-Q(t)\tilde{z}_*(t)=\int_t^T \mathcal{U}_Q^*(s,t)Q(s)G(s)\,dr(s).$$

Now the optimal control $u_*(t)$ is given by

$$u_*(t)=-N(t)^{-1}B^*(t)[Q(t)z_*(t)+\bar{p}(t)], \tag{2.31}$$

where

$$\bar{p}(t)=\int_t^T \mathcal{U}_Q^*(s,t)Q(s)G(s)\,dr(s). \tag{2.32}$$

This is the same feedback control law as in the deterministic case.

Remark. In the $r \equiv 0$ case if we define $Q(t)$ by

$$Q(t)h = \mathcal{U}^*(T, t)RE_{v_t}\{z(T, t, h)\} + \int_t^T \mathcal{U}^*(s, t)M(s)E_{v_t}\{z(s, t, h)\}\, ds, \qquad h \in \mathcal{H}_{v_t},$$

where $z(\cdot, t, h)$ is the unique optimal signal for the problem described by

$$z(s) = \mathcal{U}(s, t)h + \int_t^s \mathcal{U}(s, \alpha)B(\alpha)u(\alpha)\, d\alpha + \int_t^s \mathcal{U}(s, \alpha)G_0(\alpha)\, dv(\alpha),$$

$$J_t(\alpha) = E\left\{\int_t^T [\langle M(s)z(s), z(s)\rangle + \langle N(s)u(s), u(s)\rangle]\, ds\right\} + E\{\langle Rz(T), z(T)\rangle\},$$

then we can show that $Q(t)$ is a self-adjoint operator on H satisfying (2.28).

We can now state our main result of this section.

THEOREM 2.1. *Consider the optimal control problem of* (2.22) *where we wish to minimize* (2.23) *over the class of admissible controls in* $\tilde{\mathcal{U}}_{\text{ad}} = \int^{\oplus} \mathcal{U}_{v_t}\, dt$. *Then there exists a unique optimal control given by*

$$\begin{aligned} u_*(t) &= -N(t)^{-1}B^*(t)[Q(t)z_*(t) + \bar{p}(t)], \\ \bar{p}(t) &= \int_t^T \mathcal{U}_Q^*(s, t)Q(s)G(s)\, dr(s), \\ z_*(t) &= \mathcal{U}_Q(t, 0)\bar{x}_0 + \int_0^t \mathcal{U}_Q(t, s)G_0(s)\, dv(s) \\ &\quad + \int_0^t \mathcal{U}_Q(t, s)G(s)\, dr(s) - \int_0^t \mathcal{U}_Q(t, s)D(s)\bar{p}(s)\, ds, \end{aligned} \tag{2.33}$$

where $Q(t)$ *is the unique solution of the integral Riccati equation* (2.29) *and* $\mathcal{U}_Q(t, s)$ *is the perturbation of the mild evolution operator* $\mathcal{U}(t, s)$ *by* $-D(t)Q(t)$.

Our optimal control is also the best feedback control law of the type $u(t) = L(t)z(t) + \bar{u}(t)$, $L \in \mathcal{B}_\infty(T, \mathcal{L}(H, U))$, $\bar{u} \in L_2(T, U)$ since for such feedback controls (2.22) has a unique solution and they are admissible.

2.5. Incomplete observations for the Gaussian case. We return to our original problem of § 2.1 of minimizing $J(u)$ over all $u \in \mathcal{U}_{\text{ad}}$ under the assumption that x_0, $g(t)$ are Gaussian. First we find the optimal control in the class $\int^{\oplus} \mathcal{U}_{v_t}\, dt$. This is a well-posed problem as $\nu(t)$ is completely specified by (2.6), (2.7) and so is independent of the controls. Using the representation (2.9) for $\nu(t)$, we have $\mathcal{U}_{\nu_t} = \mathcal{U}_{v_t}$ (since $F^{-1}(t)$ exists) and the expression (2.17) for $\hat{x}(t)$ becomes

$$\begin{aligned} \hat{x}(t) &= \mathcal{U}(t, 0)\bar{x}_0 + \int_0^t \mathcal{U}(t, s)B(s)u(s)\, ds \\ &\quad + \int_0^t \mathcal{U}(t, s)K(s)F(s)\, dv(s) + \int_0^t \mathcal{U}(t, s)G(s)\, dr(s), \end{aligned} \tag{2.34}$$

which gives a complete description of the system and the cost to be minimized is

$$J_0(u) = E\left\{\int_0^T [\langle M(t)\hat{x}(t), \hat{x}(t)\rangle + \langle N(t)u(t), u(t)\rangle]\, dt\right\} + E\{\langle R\hat{x}(T), \hat{x}(T)\rangle\}. \tag{2.21}$$

The following theorem is a direct consequence of the results in § 2.4.

THEOREM 2.2. *The problem described by* (2.34), (2.21) *has a unique optimal control* u_* *given by*

$$u_*(t) = -N(t)^{-1}B^*(t)[Q(t)\hat{x}_*(t) + \bar{\rho}(t)], \tag{2.35}$$

where $Q(t)$ *is the unique solution of* (2.29), $\hat{x}_*(t)$, $\bar{\rho}(t)$ *are given by*

$$\bar{\rho}(t) = \int_t^T \mathcal{U}_Q^*(s, t)Q(s)G(s)\, dr(s), \tag{2.36}$$

$$\hat{x}_*(t) = \tilde{\mathcal{U}}(t, s)\bar{x}_0 + \int_0^t \tilde{\mathcal{U}}(t, s)G(s)\, dr(s) - \int_0^t \tilde{\mathcal{U}}(t, s)D(s)\bar{\rho}(s)\, ds + \int_0^t \tilde{\mathcal{U}}(t, s)P(s)C^*(s)[F(s)WF^*(s)]^{-1}\, dy(s), \tag{2.37}$$

and $\tilde{\mathcal{U}}(t, s)$ *is the perturbation of the mild evolution operator* $\mathcal{U}(t, s)$ *by* $-D(t)Q(t) - P(t)C^*(t)[F(t)WF(t)^*]^{-1}C(t)$. *Furthermore* $u_* \in \mathcal{U}_{\text{ad}}$.

Proof. Equations (2.35), (2.36) are direct results of § 2.4, and $\hat{x}_*(t)$ is given by

$$\hat{x}_*(t) = \mathcal{U}_Q(t, 0)x_0 + \int_0^t \mathcal{U}_Q(t, s)G(s)\, dr(s) - \int_0^t \mathcal{U}_Q(t, s)D(s)\bar{\rho}(s)\, ds + \int_0^t \mathcal{U}_Q(t, s)K(s)F(s)\, dv(s). \tag{2.38}$$

Substituting $\nu(t) = \int_0^t F(s)\, dv(s)$ and for ν in terms of y from (2.18) yields

$$\hat{x}_*(t) = \mathcal{U}_Q(t, 0)\bar{x}_0 + \int_0^t \mathcal{U}_Q(t, s)G(s)\, dr(s) - \int_0^t \mathcal{U}_Q(t, s)D(s)\bar{\rho}(s)\, ds + \int_0^t \mathcal{U}_Q(t, s)K(s)\, dy(s) - \int_0^t \mathcal{U}_Q(t, s)K(s)C(s)\hat{x}_*(s)\, ds,$$

and (2.37) follows from the definition of $\tilde{\mathcal{U}}(t, s)$ as a perturbation of $\mathcal{U}(t, s)$ by $-D(t)Q(t) - P(t)C^*(t)[F(t)WF^*(t)]^{-1}$ (see (1.1)). So $\hat{x}_*(t) \in \mathcal{H}_{y_t}$ a.a. t and $u_*(t) \in \mathcal{U}_{y_t}$ a.a. t from (2.10), $\mathcal{U}_{\nu_t} = \mathcal{U}_{\eta_t}$, and so u_* is admissible.

In [4] it is proved that $u \in \int^t \mathcal{U}_{\nu_t}\, dt \cap \int^t \mathcal{U}_{y_t}\, dt$, then $E_\nu\{\cdot\} = E_{y_t}\{\cdot\}$ and so we have

COROLLARY 2.2.

$$\hat{x}_*(t) = E_{y_t}\{x(t)\}.$$

Summarizing, we state our main result.

THEOREM 2.3 (separation theorem). *Consider the problem of minimizing* $J(u)$ *given by* (2.3) *subject to* (2.1), (2.2) *where* x_0 *and* $g(t)$ *are Gaussian over the class of controls from* $\mathcal{U}_{\text{ad}} = \int^{\oplus} \mathcal{U}_{\eta_t}\, dt \cap \int^{\oplus} \mathcal{U}_{y_t}\, dt$. *Then there exists a unique optimal*

control $u_*(t)$ *given by*

$$u_*(t) = -N(t)^{-1}B^*(t)[Q(t)\hat{x}_*(t) + \bar{\rho}(t)], \tag{2.39}$$

$$\begin{aligned}\hat{x}_*(t) = \tilde{\mathcal{U}}(t,0)\bar{x}_0 &+ \int_0^t \tilde{\mathcal{U}}(t,s)G(s)\,dr(s) - \int_0^t \tilde{\mathcal{U}}(t,s)D(s)\bar{\rho}(s)\,ds \\ &+ \int_0^t \tilde{\mathcal{U}}(t,s)P(s)C^*(s)[F(s)WF^*(s)]^{-1}\,dy(s),\end{aligned} \tag{2.40}$$

where P, Q *are the unique solution of Riccati equations* (2.8) *and* (2.29), *respectively, and* $\tilde{\mathcal{U}}(t,s)$ *is the perturbation of the mild evolution operator* $\mathcal{U}(t,s)$ *by* $-D(t)Q(t) - P(t)C^*(t)[F(t)WF^*(t)]^{-1}C(t)$. *The optimal cost is given by*

$$\begin{aligned}J(u_*) = \operatorname{trace}\{RP(T)\} &+ \int_0^T \operatorname{trace}\{M(t)P(t)\}\,dt + \langle R\bar{x}(T), \bar{x}(T)\rangle \\ &+ \int_0^T \langle [M(t) + Q(t)D(t)Q(t)]\bar{x}(t), \bar{x}(t)\rangle\,dt \\ &+ \operatorname{trace}\int_0^T F^*(t)K^*(t)Q(t)K(t)F(t)W\,dt \\ &+ 2\int_0^T \langle D(t)Q(t)\bar{x}(t), \bar{\rho}(t)\rangle\,dt + \int_0^T \langle D(t)\bar{\rho}(t), \bar{\rho}(t)\rangle\,dt,\end{aligned} \tag{2.41}$$

where

$$\bar{x}(t) = E\{\hat{x}_*(t)\} = \mathcal{U}_Q(t,0)\bar{x}_0 + \int_0^t \mathcal{U}_Q(t,s)G(s)\,dr(s) - \int_0^t \mathcal{U}_Q(t,s)D(s)\bar{\rho}(s)\,ds.$$

Proof. From (2.20),

$$\begin{aligned}J(u) &= J_0(u) + E\{\langle \operatorname{Re}(T), e(T)\rangle\} + E\left\{\int_0^T \langle M(t)e(t), e(t)\rangle\,dt\right\} \\ &= J_0(u) + \operatorname{trace}\{RP(T)\} + \int_0^T \operatorname{trace}\{M(t)P(t)\}\,dt\end{aligned}$$

since

$$\operatorname{Cov}\{e(t)\} = P(t).$$

$$\begin{aligned}J_0(u_*) &= E\left\{\int_0^T [\langle M(t)\hat{x}_*(t), \hat{x}_*(t)\rangle + \langle N(t)u_*(t), u_*(t)\rangle]\,dt\right\} + E\{\langle R\hat{x}_*(T), \hat{x}_*(T)\rangle\} \\ &= E\left\{\int_0^T \langle [M(t) + Q(t)D(t)Q(t)]\hat{x}_*(t), \hat{x}_*(t)\rangle\,dt\right\} + E\{\langle R\hat{x}_*(T), \hat{x}_*(T)\rangle\} \\ &\quad + E\left\{\int_0^T 2\langle B^*(t)Q(t)\hat{x}_*(t), N(t)^{-1}B^*(t)\bar{\rho}(t)\rangle\,dt\right\} + \int_0^T \langle D(t)\bar{\rho}(t), \bar{\rho}(t)\rangle\,dt.\end{aligned}$$

Since

$$\hat{x}_*(t) = \bar{x}(t) + \int_0^t \mathcal{U}_Q(t, s)K(s)F(s)\, dv(s),$$

$$\begin{aligned} J_0(u_*) &= \int_0^T \langle [M(t) + Q(t)D(t)Q(t)]\bar{x}(t), \bar{x}(t)\rangle\, dt + \langle R\bar{x}(T), \bar{x}(T)\rangle \\ &\quad + \text{trace} \int_0^T F^*(t)K^*(t)\mathcal{U}_Q^*(T, t)R\mathcal{U}_Q(T, t)K(t)F(t)W\, dt \\ &\quad + \text{trace} \int_0^T \int_0^t F^*(s)K^*(s)\mathcal{U}_Q^*(t, s) \\ &\qquad \cdot [M(t) + Q(t)D(t)Q(t)]\mathcal{U}_Q(t, s)K(s)F(s)W\, ds\, dt \\ &\quad + 2\int_0^T \langle B^*(t)Q(t)\bar{x}(t), N^{-1}(t)B^*(t)\bar{\rho}(t)\rangle\, dt + \int_0^T \langle D(t)\bar{\rho}(t), \bar{\rho}(t)\rangle\, dt \\ &= \int_0^T \langle [M(t) + Q(t)D(t)Q(t)]\bar{x}(t), \bar{x}(t)\rangle\, dt + \langle R\bar{x}(T), \bar{x}(T)\rangle \\ &\quad + \text{trace} \int_0^T F^*(t)K^*(t)Q(t)K(t)F(t)W\, dt \\ &\quad + 2\int_0^T \langle D(t)Q(t)\bar{x}(t), \bar{\rho}(t)\rangle\, dt + \int_0^T \langle D(t)\bar{\rho}(t), \bar{\rho}(t)\rangle\, dt. \end{aligned}$$

Thus (2.41) follows.

Remark 1. If $g(t)$ is centered, then $r(t) = 0$, $\bar{\rho}(t) = 0$ and (2.41) becomes

$$\begin{aligned} J(u_*) &= \text{trace}\,\{RP(T)\} + \int_0^T \text{trace}\,\{M(t)P(t)\}\, dt + \langle Q(0)\bar{x}_0, \bar{x}_0\rangle \\ &\quad + \int_0^T \text{trace}\, F^*(t)K^*(t)Q(t)K(t)F(t)W\, dt. \end{aligned}$$

Remark 2. If our signal model (2.1) includes an extra deterministic forcing term $\int_0^t \mathcal{U}(t, s)f(s)\, ds$, where $f \in L_2(T; H)$, then Theorem 2.3 holds replacing $\bar{\rho}(t)$ in (2.39), (2.40), (2.41) by

$$\bar{\rho}_1(t) = \int_t^T \mathcal{U}_Q^*(s, t)Q(s)f(s)\, ds + \int_t^T \mathcal{U}_Q^*(s, t)Q(s)G(s)\, dr(s)$$

and adding

$$\int_0^t \tilde{\mathcal{U}}(t, s)f(s)\, ds \quad \text{to } \hat{x}_*(t) \quad \text{in (2.40).}$$

3. Applications and concluding remarks. As in Bensoussan and Viot [4] it is possible to extend the separation principle to more general classes of cost functionals and to allow for u to take its values in a compact convex subset of U. Although our class of admissible controls is difficult to specify, it does include all

feedback controls of a measurable, nonanticipative and Lipschitzian type as defined in § 2.3. In particular our control u_* is optimal in the class of linear feedback controls of the form

$$u(t) = \int_0^t L(t, s)\, dy(s) + \bar{u}(t)$$

from Lemma 2.1.

The types of systems covered by the theory is large, including the parabolic partial differential systems considered by Bensoussan and Viot [4], hyperbolic partial differential equations and linear partial-integro-differential equations; for several examples of distributed systems described by evolution equations see [6].

Another large class of systems included in this theory is linear delay equations and so this paper generalizes some results of Lindquist [12], [13], although here we cannot allow for time delays in the control or in the cost. For details of representing stochastic delay equations by stochastic evolution equations we refer the reader to [8], [9], [10].

For complete observations the noise disturbance in the system is not restricted to be of Gaussian white noise type, but can allow for jump type disturbances, for example Poisson-type, and colored measurement noise. Several examples of different types of noise disturbances may be found in [8], [9], [10]. However, for incomplete observations we do need the Gaussian assumption.

Acknowledgment. We thank the referee for suggesting improvements in our original manuscript.

REFERENCES

[1] A. V. BALAKRISHNAN, *Stochastic control: A function space approach*, this Journal, 10 (1972), pp. 285–297.

[2] ———, *Stochastic Optimization Theory in Hilbert Spaces. I*, Appl. Math. and Opt., 1 (1974), pp. 97–120.

[3] A. BENSOUSSAN, *On the separation principle for distributed parameter systems*, IFAC Conference on Distributed Parameter Systems, Banff, Canada, 1971.

[4] A. BENSOUSSAN AND M. VIOT, *Optimal control of stochastic linear distributed parameter systems*, this Journal, 13 (1975), pp. 904–926.

[5] R. A. BROOKS, *Linear stochastic control: An extended separation principle*, J. Math. Anal. Appl., 38 (1972), pp. 569–587.

[6] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Riccati equation for systems parameter described by evolution operators*, this Journal, 14 (1976), pp. 951–983.

[7] R. F. CURTAIN, *Infinite-dimensional filtering*, this Journal (1975), pp. 89–104.

[8] ———, *Infinite dimensional estimation theory for linear systems*, Rep. 38, Control Theory Centre, Univ. of Warwick, Coventry, England, 1975.

[9] ———, *Estimation theory of abstract evolution equations excited by general white noise processes*, Rep. 40, Control Theory Centre, Univ. of Warwick, Coventry, England, 1975; this Journal, 14 (1976), pp. 1124–1150.

[10] ———, *Stochastic evolution equations with general white noise disturbance*, Rep. 41, Control Theory Centre, Univ. of Warwick, Coventry, England, 1975; J. Math. Anal. Appl., to appear.

[11] I. GOHBERG AND M. G. KREIN, *Theory and applications of Volterra operators in Hilbert space*, Amer. Math. Soc. Transl. Monograph 24, American Mathematical Society, Providence, R.I., 1970.

[12] A. LINDQUIST, *Optimal control of linear stochastic systems with applications to time lag systems*, Information Sci., 5 (1973), pp. 81–126.
[13] ———, *On feedback control of linear stochastic systems*, this Journal, 11 (1973), pp. 323–343.
[14] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, A. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131–212.
[15] A. ICHIKAWA, *Linear quadratic differential games in a Hilbert space*, this Journal, 14 (1976), pp. 120–136.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

A HYBRID ALGORITHM FOR NONLINEAR PROGRAMMING*

R. W. CHANEY†

Abstract. In this paper we present a "hybrid" first-order algorithm, designed to solve finite-dimensional nonlinear programming problems having both equality and inequality constraints. The algorithm is built up from the method of exterior penalty functions, the Pironneau–Polak method of centers, and a quasi-Newton method of Luenberger. The algorithm is shown to generate sequences which, under appropriate hypothesis, will converge linearly at a rate "asymptotically" independent of the penalty coefficient. The main technique employed in the convergence analysis is the finite-dimensional version of the indirect method of Hestenes. The algorithm seems well suited to a situation in which there are many variables and rather few constraints; for then, the demands made on storage and the direction finding subprocedure would be relatively modest.

1. Introduction. Suppose $f, f_1, f_2, \cdots, f_m, \phi_1, \cdots, \phi_\nu$ are continuously differentiable functions on n-dimensional real Euclidean space R^n. Define

$$S = \bigcap_{k=1}^{m} \{x \in R^n : f_k(x) \leqq 0\} \cap \bigcap_{j=1}^{\nu} \{x \in R^n : \phi_j(x) = 0\}.$$

We formulate the problem

$$\text{P:} \quad \text{minimize } f \text{ over } S.$$

Next, we define

$$S_1 = \bigcap_{k=1}^{m} \{x \in R^n : f_k(x) \leqq 0\}.$$

Given $t > 0$, we set $f_0 = f + \frac{1}{2}t \sum_{j=1}^{\nu} \phi_j^2$ and formulate the problem

$$\text{P}_t\text{:} \quad \text{minimize } f_0 \text{ over } S_1.$$

Throughout the paper, the function f_0 will be understood to depend upon t, although our notation suggests otherwise.

We shall present below a first-order algorithm for solving problem P_t. The algorithm is a blend of a quasi-Newton method due to Luenberger [7] and the Pironneau–Polak method of centers [10]. The purpose of the former method is to circumvent the ill-conditioning which often plagues penalty function approaches; the latter method deals with the inequality constraints in such a way as to maintain feasibility for problem P_t at every iteration. The direction-finding procedure in the Pironneau–Polak algorithm is a quadratic programming problem. Its dual has $m+1$ variables, which are constrained only to be nonnegative and to add up to 1 (see [9]).

The algorithm is presented below in § 2, its linear convergence is established in § 3, and two numerical examples are given in § 4. The rest of this section is devoted to a discussion of several connections between problems P_t and P. Of course, many connections are well-known; see, e.g., [4], [7] or [11]. First, we must

* Received by the editors December 10, 1975, and in revised form July 8, 1976.

† Department of Mathematics and Computer Science, Western Washington State College, Bellingham, Washington 98225.

set some notations and hypotheses. We shall assume throughout the paper that the functions f, f_k, and ϕ_j are continuously differentiable, that S is nonempty, and that S_1 has a nonvoid interior.

Notations. We denote by $\langle \cdot\,, \cdot \rangle$ and $|\cdot|$ respectively the Euclidean inner product and norm in Euclidean space R^n. If A is a $n \times n$ matrix, $\|A\|$ denotes the operator norm of A. If F is a real-valued function on R^n, then ∇F and $\nabla^2 F$ denote respectively the gradient of F and the Hessian of F. We shall, in fact, always use ∇ and ∇^2 in place of ∇_x and ∇^2_{xx} respectively; that is, gradients and Hessians are always taken with respect to x. We let $\nabla\phi(x)$ be the matrix whose columns are the gradients $\nabla\phi_1(x), \nabla\phi_2(x), \cdots, \nabla\phi_\nu(x)$.

Next, we define W_m to be the set of all points $(w_0, w_1, \cdots, w_m)$ in R^{1+m} such that $w_0 + w_1 + \cdots + w_m = 1$ and $w_k \geqq 0$ for each k. W_m is the set of all "multiplier candidates" for problems P_t. Given $t > 0$, we define functions G and L_t on $R^n \times W_m$ by the formulas $G(x, w) = \sum_{k=1}^m w_k f_k(x)$ and $L_t(x, w) = w_0 f_0(x) + G(x, w)$. We define the function L on $R^n \times R_\nu \times W_m$ by

$$L(x, \lambda, w) = w_0 f(x) + \sum_{j=1}^{\nu} \lambda_j \phi_j(x) + G(x, w).$$

The functions L_t and L are Lagrangians for problems P_t and P respectively. And, given $t > 0$ and x in S_1, we define the set $\Lambda_t(x)$ of "optimal multipliers" to consist of all w in W_m such that $G(x, w) = 0$ and $\nabla L_t(x, w) = 0$. Given x in S_1, we define the set $A(x)$ of active constraints as follows: $A(x) = \{k \geqq 1 : f_k(x) = 0\}$. And, for w in W_m, we put $K(w) = \{k \geqq 1 : w_k > 0\}$.

Finally, following Hestenes [5, p. 25], we let $C(x)$ denote the tangent cone of S at x. Accordingly, $C(x)$ is the set of all *unit* vectors u in R^n for which there exists a sequence $\{x_q\}_{q=1}^\infty$ in S such that $x_q \neq x$ for each q, such that $\{x_q\}$ converges to x, and such that $\{|x_q - x|^{-1}(x_q - x)\}$ converges to u; we are making the (harmless) abuse of identifying a set of unit vectors as a cone.

Similarly, we let $C_1(x)$ denote the tangent cone of S_1 at x.

HYPOTHESIS I. Assume that problem P has a unique solution $\hat{x}$. Assume also that there exists a unique pair $(\hat{\lambda}, \hat{w})$ in $R^\nu \times W_m$ such that $G(\hat{x}, \hat{w}) = 0$ and $\nabla L(\hat{x}, \hat{\lambda}, \hat{w}) = 0$. Assume moreover that $\hat{w}_0 > 0$, that the matrix $\nabla\phi(\hat{x})$ has rank ν, and that

$$f(x) \to \infty \quad \text{as } |x| \to \infty. \tag{1.1}$$

We conclude this section with two basic results about an exterior penalty function method.

PROPOSITION 1.1. *Assume that Hypothesis* I *is satisfied.*

(i) *Then each problem* P_t *has a solution.*

(ii) *For every neighborhood* U_1 *of* $\hat{x}$ *there exists* $t(U_1) \geqq 0$ *such that* x^t *is in* U_1 *whenever* $t \geqq t(U_1)$ *and* x^t *solves problem* P_t.

Proof. (i) is an immediate consequence of (1.1) and the inequality $f_0 \geqq f$.

Now suppose U_1 is a neighborhood of $\hat{x}$ for which the conclusion in (ii) is false. Then there is a sequence $\{t_q\}_{q=1}^\infty$ such that $\lim t_q = +\infty$ and such that, for each q, some solution y_q to problem P_{t_q} does not lie in U_1. By (1.1) the sequence $\{y_q\}$ is bounded. We may assume that $\{y_q\}$ converges to $\hat{y}$. Clearly, $\hat{y} \neq \hat{x}$ and $\hat{y}$ is in S_1.

If we can now show that $\hat{y}$ must solve problem P, we shall have a contradiction to Hypothesis I. Now, for each q,

$$f(\hat{x}) \geqq f(y_q) + \tfrac{1}{2} t_q \sum_{j=1}^{\nu} \phi_j(y_q)^2. \tag{1.2}$$

Taking the limit on q in (1.2), we find that $\hat{y}$ is in S.

Finally, let x be any element of S. Since $f(y_q) \leqq f_0(y_q) \leqq f_0(x) = f(x)$, we see that $f(\hat{y}) \leqq f(x)$. Hence, $\hat{y}$ would have to be a solution to problem P.

PROPOSITION 1.2. *Assume that Hypothesis* I *is satisfied. For* $t > 0$, *let* x^t *solve problem* P_t. *Suppose* w^t *is in* $\Lambda_t(x^t)$. *Then*

(i) $\lim_{t\to\infty} w_0^t t\phi_j(x^t) = \hat{\lambda}_j$ *for each* $j = 1, \cdots, \nu$

and

(ii) $\lim_{t\to\infty} w^t = \hat{w}$.

Proof. We have, by assumption,

$$0 = w_0^t \nabla f(x^t) + w_0^t \sum_{j=1}^{\nu} t\phi_j(x^t) \nabla \phi_j(x^t) + \nabla G(x^t, w^t). \tag{1.3}$$

For each $t > 0$, let $\psi^t \in R^\nu$ have components $w_0^t \phi_1(x^t)$, $w_0^t t\phi_2(x^t), \cdots, w_0^t t\phi_\nu(x^t)$. Then (1.3) becomes

$$0 = b^t + \nabla \phi(x^t) \psi^t, \tag{1.4}$$

where

$$b^t = w_0^t \nabla f(x^t) + \nabla G(x^t, w^t). \tag{1.5}$$

hence, for t sufficiently large, we have, in view of Proposition 1.1,

$$\psi^t = -\{\nabla \phi(x^t)^T \nabla \phi(x^t)\}^{-1} \nabla \phi(x^t)^T b^t. \tag{1.6}$$

Now suppose that (ii) is false. Hence there exists $\{t_q\}_{q=1}^{\infty}$ such that $\lim_{q\to\infty} t_q = +\infty$ and $\{w^{t_q}\}$ converges to some w in W_m for which $w \neq \hat{w}$. From (1.6), we see that $\{\psi^{t_q}\}$ must converge to some λ in R^ν. Then, from (1.4) and (1.5), we get

$$0 = w_o \nabla f(\hat{x}) + \nabla G(\hat{x}, w) + \sum_{j=1}^{\nu} \lambda_j \nabla \phi_j(\hat{x}) = \nabla L(\hat{x}, \lambda, w).$$

Since $G(x^t, w^t) = 0$ for each t, we have $G(\hat{x}, w) = 0$. It follows from one of the uniqueness provisions of Hypothesis I that $w = \hat{w}$ and $\lambda = \hat{\lambda}$. We have a contradiction and so (ii) is proved.

Now, returning to (1.6), we see that $\bar{\lambda} = \lim \psi^t$ exists. From (1.4) and (1.5), we see again that $\nabla L(\hat{x}, \bar{\lambda}, \hat{w}) = 0$ and so, again, by uniqueness, $\bar{\lambda} = \hat{\lambda}$.

2. The algorithm. Throughout this section, we shall hold the penalty coefficient t fixed.

DEFINITIONS. Given x and y in S_1, we put

$$d(x, y) = \max\{f_0(x) - f_0(y), f_1(x), f_2(x), \cdots, f_m(x)\}$$

and

$$R_t(x) = \{I_n + t\nabla \phi(x) \nabla \phi(x)^T\}^{-1}, \tag{2.1}$$

where I_n denotes the $n \times n$ identity matrix.

Remarks. If $\nabla\phi(x)$ has rank ν then $\nabla\phi(x)^T\nabla\phi(x)$ is nonsingular and we have the formula

$$R_t(x) = I_n - \nabla\phi(x)\left\{\frac{1}{t}I_m + \nabla\phi(x)^T\nabla\phi(x)\right\}^{-1}\nabla\phi(x)^T. \tag{2.2}$$

Thus, in (2.2) we must invert a $\nu\times\nu$ matrix, which is presumably not too ill-conditioned, whereas, in (2.1), an ill-conditioned $n\times n$ matrix must be inverted.

Luenberger [7, p. 291] uses the matrix R_t in the design of a first-order quasi-Newton algorithm for minimizing unconstrained exterior penalty functions. The matrix R_t is used to counteract the potential ill-conditioning caused by the penalty coefficient.

ALGORITHM FOR PROBLEM P_t $(t \geqq 0)$.

Step 0. Choose x_0 in S_1 and set $i = 0$. Choose β in the open interval $(0, 1)$.

Step 1. Obtain a solution h_i^0 in R^1 and h_i in R^n to the convex quadratic programming problem $QP_t(x_i)$:

$$\text{minimize } h^0 + \tfrac{1}{2}\langle h, R_t(x_i)h\rangle$$

subject to

$$h^0 \in R^1, \quad h \in R^n, \quad \langle R_t(x_i)\nabla f_0(x_i), h\rangle \leqq h^0$$

and

$$f_k(x_i) + \langle R_t(x_i)\nabla f_k(x_i), h\rangle \leqq h^0 \quad \text{for } k = 1, \cdots, m.$$

Step 2. If $h_i^0 = 0$, stop. Otherwise, set $v_i = R_t(x_i)h_i$. Compute $\mu_i > 0$ according to this subprocedure:

step (a). Set $\mu = 1$;

step (b). If $d(x_i + \mu v_i, x_i) \leqq \frac{1}{2}\mu h_i^0$, then set $\mu_i = \mu$ and go to Step 3; otherwise, go to step (c);

step (c). Replace μ by $\beta\mu$ and go to step (b).

Step 3. Set $x_{i+1} = x_i + \mu_i v_i$, increase i by 1, and go to Step 1.

Remarks. It should be pointed out that, in practice, one would solve the dual of problem $QP_t(x_i)$ rather than the primal problem itself. The dual has only one constraint; in fact, the dual problem is this:

$$\text{minimize } G(x_i, w) - \tfrac{1}{2}\langle \nabla L_t(x_i, w), R_t(x_i)\nabla L_t(x_i, w)\rangle \quad \text{over } w \text{ in } W_m.$$

Also, it should be observed that the linear search procedure in Step 2 is due originally to Armijo [1]. When there are no equality constraints, the algorithm reduces to the Pironneau–Polak method of centers [10] with an Armijo linear search. But it would not quite be the same as the version given in [9] by Pironneau and Polak, for, in [9], Pironneau and Polak use in Step 2(b) the condition "if $d(x_i + \mu v_i, x_i) \leqq \frac{1}{2}(h_i^0 + \langle h_i, R_t(x_i)h_i\rangle)\ldots$" in place of "If $d(x_i + \mu v_i, x_i) \leqq \frac{1}{2}h_i^0\ldots$". The author knows of no decisive result which would cause one to prefer one of these slightly differing versions to the other.

Next, we present some basic results about the algorithm.

LEMMA 2.1. *Let $\{y_j\}_{j=1}^{\infty}$ be a sequence in S_1 which converges to y_0. For each $j \geqq 0$, let h_j^0 and h_j be optimal for $\mathrm{QP}_t(y_j)$, where $\mathrm{QP}_t(\cdot)$ is the problem described in Step 1 of the algorithm. Assume that $\{h_j^0\}_{j=1}^{\infty}$ converges to $\bar{h}^0$ and $\{h_j\}_{j=1}^{\infty}$ converges to $\bar{h}$. Then $\bar{h}^0$ and $\bar{h}$ are feasible and optimal for problem $\mathrm{QP}_t(y_0)$.*

Proof. Let $\varepsilon > 0$ be given. Because R_t, f and ∇f_k are all continuous, it follows that $h_0^0 + \varepsilon$ and h_0 are feasible for problem $\mathrm{QP}_t(y_t)$, provided j is large. Therefore

$$h_j^0 + \tfrac{1}{2}\langle h_j, R_t(y_j)h_j\rangle \leqq h_0^0 + \varepsilon + \tfrac{1}{2}\langle h_0, R_t(y_j)h_0\rangle \quad \text{for large } j. \tag{2.3}$$

From (2.3), we get, letting j approach infinity,

$$\bar{h}^0 + \tfrac{1}{2}\langle \bar{h}, R_t(y_0)\bar{h}\rangle \leqq h_0^0 + \varepsilon + \tfrac{1}{2}\langle h_0, R_t(y_0)h_0\rangle, \qquad \varepsilon > 0. \tag{2.4}$$

Now $\bar{h}^0$ and $\bar{h}$ are clearly feasible for problem $\mathrm{QP}_t(y_0)$; in view of (2.4), they are also optimal.

Remarks. If we apply the Kuhn–Tucker theorem to problem $\mathrm{QP}_t(x_i)$, we find that nonnegative numbers $\alpha_{i0}, \alpha_{i1}, \cdots, \alpha_{im}$ exist so that

$$\alpha_{i0} + \alpha_{i1} + \cdots + \alpha_{im} = 1, \qquad R_t(x_i)h_i + \sum_{k=0}^{m} \alpha_{ik} R_t(x_i)\nabla f_k(x_i) = 0, \tag{2.5}$$

$$\alpha_{i0}\{\langle R_t(x_i)\nabla f_0(x_i), h_i\rangle - h_i^0\} = 0 \tag{2.6}$$

and

$$\alpha_{ik}\{f_k(x_i) + \langle R_t(x_i)\nabla f_k(x_i), h_i\rangle - h_i^0\} = 0 \quad \text{for } k = 1, \cdots, m. \tag{2.7}$$

For each i, define α_i in W_m to have components $\alpha_{i0}, \cdots, \alpha_{im}$. From (2.5)–(2.7), we get

$$G(x_i, \alpha_i) - \langle h_i, R_t(x_i)h_i\rangle = h_i^0. \tag{2.8}$$

CONVERGENCE THEOREM. *Suppose that the algorithm stops at $\hat{y} = x_i$ or that $\hat{y}$ is a limit point of an infinite sequence $\{x_i\}_{i=0}^{\infty}$ generated by the algorithm. Then the zero solution is optimal for problem $\mathrm{QP}_t(\hat{y})$. Moreover, the John conditions hold at $\hat{y}$; i.e., the set $\Lambda_t(\hat{y})$ is nonvoid.* (Here, we maintain the assumptions made at the start of the paper. Moreover, we assume $\{x_i\}_{i=0}^{\infty}$ is bounded.)

Remarks. A direct proof can be given or the proof can be made to fit one of the models discussed in Polak [11].

The proof is very similar to the proof of Theorem 3.12 in [9].

LEMMA 2.2. *Assume that the algorithm generates a sequence $\{x_i\}_{i=0}^{\infty}$ which converges to a solution x^t of problem P_t. Every limit point of $\{\alpha_i\}_{i=0}^{\infty}$ is in $\Lambda_t(x^t)$. Also, there is a positive integer i_1 such that $\alpha_{ik} = 0$ whenever $i \geqq i_1$ and k is a positive integer not in $A(x^t)$.*

Proof. By the convergence theorem, $\hat{h}^0 = 0$ and $\hat{h} = 0$ solve problem $\mathrm{QP}_t(x^t)$. Moreover, this problem admits no other solution. For, if $\bar{h}^0$ and $\bar{h}$ were also optimal for $\mathrm{QP}_t(x^t)$, then, proceeding as in the derivation of (2.5)–(2.8), we could obtain $\bar{w}$ in W_m such that

$$G(x^t, \bar{w}) - \tfrac{1}{2}\langle \bar{h}, R_t(x^t)\bar{h}\rangle = \bar{h}^0 + \tfrac{1}{2}\langle \bar{h}, R_t(x^t)\bar{h}\rangle = 0; \tag{2.9}$$

since $G(x^t, \bar{w}) \leqq 0$ and $\langle \bar{h}, R_t(x^t)\bar{h}\rangle \geqq 0$, we would then get from (2.9) the equation

$\langle \bar{h}, R_t(x^t)\bar{h}\rangle = 0$; since $R_t(x)$ is positive-definite it would follow that $\bar{h} = 0$ and hence that $\bar{h}^0 = 0$.

Now, according to (2.5), the sequence $\{h_i\}$ is bounded. It follows from (2.8) that $\{h_i^0\}$ is also bounded. By Lemma 2.1 (and the fact that $QP_t(x^t)$ has but one solution), we see that $\{|h_i|\}$ and $\{h_i^0\}$ both converge to 0.

If $k \geqq 1$ and if k is not in $A(x^t)$, it follows now from (2.7) that $\alpha_{ik} = 0$ for all large i. Moreover, since $R_t(x^t)$ is nonsingular, it follows from (2.5)–(2.7) that any limit point of $\{\alpha_i\}$ is in $\Lambda_t(x^t)$.

3. Rate of convergence. In this section, we shall show that the algorithm just presented will often generate sequences $\{x_i\}_{i=0}^{\infty}$ which converge at a linear rate. In fact, it will be shown that if t is large enough and if x^t solves problem P_t then $\limsup_{i\to\infty} [f_0(x_{i+1}) - f_0(x^t)]/[f_0(x_i) - f_0(x^t)] \leqq \theta(t)$, where $\lim_{t\to\infty} \theta(t) < 1$.

Before presenting the technical development, we shall provide an intuitive, preliminary discussion. Under proper hypotheses, including second-order conditions, we establish three main lemmas. We show in Lemma 3.2 that there exists $M_1 > 0$ such that $\|R_t(x)\nabla^2 f(x)\| \leqq M_1$ for each t, so long as x is sufficiently close to a solution x^t to problem P_t; this inequality shows how the matrix $R_t(x)$ counteracts the ill-conditioning caused by the large eigenvalues of $\nabla^2 L_t$. Then, we prove in Lemma 3.3 that

$$f_0(x_{i+1}) - f_0(x_i) \leqq \frac{\beta}{M}(h_i^0 + \tfrac{1}{2}\langle h_i, R_t(x_i)h_i\rangle)$$

or

$$f_0(x_{i+1}) - f_0(x_i) \leqq \tfrac{1}{2}h_i^0,$$

so that we may be certain that the move in the direction v_i is a comparatively good one. Then we show in Lemma 3.4 that, in turn, the quantity $h_i^0 + \frac{1}{2}\langle h_i, R_t(x_i)h_i\rangle$ is "sufficiently negative" relative to $f_0(x^t) - f_0(x_i)$. It is easy to combine these last two lemmas to prove that $\{f_0(x_i)\}_{i=1}^{\infty}$ is Q-linearly convergent to $f_0(x^t)$ as mentioned above. The section concludes with another theorem from which it follows that $|x_i - x^t|$ and $\{t|\nabla\phi(x_i)^T(x_i - x^t)|^2\}$ are both R-linearly convergent to 0, with the rate being again asymptotically independent of t. The terms "Q-linearly convergent" and "R-linearly convergent" are used as in [8].

Now we state some lengthy, but not unusual, hypotheses.

HYPOTHESIS II. Assume Hypothesis I holds. Assume that strict complementarity holds at $\hat{x}$; i.e., assume $K(\hat{w}) = A(\hat{x})$. Moreover assume that $\hat{x}$ is a regular point of S [5, p. 29]; i.e., assume that $C(\hat{x})$ is the set of all unit vectors u such that $\langle \nabla f_k(\hat{x}), u\rangle \leqq 0$ whenever k is in $A(\hat{x})$ and $\langle \nabla\phi_j(\hat{x}), u\rangle = 0$ for each $j = 1, \cdots, \nu$.

HYPOTHESIS III. Assume Hypothesis I holds and that the functions $f, f_1, \cdots, f_m, \phi_1, \cdots, \phi_\nu$ are twice continuously differentiable on some neighborhood U_1 of $\hat{x}$. Assume moreover that a number m_0 exists in the interval $(0, 1]$ for which $\langle u, \nabla^2 L(\hat{x}, \hat{\lambda}, \hat{w})u\rangle > m_0$ whenever u is a unit vector in $C(\hat{x})$ and $\langle \nabla f_k(\hat{x}), u\rangle = 0$ for all k in $K(\hat{w})$.

Remarks. If Hypothesis III holds, there are neighborhoods U_2, V_2, and W_2 of $\hat{x}, \hat{\lambda}$ and $\hat{w}$ respectively such that $\nabla\phi(x)$ has rank ν whenever x is in U_2 and such that $\langle u, \nabla^2 L(x, \lambda, w)u\rangle > m_0$ whenever x is in U_2, λ is in V_2, w is in W_2, u is a unit vector in $C(\hat{x})$, and $\langle \nabla f_k(\hat{x}), u\rangle = 0$ for every k in $K(\hat{w})$.

HYPOTHESIS IV. Assume that Hypotheses I and III hold and that U_1 is as in Hypothesis III. Assume that a number $t_1 \geqq 0$ exists such that for every $t \geqq t_1$ it is true that the algorithm generates a sequence which converges to a point x^t in U_1 which solves problem P_t.

Remarks. With the listing of the hypotheses now complete, we shall proceed to the discussion of five lemmas. We let H_t denote the Hessian of the Lagrangian L_t for problem P_t. We have

$$H_t(x, w) = w_0 \nabla^2 f(x) + \nabla^2 G(x, w) + \sum_{j=1}^{\nu} (t w_0 \phi_j(x)) \nabla^2 \phi_j(x) + t w_0 \nabla \phi(x) \{\nabla \phi(x)\}^T. \tag{3.1}$$

LEMMA 3.1. *Assume Hypotheses* II *and* III *hold. Suppose* $0 < \gamma < 1$. *There exists* $t_0 \geqq 0$ *such that if* $t \geqq t_0$ *and* x^t *solves problem* P_t *then*

$$\langle u, H_t(x^t, w^t) u \rangle \geqq m_0 + \gamma t w_0^t |\nabla \phi(x_t)^T u|^2$$

whenever w^t *is in* $\Lambda_t(x^t)$ *and* u *is a unit vector in* $C_1(x^t)$ *for which* $\langle \nabla f_k(x^t), u \rangle = 0$ *for all* k *in* $K(w^t)$.

Proof. Suppose that the conclusion is false. There are sequences $\{t_p\}_{p=1}^{\infty}$, $\{x^{t_p}\}$, $\{w^{t_p}\}$, and $\{u_p\}$ such that $\lim_{p \to \infty} t_p = +\infty$, x^{t_p} solves problem P_{t_p}, w^{t_p} is in $\Lambda_{t_p}(x^{t_p})$, u_p is a unit vector in $C_1(x^{t_p})$, $\langle \nabla f_k(x^{t_p}), u_p \rangle = 0$ for all k in $K(w^{t_p})$ and yet

$$\langle u_p, H_{t_p}(x^{t_p}, w^{t_p}) u_p \rangle < m_0 + \gamma t_p w_0^{t_p} |\nabla \phi(x^{t_p})^T u_p|^2, \qquad p \geqq 1. \tag{3.2}$$

We may assume that $\{u_p\}$ converges to u.

Now suppose k is in $K(\hat{w}) = A(\hat{x})$. Then $\hat{w}_k > 0$. By Proposition 1.2, $w_k^{t_p} > 0$ for large p and so $\langle \nabla f_k(x^{t_p}), u_p \rangle = 0$. Hence $\langle \nabla f_k(\hat{x}), u \rangle = 0$. In short,

$$\langle \nabla f_k(\hat{x}), u \rangle = 0 \quad \text{for all } k \text{ in } K(\hat{w}) = A(\hat{x}). \tag{3.3}$$

Now, from (3.1) and (3.2), we get

$$\langle u_p, \nabla^2 L(x^{t_p}, \psi^{t_p}, w_0^{t_p}) u_p \rangle + t_p w^{t_p} |\nabla \phi(x^{t_p})^T u_p|^2 < m_0 + \gamma t_p w_0^{t_p} |\nabla \phi(x^{t_p})^T u_p|^2, \tag{3.4}$$

where we have defined ψ^{t_p} as in the proof of Proposition 1.2.

By Proposition 1.2, we derive from (3.4)

$$0 \leqq \limsup_{p \to \infty} (1 - \gamma) t_p w_0^{t_p} |\nabla \phi(x^{t_p})^T u_p|^2 \leqq m_0 - \langle u, \nabla^2 L(\hat{x}, \hat{\lambda}, \hat{w}) u \rangle. \tag{3.5}$$

From (3.5), we deduce, first of all, that $|\nabla \phi(\hat{x})^T u| = 0$ and so

$$\langle \nabla \phi_j(\hat{x}), u \rangle = 0 \quad \text{for } j = 1, \cdots, \nu. \tag{3.6}$$

Since $\hat{x}$ is assumed to be a regular point of S, it follows from (3.3) and (3.6) that u must be a unit vector in $C(\hat{x})$. But then, (3.3) and (3.5) yield a contradiction of Hypothesis III.

LEMMA 3.2. *Assume that Hypothesis* II *is valid. There is a number* $M_1 > 0$ *with this property: Given* $t \geqq 0$, *there exists* $\delta_t > 0$ *such that if* x^t *solves problem* P_t *and* $|x - x^t| < \delta_t$, *then* $\|R_t(x)\nabla^2 f_o(x)\| \leqq M_1$.

Proof. Let U_1 be as in Hypothesis III. Given x in U_1, we find a formula for $R_t(x)\nabla^2 f_0(x)$ as follows (where we have omitted the argument x):

$$R_t(x)\nabla^2 f_0(x) = R_t\left\{\nabla^2 f + \sum_{j=1}^{\nu} t\phi_j \nabla^2 \phi_j + t\nabla\phi\nabla\phi^T\right\}$$

$$= R_t\left\{\nabla^2 f - I_n + R_t^{-1} + \sum_{j=1}^{\nu} t\phi_j \nabla^2 \phi_j\right\}$$

$$= R_t\left\{\nabla^2 f - I_n + \sum_{j=1}^{\nu} t\phi_j \nabla^2 \phi_j\right\} + I_n.$$

By Proposition 1.2, the numbers $t\phi_j(x^t)$, $t \geqq 0$, are uniformly bounded and so the norms of the matrices $R_t(x^t)\nabla^2 f_0(x^t)$ will be uniformly bounded, say by N. Let M_1 be a number larger than N. Then, given $t \geqq 0$, there is a number $\delta_t > 0$ such that if $|x - x^t| < \delta_t$, then the numbers $t\phi_j(x)$ will be close enough to $t\phi_j(x^t)$ that $R_t(x)\nabla^2 f_0(x)$ will be bounded by M_1.

Remark. In the next three lemmas, we shall suppose t to be fixed. Accordingly, we may use the notation $R_i = R_t(x_i)$ without ambiguity.

LEMMA 3.3. *Assume Hypotheses* III *and* IV *and let* M *be chosen so that* $M \geqq 1$, $M > M_1$, *and so that* $\|\nabla^2 f_k(x)\| \leqq M$ *whenever* $1 \leqq k \leqq m$ *and* x *is in* U_1. *Let* $t \geqq \max(t_0, t_1)$ *be fixed. Here* t_0 *and* t_1 *are as in Lemma* 1.3 *and Hypothesis* IV *respectively.*

For each i, *let* x_i *and* μ_i *be as in the algorithm. If* $\mu_i = 1$, *then* $d(x_{i+1}, x_i) \leqq \frac{1}{2}h_i^0$. *There exists an integer* $i_2(t)$ *such that*

$$Md(x_{i+1}, x_i) \leqq \beta(h_i^0 + \tfrac{1}{2}\langle h_i, R_i h_i\rangle),$$

whenever $i \geqq i_2(t)$ *and* $\mu_i < 1$.

Proof. By construction, we must have $d(x_i + \mu_i v_i, x_i) \leqq \frac{1}{2}h_i^0$ in case $\mu_i = 1$. Now, suppose $\mu_i < 1$. Then μ_i is a positive integral power of β. We have

$$d(x_i + \mu_i v_i, x_i) \leqq \tfrac{1}{2}\mu_i h_i^0 \tag{3.7}$$

while

$$f_0(x_i + \mu_i\beta^{-1}v_i) - f_0(x_i) > \tfrac{1}{2}\mu_i\beta^{-1}h_i^0, \tag{3.8}$$

or, for some $k \geqq 1$,

$$f_k(x_i + \mu_i\beta^{-1}v_i) > \tfrac{1}{2}\mu_i\beta^{-1}h_i^0. \tag{3.9}$$

Since $\{x_i\}_{i=0}^{\infty}$ converges to x^t and $\{h_i\}_{i=0}^{\infty}$ converges to 0, then there exists i_3^* such that $x_i + \mu v_i$ is in U_1 and $|x_i + \mu v_i - x^t| < \delta_t$ whenever $i \geqq i_3^*$ and $0 \leqq \mu \leqq 1$; here, δ_t is chosen as in Lemma 3.2.

Now, if (3.9) holds, then, we have, for large i, $\frac{1}{2}\mu_i\beta^{-1}h_i^0 < f_k(x_i) + \mu_i\beta^{-1}\langle\nabla f_k(x_i), v_i\rangle + \frac{1}{2}(\mu_i\beta^{-1})^2\langle v_i, \nabla^2 f_k(x_i + \xi_i\mu_i\beta^{-1}v_i)v_i\rangle$, where $0 < \xi_i < 1$. Since $\mu_i\beta^{-1} \leqq 1$ and $f_k(x_i) \leqq 0$, we get

$$\tfrac{1}{2}\mu_i\beta^{-1}h_i^0 \leqq \mu_i\beta^{-1}h_i^0 + \tfrac{1}{2}(\mu_i\beta^{-1})^2\langle v_i, \nabla^2 f_k(x_i + \xi_i\mu_i\beta^{-1}v_i)v_i\rangle. \tag{3.10}$$

Let Q_i denote the positive-definite square root of R_i and put $S_i = Q_i\nabla^2 f_k(x_i+\xi_i\mu_i\beta^{-1}v_i)Q_i$. We get

$$\langle v_i, \nabla^2 f_k(x_i+\xi_i\mu_i\beta^{-1}v_i)v_i\rangle = \frac{\langle Q_ih_i, S_iQ_ih_i\rangle}{\langle Q_ih_i, Q_ih_i\rangle}\cdot\langle h_i, R_ih_i\rangle.$$

Since $\|Q_i\|\leqq 1$, it follows that $\|S_i\|\leqq M$. Hence, from (3.10), we obtain

$$-\beta h_i^0 \leqq M\mu_i\langle h_i, R_ih_i\rangle \quad \text{for large } i. \tag{3.11}$$

On the other hand, if (3.8) holds, then, for large i, $\frac{1}{2}\mu_i\beta^{-1}h_i^0 < \mu_i\beta^{-1}\langle\nabla f_0(x_i), v_i\rangle + \frac{1}{2}(\mu_i\beta^{-1})^2\langle v_i, \nabla^2 f_0(x_i+\xi_i\mu_i\beta^{-1}v_i)v_i\rangle$, where $0<\xi_i<1$. It follows that

$$-\beta h_i^0 \leqq \mu_i\langle v_i, \nabla^2 f_0(x_i+\xi_i\mu_i\beta^{-1}v_i)v_i\rangle \quad \text{for large } i. \tag{3.12}$$

Setting $T_i = Q_i\nabla^2 f_0(x_i)Q_i$, we have

$$-\beta h_i^0 \leqq \mu_i\langle Q_ih_i, T_iQ_ih_i\rangle + \mu_i\langle v_i, \{\nabla^2 f_0(x_i+\xi_i\mu_i\beta^{-1}v_i)-\nabla^2 f_0(x_i)\}v_i\rangle.$$

Now T_i is similar to $R_i\nabla^2 f_0(x_i)$ and so, by Lemma 3.2,

$$-\beta h_i^0 \leqq \mu_i M_1|R_i^{1/2}h_i|^2 + \mu_i\langle v_i, \{\nabla^2 f_0(x_i+\xi_i\mu_i\beta^{-1}v_i)-\nabla^2 f_0(x_i)\}v_i\rangle. \tag{3.13}$$

Since the final term in (3.13) is $o(|v_i|^2)$ and since $M>M_1$, we see that (3.11) must hold in this case also.

From (3.7) and (3.11), we have

$$d(x_i+\mu_iv_i, x_i) \leqq -\frac{\beta(h_i^0)^2}{2M\langle h_i, R_ih_i\rangle} \quad \text{for large } i. \tag{3.14}$$

By some elementary algebra, we have

$$\begin{aligned}(h_i^0)^2 &= 2|h_i^0|\langle h_i, R_ih_i\rangle - \langle h_i, R_ih_i\rangle^2 + (|h_i^0|-\langle h_i, R_ih_i\rangle)^2\\ &\geqq 2|h_i^0|\langle h_i, R_ih_i\rangle - \langle h_i, R_ih_i\rangle^2\end{aligned}$$

and so

$$\frac{(h_i^0)^2}{\langle h_i, R_ih_i\rangle} \geqq 2|h_i^0| - \langle h_i, R_ih_i\rangle = -2h_i^0 - \langle h_i, R_ih_i\rangle. \tag{3.15}$$

If we combine (3.14) and (3.15), we get,

$$d(x_i+\mu_iv_i, x_i) \leqq \frac{\beta}{M}(h_i^0 + \tfrac{1}{2}\langle h_i, R_ih_i\rangle \quad \text{for large } i.$$

LEMMA 3.4. *Assume Hypotheses* II, III *and* IV *hold. Suppose* $t\geqq t_0$, *where* t_0 *is as in Lemma* 1.3, *and* $t\geqq t_1$, *where* t_1 *is as in Hypothesis* IV. *Suppose* γ *is in* (0, 1). *Let* $s_t = \min\{w_0^t: w^t\in\Lambda_t(x^t)\}$ *and assume* $s_t>0$. *Set* $m_t = \min(s_t, m_0)$. *Then there exists a positive integer* $i_4(t,\gamma)$ *such that*

$$h_i^0 + \tfrac{1}{2}\langle h_i, R_ih_i\rangle \leqq -m_ts_t\gamma^2\{f_0(x_i)-f_0(x^t)\}, \qquad i\geqq i_4(t,\gamma).$$

Proof. Since $0<m_t\leqq 1$ and $G(x_i,\alpha_i)\leqq 0$, we see from (2.8) that

$$\frac{1}{m_t\gamma}\{h_i^0+\tfrac{1}{2}\langle h_i, R_ih_i\rangle\} \leqq G(x_i,\alpha_i) - \frac{1}{2m_t\gamma}\langle h_i, R_ih_i\rangle. \tag{3.16}$$

If we now add $(1/(2m_t\gamma))|R_i^{1/2}h_i + m_t\gamma R_i^{-1/2}2(x_i - x^t)|^2$ to the right member of (3.16), we get

$$\frac{1}{m_t\gamma}\{h_i^0 + \tfrac{1}{2}\langle h_i, R_i h_i\rangle\} \leqq G(x_i, \alpha_i) + \langle h_i, x_i - x^t\rangle\rangle + \frac{m_t\gamma}{2}\langle x_i - x^t, R_i^{-1}(x_i - x^t)\rangle\rangle.$$

But, from (2.1) and (2.5), we then obtain

$$\frac{1}{m_t\gamma}\{h_i^0 + \tfrac{1}{2}\langle h_i, R_i h_i\rangle\} \leqq G(x_i, \alpha_i) - \langle \nabla L_t(x_i, \alpha_i), x_i - x^t\rangle + \frac{m_t\gamma}{2}|x_i - x^t|^2 + \frac{m_\gamma\gamma t}{2}|\nabla\phi(x_i)^T(x_i - x^t)|^2. \tag{3.17}$$

Now suppose that there exists an infinite subsequence K of the positive integers such that

$$G(x_q, \alpha_q) - \langle \nabla L_t(x_q, \alpha_q), x_q - x^t\rangle + \frac{m_t\gamma}{2}\{|x_q - x^t|^2 + t|\nabla\phi(x_q)^T(x_q - x^t)|^2\} > -s_t\gamma\{f_0(x_q) - f_0(x^t)\}, \qquad q \in K. \tag{3.18}$$

In view of (3.17), the proof would be complete if we could show that (3.18) leads to a contradiction.

For each q in K, define u_q by $|x_q - x^t|u_q = x_q - x^t$. We may assume that $\{u_q : q \in K\}$ converges to u in $C_1(x^t)$, and, by Lemma 2.2, we may assume that $\{\alpha_q\}$ converges to w^t in $\Lambda_t(x^t)$. Let i_1 be as in Lemma 2.2. For $q \geqq i_1$, we have $G(x^t, \alpha_q) = 0$; hence, from (3.18), we get

$$L_t(x_q, \alpha_q) - L_t(x^t, \alpha_q) - \langle \nabla L_t(x_q, \alpha_q), x_q - x^t\rangle + \frac{m_t\gamma}{2}\{|x_q - x^t|^2 + t|\nabla\phi(x_q)^T(x_q - x^t)|^2\} > (\alpha_{q0} - s_t\gamma)\{f_0(x_q) - f_0(x^t)\}, \qquad q \geqq i_1. \tag{3.19}$$

If we divide both sides of (3.19) by $|x_q - x^t|$ and let q approach infinity, we obtain

$$\langle \nabla L_t(x^t, w^t), u\rangle - \langle \nabla L_t(x^t, w^t), u\rangle \geqq (w_0^t - s_t\gamma)\langle \nabla f_0(x^t), u\rangle. \tag{3.20}$$

Now $\nabla L_t(x^t, w^t) = 0$. Hence, if $\langle \nabla f_k(x^t), u\rangle < 0$ for some k in $K(w^t)$, then we should have $\langle \nabla f_0(x^t), u\rangle > 0$; but this would contradict (3.20).

By Lemma 3.1, we have

$$\langle u, H_t(x^t, w^t)u\rangle \geqq m_0 + \gamma t w_0^t|\nabla\phi(x^t)^T u|^2. \tag{3.21}$$

In view of (3.19) and Lemma 2.2, we have

$$-\int_0^1 (1-s)\langle x_q - x^t, H_t(sx^t + (1-s)x_q, \alpha_q)(x_q - x^t)\rangle\, ds + \frac{m_t\gamma}{2}\{|x_q - x^t|^2 + t|\nabla\phi(x_q)^T(x_q - x^t)|^2\} > 0 \quad \text{for large } q. \tag{3.22}$$

If we divide both sides of (3.22) by $|x_q - x^t|^2$ and then let q approach infinity, we get

$$\frac{m_t\gamma}{2}\{1 + t|\nabla\phi(x^t)^T u|^2\} \geqq \tfrac{1}{2}\langle u, H_t(x^t, w^t)u\rangle,$$

and so, by (3.21),

$$m_t\gamma\{1 + t|\nabla\phi(x^t)^T u|^2 \geqq m_0 + \gamma t w_0^t |\nabla\phi(x^t)^T u|^2. \tag{3.23}$$

Since $m_t\gamma < m_0$ and $m_t \leqq s_t \leqq w_0^t$, we find that (3.23) cannot possibly hold. We have obtained a contradiction.

LEMMA 3.5. *Under the assumptions made in Lemma* 3.4, *there exists a positive integer* $i_5(t, \gamma)$ *such that*

$$h_i^0 \leqq -2p_t s_t \gamma^2\{f_0(x_i) - f_0(x^t)\}, \qquad i \geqq i_5(t, \gamma):$$

here, $p_t = \min(\frac{1}{2}, m_t)$.

Proof. There is little difference between this proof and the preceding one. In place of (3.16), we have

$$\frac{1}{p_t\gamma} h_i^0 \leqq 2G(x_i, \alpha_i) - \frac{1}{p_t\gamma}\langle h_i, R_i h_i\rangle;$$

Then, in place of (3.17), we get

$$\frac{1}{p_t\gamma} h_i^0 \leqq 2G(x_i, \alpha_i) - 2\langle \nabla L_t(x_i, \alpha_i), x_i - x^t\rangle + p_t\gamma\{|x_i - x^t|^2 + t|\nabla\phi(x_i)^T(x_i - x^t)|^2\}.$$

The counterpart to (3.18) is

$$2G(x_q, \alpha_q) - 2\langle \nabla L_t(x_q, \alpha_q), x_q - x^t\rangle + p_t\gamma\{|x_q - x^t|^2 + t|\nabla\phi(x_q)^T(x_q - x^t)|^2\} > -2s_t\gamma\{f_0(x_q) - f_0(x^t)\}, \qquad q \in K.$$

The rest of the proof proceeds as before.

THEOREM 3.1. *Assume Hypotheses* I–IV, *suppose* M *is as in Lemma* 3.3, *and* $s_t = \min\{w_0^t: w^t \in \Lambda_t(x^t)\} > 0$. *As in Lemmas* 3.4 *and* 3.5, *put* $m_t = \min(s_t, m_0)$ *and* $p_t = \min(\frac{1}{2}, m_t)$. *For* $t \geqq \max(t_0, t_1)$, *define*

$$\theta(t) = \max\left\{1 - p_t s_t, 1 - \frac{\beta m_t s_t}{M}\right\}.$$

Then, for $t \geqq \max(t_0, t_1)$,

$$\limsup_{i\to\infty} \frac{f_0(x_{i+1}) - f_0(x^t)}{f_0(x_i) - f_0(x^t)} \leqq \theta(t). \tag{3.24}$$

Remarks. Observe $\lim_{t\to\infty}\theta(t) = \max(1 - m^{**}\hat{w}_0, 1 - \beta m^*\hat{w}_0/M)$, where $m^* = \min(\hat{w}_0, m_0)$ and $m^{**} = \min(\frac{1}{2}, m^*)$. Hence the rate of convergence of $\{f_0(x_i)\}_{i=0}^{\infty}$ is independent of t, for large t.

If we were to replace the Armijo linear search in Step 2 of the algorithm by an exact linear search, we could obtain (3.24) with $\theta(t)$ taken to be $1 - m_t s_t/M$.

Proof of Theorem 3.1. Let γ in $(0, 1)$ and $t \geqq \max(t_0, t_1)$ be given. Let $i_6(t, \gamma)$ be the largest of the integers $i_2(t)$, $i_4(t, \gamma)$, and $i_5(t, \gamma)$, which are given in the three

preceding lemmas. Suppose $i \geqq i_6(t, \gamma)$. If $\mu_i = 1$, we get

$$f_0(x_{i+1}) - f_0(x_i) \leqq d(x_{i+1}, x_i) \leqq \tfrac{1}{2} h_i^0 \leqq -p_t s_t \gamma^2 \{f_0(x_i) - f_0(x^t)\}. \tag{3.25}$$

If $\mu_i < 1$, we get

$$f_0(x_{i+1}) - f_0(x_i) \leqq \frac{\beta}{M}(h_i^0 + \tfrac{1}{2}\langle h_i, R_i h_i\rangle) \leqq -\frac{\beta m_t s_t \gamma^2}{M}\{f_0(x_i) - f_0(x^t)\}. \tag{3.26}$$

If we add $f_0(x_i) - f_0(x^t)$ to both sides of (3.25) and (3.26), it is clear that (3.24) will follow.

THEOREM 3.2. *Assume that all of the hypotheses of Theorem* 3.1 *are satisfied. Let* $\theta(t)$ *be defined as in Theorem* 3.1. *Then,*

$$\limsup_{i \to \infty} \sqrt[i]{m_0 |x_i - x^t|^2 + t w_0^t |\nabla \phi(x_i)^T (x_i - x^t)|^2} \leqq \theta(t).$$

Remarks. This theorem is an immediate consequence of Theorem 3.1 and the lemma which follows.

We see, in particular, from Theorem 3.2, that $\{x_i\}$ is R-linearly convergent to x^t, at a rate which is essentially independent of t, provided that t is large.

LEMMA 3.6. *Assume the Hypotheses* II *and* III *hold. Suppose* $0 < \gamma < 1$ *and let* t_0 *again be as in Lemma* 3.1. *Assume that* x^t *is a solution to problem* P_t, *for* $t \geqq t_0$. *Then there is a neighborhood* U_{3t} *of* x^t *such that*

$$2 w_0^t \{f_0(x) - f_0(x^t)\} > m_0 \gamma |x - x^t|^2 + \gamma t w_0^t |\nabla \phi(x)^T (x - x^t)|^2$$

whenever w^t *is in* $\Lambda_t(x^t)$, x *is in* $S_1 \cap U_{3t}$, *and* $x \neq x^t$.

Proof. Suppose that the conclusion is false. Then, for every positive integer q, there exists x_q in S_1 so that $0 < |x_q - x^t| < 1/q$ and

$$w_0^t \{f_0(x_q) - f_0(x^t)\} \leqq \left(\frac{m_0 \gamma}{2} + \frac{1}{q}\right)|x_q - x^t|^2 + \frac{1}{2}\gamma t w_0^t |\nabla \phi(x_q)^T (x_q - x^t)|^2.$$

For each q, define u_q by $|x_q - x^t| u_q = x_q - x^t$. We may assume that $\{u_q\}_{q=1}^{\infty}$ converges to a unit vector u in $C_1(x^t)$. Since $G(x^t, w^t) = 0$ and $\nabla L_t(x^t, w^t) = 0$, we have

$$\begin{aligned} &L_t(x_q, w^t) - L_t(x^t, w^t) - \langle \nabla L_t(x^t, w^t), x_q - x^t\rangle - G(x_q, w^t) \\ &\qquad \leqq \left(\frac{m_0 \gamma}{2} + \frac{1}{q}\right)|x_q - x^t|^2 + \frac{1}{2}\gamma t w_0^t |\nabla \phi(x_q)^T (x_q - x^t)|^2 \end{aligned} \tag{3.27}$$

Now we divide both sides of (3.27) by $|x_q - x^t|^2$ and take the limit as q approaches infinity; we obtain

$$\frac{1}{2}\langle u, H_t(x^t, w^t) u\rangle + \limsup_{q \to \infty} \frac{-G(x_q, w^t)}{|x_q - x^t|^2} \leqq \frac{m_0 \gamma}{2} + \frac{1}{2}\gamma t w_0^t |\nabla \phi(x^t)^T u|^2. \tag{3.28}$$

Since each x_q is in S_1, each number $-G(x_q, w^t)$ is nonnegative, and so, by (3.28),

the sequence $\{-G(x_q, w^t)/|x_q - x^t|^2\}_{q=1}^{\infty}$ is bounded. Consequently,

$$0 = \lim_q \frac{G(x_q, w^t)}{|x_q - x^t|} = \lim_q \frac{G(x_q, w^t) - G(x^t, w^t)}{|x_q - x^t|} = \sum_{k=1}^{m} w_k^t \langle \nabla f_k(x^t), u \rangle.$$

Since $\langle \nabla f_k(x^t), u \rangle \leqq 0$ for every k in $K(w^t)$, we see that $\langle \nabla f_k(x^t), u \rangle = 0$ for every k in $K(w^t)$. Hence, by Lemma 3.1,

$$\langle u, H_t(x^t, w^t)u \rangle \geqq m_0 + \gamma t w_0^t |\nabla \phi(x^t)^T u|^2. \tag{3.29}$$

Since each $-G(x_q, w^t)$ is nonnegative and $m_0\gamma < m_0$, we see that (3.29) and (3.28) are incompatible.

Remark. The proof of Lemma 3.6 is a variant of the proof by Hestenes [5, p. 37], which contains the essence of the finite-dimensional indirect method. A slightly sharper form of Hestenes' theorem is given by Chaney [3, Lem. 3.1]. Lemma 3.6 is a sufficiency theorem for a local minimum.

4. Numerical results. The algorithm presented in § 2 has been used to obtain numerical solutions to several problems of small dimension. The two examples discussed here are representative.

In each case, we found "approximate" solutions to problems P_t for successively larger values of t. A point x_i was accepted as an approximate solution as soon as $|h_i|^2 < \varepsilon$, where ε was a certain positive number depending upon t. In particular, we began with $t = 20$ and then took, in order, $t = 100$, $t = 500$, and $t = 2500$. At each transition, t was therefore increased by a factor of 5. Moreover, at each transition, ε was *reduced* by a factor of 5. And, after the transition, the approximate solution for the previous problem became the x_0 for the new problem.

Example 4.1. The first problem considered was the following:

$$\text{minimize } x_2^2 + x_2 + 2x_4^2 + 4x_5$$

subject to

$$6 \leqq -x_1^2 + x_2^2 + 2x_3^2 + x_4^2 + x_5^2, \qquad 3 \leqq 5x_5 + x_3 x_5$$

and

$$x_1^2 + 2x_2^2 + x_2 = 5, \qquad x_3^2 + x_4^2 = 4.$$

Now the solution to this problem, written as a row vector, is $(2.0, -1.0, -2.0, 0, 1.0)$. In one experimental "run" on this problem, we chose to begin from the initial point $x_0 = (1.5, 0, -1.5, 0.5, 2.0)$. The results are summarized in Table 1.

TABLE 1

t	ε	Number of iterations	Final value of $\lvert h_i \rvert^2$
20	.01000	17	.00843773
100	.00200	8	.00165843
500	.00040	10	.00024239
2500	.00008	11	.00007297

The final point, obtained after 46 iterations in all, was (2.00000503, −0.99996611, −2.00001019, 0.0, 1.00000380).

Example 4.2. Another problem considered was this:

$$\text{minimize } x_1^2+x_2^2+x_3^2+x_1x_2+3x_2x_3$$

subject to

$$9\leqq 3x_1^2+2x_2^2+x_3^2, \qquad x_2^2-x_3\leqq 0$$

and

$$x_1^2+2x_2^2=6.$$

The choice of the initial feasible point, as required in Step 0 was (1.0, 1.5, 2.5).

This problem was chosen with less contrivance than the previous one, inasmuch as the author devised it with little thought and with no attempt to gain prior knowledge of the solution.

The results are summarized according to the pattern set in the previous example. See Table 2.

TABLE 2

t	ε	Number of iterations	Final value of $\lvert h_i\rvert^2$
20	.01000	42	.00326915
100	.00200	7	.00120892
500	.00040	8	.00016816
2500	.00008	9	.00002097

The final point, obtained after 66 iterations in all, was (0.55125710, −1.68765049, 2.84816566).

5. Concluding remarks. There are other methods for solving nonlinear programming problems, which might suitably be termed "hybrid". We shall not attempt to compare the present method with others in any detail. We shall merely mention two other methods, these two being the combined penalty function and gradient projection method of Luenberger [6] and the extensively developed method of multipliers, which was originated by Hestenes and by Powell. These two methods do have several features in common with the algorithm presented in this paper: Exterior penalty functions are used in conjunction with another technique, and an effort is made to nullify the ill-conditioning associated with the use of exterior penalty functions.

To solve problem P by the method of multipliers one can form the augmented Lagrangian

$$K(x, w, \lambda, z, t)=f(x)+\sum_{k=1}^{m} w_k(f_k(x)+z_k^2)+\frac{1}{2}t\sum_{k=1}^{m}(f_k(x)+z_k^2)^2 + \sum_{j=1}^{\nu}\lambda_j\phi_j(x)+\frac{1}{2}t\sum_{j=1}^{\nu}\phi_j(x)^2; \tag{5.1}$$

cf. Bertsekas [2] and Rockafellar [12]. In (5.1), the z_k are slack variables and the vectors w and λ are multiplier candidates (but w is not constrained to be in W_m). The method of multipliers proceeds as follows: Given iterates w^i and λ^i and a penalty coefficient t_i, one minimizes (5.1) over all x and z to obtain x_i and z_i. The values of x_i and z_i are then used to define w^{i+1}, λ^{i+1}, and t_{i+1}. In practice, the minimization over z is carried out explicitly in advance and z disappears from the problem. For this detail and many others, one should see [2] and [12] and the many references cited therein.

Bertsekas has shown [2] that the classical method of multipliers can profitably be viewed as a steepest ascent method for maximizing a certain dual function. Bertsekas has [2] used this observation to show that if certain typical sufficiency conditions are satisfied and if $\{t_i\}$ stays bounded, then the method of multipliers will construct sequences $\{w^i\}$ and $\{\lambda^i\}$ which will converge Q-linearly to the optimal Lagrange multiplier. Rockafellar has shown [12] that the (approximate) minimizers x_i of the augmented Lagrangians must converge "at least as fast" as the sequences $\{\lambda^i\}$ and $\{w^i\}$; Rockafellar's result, which is established in [12] for the convex case, implies that the sequence $\{x_i\}$ will converge R-linearly to the optimal solution. This result is similar, at least in qualitative terms, to Theorem 3.2 in the present paper. (No attempt is made to compare the value of $\theta(t)$ with the analogous value for the multiplier method; cf. [2, p. 531].)

To an extent, then, a comparison of the multiplier method with the method developed here may depend on the amount of work required to generate x_{i+1} from x_i (in the respective methods). It becomes difficult to set fair standards for a direct comparison, because the method for unconstrained minimization is not specified in the multiplier method. Needless to say, since the multiplier method has been tested considerably (see references cited in [2] and [12]), it certainly does not bear the burden of proof in this matter.

Luenberger's combined method generates [6] a sequence $\{x_i\}$ as follows. Given x_i, a Newton-like step is made with respect to certain dual variables, which results in an intermediate point z_i; x_{i+1} is obtained from z_i by a steepest descent move applied to an exterior penalty function. Given certain approximations, the work required to obtain x_{i+1} from x_i in [6] seems much less than in the present paper. On this basis, the present method seems poor by comparison. Furthermore, Luenberger presents a sharp result on the rate of convergence of the sequence $\{x_i\}$, at least for an idealized version of the algorithm applied to problems having linear *equality* constraints only. Again, however, it seems difficult to set definitive standards for comparison. It would be desirable to have computational examples for Luenberger's combined method. This would be particularly true for the case in which there are inequality constraints, inasmuch as the convergence analysis in [6] deals mainly with equality constraints.

By contrast to the two methods just discussed, the algorithm presented here gives, at every iteration, a point x_i which does satisfy the inequality constraints. Moreover, the dual of the direction-finding problem does depend mainly on the number of inequality constraints, hence, the present method may be of interest particularly when there are many variables, few constraints, and when it is important that the inequality constraints be satisfied. And, some version of it may be suitable for certain infinite-dimensional problems like those considered in [9] and [3].

REFERENCES

[1] L. ARMIJO, *Minimization of functions having continuous partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[2] D. P. BERTSEKAS, *Combined primal-dual and penalty methods for constrained minimization*, this Journal, 13 (1975), pp. 521–544.

[3] R. W. CHANEY, *On the Pironneau–Polak method of centers*, J. Optimization Theory Appl., 20 (1976), pp. 269–295.

[4] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York and London, 1968.

[5] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York and London, 1966.

[6] D. G. LUENBERGER, *A combined penalty function and gradient projection method for nonlinear programming*, J. Optimization Theory Appl., 14 (1974), pp. 477–495.

[7] ———, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass., 1973.

[8] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York and London, 1970.

[9] 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.

[10] ———, *On the rate of convergence of certain methods of centers*, Math. Programming, 2 (1972), pp. 230–257.

[11] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York and London, 1971.

[12] R. T. ROCKAFELLAR, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Programming, 5 (1973), pp. 354–373.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

A FINITELY SOLVABLE CLASS OF APPROXIMATING PROBLEMS*

GERARD G. L. MEYER†

Abstract. Let P be the following nonlinear programming problem: given $m+1$ continuously differentiable convex maps $f^0(\cdot), f^1(\cdot), \ldots, f^m(\cdot)$ from E^n into E, minimize $f^0(z)$ subject to $f^j(z) \leqq 0$, $j=1, 2, \ldots, m$. A well known approach for solving P consists of embedding P into a family of approximate problems $P(\alpha)$. Given $\alpha > 0$, the problem $P(\alpha)$ is to find a point z such that $f^j(z) \leqq 0$, $j=1, 2, \ldots, m$, and such that for every h in E^n, there exists j in $J(z, \alpha)$, j depending on h, satisfying $\langle \nabla f^j(z), h \rangle \geqq 0$, with $J(z, \alpha) = \{j \in \{1, 2, \ldots, m\} | f^j(z) + 1/\alpha \geqq 0\} \cup \{0\}$. In general, $P(\alpha)$ cannot be solved in a finite number of iterations and therefore one is obliged to use antizigzagging schemes of varying complexity. The purpose of this paper is to describe a class C of problems P such that the approximating problems $P(\alpha)$ may be solved in a finite number of steps. It is shown that if P is in C, then its solution is unique and is stable with respect to variation in the cost function. There are indications that this phenomenon is not restricted to the particular case under study and that there is a definite connection between the stability of the solution of a problem and the existence of a finite procedure for solving it.

Introduction. The class of feasible directions methods is a powerful tool for solving constrained minimization problems, min-max problems and unconstrained minimization problems in the absence of continuity of the gradient [1], [2], [3], [4], [7], [8], [9], [10]. The different versions proposed either involve all the constraints [3], [7], [9] and do not require "antizigzagging precautions," or involve only the constraints in a neighborhood of the current point and do require "antizigzagging precautions" [1], [3], [7], [9], [10]. In this paper only the latter case will be considered.

It is known that a large amount of complexity in the existing methods is due to the arbitrariness of the antizigzagging procedures. Once the unnecessary complexity is removed, one obtains a family of parametrized algorithms called drivable methods of feasible directions [5], [6]. These algorithms are simpler than the classical ones but still retain an antizigzagging procedure of sorts. This is due to the fact that the family of problems $P(\alpha)$ in which the original problem P is embedded may not be solved in a finite number of iterations.

This paper addresses itself to the question of characterizing a family C of problems P such that the approximating problems $P(\alpha)$ may be solved in a finite number of steps. It will be shown that C is not empty and that if P is in C, then its solution is unique and stable with respect to a family of perturbations of P.

Approximating problems. Given $m+1$ continuously differentiable convex maps $f^0(\cdot), f^1(\cdot), \cdots, f^m(\cdot)$ from E^n into E, let T be the subset of E^n defined by

$$T = \{z | f^j(z) \leqq 0, j = 1, 2, \cdots, m\}.$$

Assume that T is nonempty, compact and that for every z in T, the set $\{\nabla f^j(z) | f^j(z) = 0, j = 1, 2, \cdots, m\}$ is linearly independent.

PROBLEM P. Find a point z in T such that for all z' in T,

$$f^0(z) \leqq f^0(z').$$

* Received by the editors September 24, 1975, and in revised form June 29, 1976.

† Department of Electrical Engineering, Johns Hopkins University, Baltimore, Maryland 21218.
This work was supported by the National Science Foundation under Grant GK-36221.

In order to solve Problem P, one usually embeds it in a family of parametrized problems $P(\alpha)$. The main requirement is that as α approaches $+\infty$, the solution set of $P(\alpha)$ converges in a suitable sense to the solution set of P. Such a possible embedding is now described.

PROBLEM $P(\alpha)$. Given a positive scalar α, find a point z in T such that for all h in E^n,

$$\max_{j\in J(z,\alpha)} \langle \nabla f^j(z), h\rangle \geqq 0,$$

with

$$J(z,\alpha)=\{j\in\{1,2,\cdots,m\}|f^j(z)+1/\alpha\geqq 0\}\cup\{0\}.$$

The properties of the family $P(\alpha)$, $\alpha>0$, are contained in the theorem below. The proofs involved are not difficult and have been deleted.

THEOREM 1. *Let $D(P)$ and $D(P(\alpha))$ be the solution sets of the problems P and $P(\alpha)$ respectively. Then*:

(i) *$D(P)$ is a nonempty and closed subset of T and for all $\alpha>0$, $D(P(\alpha))$ is a nonempty and closed subset of T;*

(ii) *$D(P)\subseteq D(P(\alpha_2))\subseteq D(P(\alpha_1))$ for all $\alpha_2>\alpha_1>0$;*

(iii) *Given any neighborhood $N(D(P))$ of $D(P)$, there exists $\bar{\alpha}>0$, depending on $N(D(P))$, such that for all $\alpha\geqq\bar{\alpha}$, $D(P(\alpha))\subseteq N(D(P))$.*

The results of Theorem 1 suggest that it is easier to find points in $D(P(\alpha))$ than in $D(P)$. One may therefore consider the following iterative procedure: given a sequence $\{\alpha_i\}$ converging to $+\infty$, generate a sequence $\{z_i\}$ by computing for each i a point z_i in $D(P(\alpha_i))$. Every cluster point z^* of $\{z_i\}$ is in $D(P)$ and therefore problem P has been transformed into an infinite sequence of approximating problems $P(\alpha)$. It is clear that the usefulness of such a scheme will depend on the availability of methods for solving $P(\alpha)$.

A class of iterative procedures. A family of algorithms parametrized by a positive scalar β is now presented. The algorithms use a compact neighborhood S of the origin in E^n and a positive scalar ρ. The set S is usually selected so that the computations in Step 1 may be conveniently performed, and ρ may be chosen arbitrarily large but must be finite.

ALGORITHM $A(\beta)$.

Step 0. Compute z_0 in T and set $i=0$.

Step 1. Compute h_i in S such that for all h in S,

$$\max_{j\in J(z_i,\beta)} \langle \nabla f^j(z_i), h_i\rangle \leqq \max_{j\in J(z_i,\beta)} \langle \nabla f^j(z_i)h\rangle$$

Step 2. Let $\lambda_i=\max\{\lambda\in[0,\rho]|z_i+\lambda h_i\in T\}$.

Step 3. Compute μ_i in $[0,\lambda_i]$ such that for all μ in $[0,\lambda_i]$,

$$f^0(z_i+\mu_i h_i)\leqq f^0(z_i+\mu h_i).$$

Step 4. Set $z_{i+1}=z_i+\mu_i h_i$.

Step 5. If $f^0(z_{i+1})=f^0(z_i)$ stop; else set $i=i+1$ and go to Step 1.

Note that Step 1 is a linear program when S is polyhedral.

It is worth noting that many methods of feasible directions proposed in the literature use sequences of algorithms $A(\beta)$ [7], [9], [10]. The successive values given to β are determined by schemes of varying complexity [5], [6]. In this paper, the innate properties of the family of algorithms $A(\beta)$ are investigated. This is in contrast to the usual approach in which the properties of iterative processes containing $A(\beta)$ as a subprocedure are examined.

Solvable approximating problems. A class of approximating problems has been defined and a family of algorithms has been given. It is clearly of interest to know which algorithms solve which problems. Before this may be done, one needs a precise and unambiguous definition of the solvability of a problem P by an iterative procedure A.

DEFINITION. P is said to be *A-finitely solvable* iff every sequence $\{z_i\}$ generated by A when applied to P satisfies:

(i) The sequence $\{z\}$ is finite;

(ii) The last element of $\{z_i\}$ is a solution of P.

P is said to be *A-asymptotically solvable* iff every sequence $\{z_i\}$ generated by A when applied to P satisfies:

(iii) The sequence $\{z_i\}$ has at least one cluster point;

(iv) Every cluster point of $\{z_i\}$ is a solution of P.

P is said to be A-solvable iff every sequence $\{z_i\}$ generated by A when applied to P satisfies either (i) and (ii) or (iii) and (iv).

The theorem below characterizes the $A(\beta)$-solvable approximating problems $P(\alpha)$. Its proof is not difficult and may be found in [6] or [7].

THEOREM 2. *Let $\beta \geqq \alpha > 0$; then $P(\alpha)$ is $A(\beta)$-solvable.*

$A(\beta)$-Finitely solvable approximating problems. An important behavioral characterization of an iterative process consists in the identification of the class of problems that it finitely solves. This helps in understanding some of the algorithm's possibilities and limitations. In particular, the knowledge of the class of problems which are finitely solved by the procedure indicates in which cases the algorithm may be used as a subprocess without antizigzagging schemes.

In this section, a class C of problems P such that the corresponding $P(\alpha)$ are $A(\beta)$-finitely solvable is described. The approach followed consists essentially of two parts. First a family of approximating problems $P(\alpha)$ which are $A(\beta)$-finitely solvable is exhibited. Then the class C is presented. It is shown that if P is in C then $P(\alpha)$ is $A(\beta)$-finitely solvable.

The following hypothesis is sufficient to ensure that $P(\alpha)$ is $A(\beta)$-finitely solvable when $\beta \geqq \alpha \geqq \bar{\alpha}$.

HYPOTHESIS 1. There exists $\bar{\alpha} > 0$ such that the origin 0 of E^n is not on the boundary of the convex hull of the set

$$\{\nabla f^j(y) | j \in K \cup \{0\}\}$$

for all y in $D(P(\bar{\alpha}))$ and for all subsets K of $I(y, \bar{\alpha}) = \{j \in \{1, 2, \ldots, m\} | f^j(y) + 1/\bar{\alpha} \geqq 0\}$.

THEOREM 3. *Suppose that Hypothesis 1 is satisfied and let $\beta \geqq \alpha \geqq \bar{\alpha} > 0$. Then $P(\alpha)$ is $A(\beta)$-finitely solvable.*

Proof. In order to prove the theorem, it is sufficient to show that if an infinite sequence $\{z_i\}$ is generated by $A(\beta)$ when applied to $P(\alpha)$, then there exists a finite k such that $z_{k+1} = z_k$.

(i) Let $\beta \geqq \alpha \geqq \bar{\alpha} > 0$ and let $\{z_i\}$ be an infinite sequence generated by $A(\beta)$ when applied to $P(\alpha)$. Then there exist an infinite subset L of the integers, a point z^* in T and a subset J^* of $\{0, 1, 2, \cdots, m\}$ satisfying:

(i(a)) $\{z_i\}_L$ converges to z^*;

(i(b)) $J(z_i, \beta) = J^*$ for all i in L.

The fact that $\beta \geqq \alpha > 0$ and the results of Theorem 2 imply that

(i(c)) z^* is in $P(\alpha)$.

(ii) The set J^* is a subset of $J(z^*, \beta)$ and clearly the index 0 belongs to both the sets J^* and $J(z^*, \beta)$. It follows from Hypothesis 1 that only two cases are possible:

(ii(a)) The origin of E^n belongs to the interior of the hull of the set $\{\nabla f^j(z^*) | j \in J^*\}$;

(ii(b)) The origin of E^n does not belong to the convex hull of the set $\{\nabla f^j(z^*) | j \in J^*\}$.

In order to show that (ii(a)) is true, one shows that (ii(b)) leads to a contradiction.

Suppose that (ii(b)) is true; then there exists h in E^n such that

(ii(c)) $\langle \nabla f^j(z^*), h \rangle < 0$ for all j in J^*.

By construction, if j is not in J^*, then for every i in L, the index j is not in $J(z_i, \beta)$. This implies that $f^j(z_i) + 1/\beta < 0$ for all i in L and using the continuity of the map $f^j(\cdot)$ one obtains,

$$\text{(ii(d))} \qquad f^j(z^*) < 0 \quad \text{for all } j \text{ not in } J^*.$$

It is known [6] that (ii(c)) and (ii(d)) contradict the fact that $f^0(z)$ is bounded from below on T. One concludes that (ii(b)) is not true and therefore (ii(a)) is true (i.e., the approximate Fritz-John necessary conditions are satisfied).

(iii) The auxiliary results needed to prove the theorem have been obtained and one proceeds towards the conclusion of the proof. The fact that (ii(a)) is true and that the maps $f^j(\cdot)$ are continuously differentiable implies that there exists a neighborhood $N(z^*)$ of z^* such that the origin of E^n belongs to the convex hull of

$$\{\nabla f^j(z) | j \in J^*\}$$

for all z in $N(z^*)$. By construction

$$J(z_i, \beta) = J^*,$$

and therefore the origin in E^n belongs to the convex hull of

$$\{\nabla f^j(z_i) | j \in J(z_i, \beta)\}$$

for all i in L and for all z in $N(z^*)$. But the sequence $\{z_i\}_L$ converges to z^* and therefore there exists k such that z_i is in $N(z^*)$ for all $i \geqq k$, i in L. It is immediate that $f^0(z_{k+1}) = f^0(z_k)$ and the theorem is proved.

The iterative procedure $A(\beta)$ finitely solves a nonempty family of problems $P(\alpha)$. In order to obtain C one must find hypotheses on P which insure that the corresponding $P(\alpha)$ are $A(\beta)$-finitely solvable. The assumption below is sufficient to guarantee that this is the case.

HYPOTHESIS 2. For every z in $D(P)$, the origin in E^n belongs to the interior of the convex hull of the set

$$\{\nabla f^j(z) | j \in I(z) \cup \{0\}\}$$

with

$$I(z) = \{j \in \{1, 2, \cdots, m\} | f^j(z) = 0\}.$$

THEOREM 4. *Suppose that P satisfies Hypothesis 2. Then there exists $\bar{\alpha} > 0$, depending on P, such that $P(\alpha)$ is $A(\beta)$-finitely solvable for all $\beta \geqq \alpha > \bar{\alpha}$.*

Proof. Assume that Hypothesis 2 is satisfied. The set $D(P)$ is a subset of T and therefore z in $D(P)$ implies that the set

$$\{\nabla f^j(z) | j \in I(z)\}$$

is linearly independent. It follows immediately that the origin in E^n is not on the boundary of the convex hull of the set

$$\{\nabla f^j(z) | j \in K \cup \{0\}\}$$

for all subsets K of $I(z)$.

The maps $f^0(\cdot), f^1(\cdot), \cdots, f^m(\cdot)$ are continuously differentiable and this fact together with part (iii) of Theorem 1 shows that Hypothesis 1 is satisfied for some $\alpha > 0$. The result of Theorem 4 is then a direct consequence of Theorem 3.

Let C be the class of all problems P satisfying Hypothesis 2. The lemma below is a direct consequence of Theorem 4.

LEMMA 1. *Let P be in C. Then there exists $\bar{\alpha} > 0$, depending on P, such that every sequence $\{z_i\}$ generated by $A(\beta)$ when applied to P with $\beta \geqq \bar{\alpha}$ satisfies:*

(i) *$\{z_i\}$ is finite;*

(ii) *The last element of $\{z_i\}$ is in $P(\beta)$.*

Stability of *P*. The family C exhibits some remarkable properties. One of them, namely the fact that $A(\beta)$ generates only finite sequences when applied to P in C when β is large enough has been presented in the preceding section. The properties of the solution set $D(P)$ of P in C are now investigated.

LEMMA 2. *Suppose that P is in C; then $D(P)$ consists of one and only one point.*

Proof. By assumption, T is nonempty and compact and $f^0(\cdot)$ is continuous. It follows that $D(P)$ is nonempty. The maps $f^0(\cdot), f^1(\cdot), \cdots, f^m(\cdot)$ are convex and therefore $D(P)$ is convex. Suppose that z_1 and z_2 are two distinct points in $D(P)$, then the entire segment $[z_1, z_2]$ is in $D(P)$. Let $h = z_2 - z_1$. Then,

(i) $h \neq 0$;

(ii) $\langle \nabla f^0(z_1), h \rangle = 0$;

(iii) $\langle \nabla f^j(z_1), h \rangle \leqq 0$ for all j such that $f^j(z_1) = 0$, $j \neq 0$.

This obviously contradicts Hypothesis 2 and therefore $D(P)$ contains only one point.

The solution set $D(P)$ of P does not vary when P is subject to small variations in the cost function. In order to define these variations precisely, it is convenient to introduce the class of problems $V(P, \bar{\varepsilon})$.

Class $V(P, \bar{\varepsilon})$. Given P and $\bar{\varepsilon} > 0$, the class $V(P, \bar{\varepsilon})$ consists of all the problems $\tilde{P}$ characterized by

$$D(\tilde{P}) = \{z \text{ in } T | \tilde{f}^0(z) \leqq \tilde{f}^0(z') \text{ for all } z' \text{ in } T\},$$

where $\tilde{f}^0(\cdot)$ is a continuously differentiable convex map from E^n into E satisfying

$$\|\nabla \tilde{f}^0(z) - \nabla f^0(z)\| \leqq \bar{\varepsilon}$$

for all z in T.

LEMMA 3. *Suppose that P is in C. Then there exists $\bar{\varepsilon} > 0$ depending on P such that, for all problems $\tilde{P}$ in $V(P, \bar{\varepsilon})$:*

(i) *The solution set $D(\tilde{P})$ of $\tilde{P}$ contains one and only one point;*

(ii) $D(\tilde{P}) = D(P)$.

Proof. P is in C and therefore satisfies Hypothesis 2. It follows that there exists $\bar{\varepsilon} > 0$ such that Hypothesis 2 is still satisfied when $f^0(\cdot)$ is replaced by $\tilde{f}^0(\cdot)$, provided that $\tilde{f}^0(\cdot)$ is a continuously differentiable convex map from E^n into E and also that

$$\|\nabla \tilde{f}^0(z) - \nabla f^0(z)\| \leqq \bar{\varepsilon}$$

for all z in T. This shows that if $\tilde{P}$ is in $V(P, \bar{\varepsilon})$, then there exists a problem, call it $\tilde{P}'$, satisfying Hypothesis 2 and such that $D(\tilde{P}) = D(\tilde{P}')$. But, since $\tilde{P}'$ satisfies Hypothesis 2, $\tilde{P}'$ is in C and so $D(\tilde{P}')$ consists of a single point. The second part of the proof is an immediate consequence of the fact that $D(P)$ is a subset of $D(\tilde{P})$ for all $\tilde{P}$ in $V(P, \bar{\varepsilon})$.

It has been shown that if P is in C, the solution of the problem is not modified by small perturbations of the cost function. It happens that this property in fact characterizes C.

LEMMA 4. *Let P be given. Suppose that there exists $\bar{\varepsilon} > 0$ depending on P such that for all $\tilde{P}$ in $V(P, \bar{\varepsilon})$, the solution set $D(\tilde{P})$ of $\tilde{P}$ contains one and only one point and $D(\tilde{P}) = D(P)$. Then P is in C.*

Proof. Let z be in $D(P)$ and suppose that there exists $\bar{\varepsilon} > 0$ such that z is in $D(\tilde{P})$ for all $\tilde{P}$ in $V(P, \bar{\varepsilon})$. Let H be the convex hull of the set

$$\{\nabla f^j(z) | f^j(z) = 0, j = 1, 2, \cdots, m\}.$$

The origin in E^n does not belong to H but belongs to the convex hull of the set

$$H \cup \{\nabla \tilde{f}^0(z)\}$$

for all continuously differentiable convex maps $\tilde{f}^0(\cdot)$ which satisfy

$$\|\nabla \tilde{f}^0(y) - \nabla f^0(y)\| \leqq \bar{\varepsilon}$$

for all y in T. It follows immediately that the origin in E^n belongs to the interior of the convex hull of the set

$$H \cup \{\nabla f^0(z)\}$$

and therefore P is in C.

Conclusion. It has been shown that a class C of problems P possesses two important properties. On one hand, there exists a class of iterative processes which terminate when applied to P in C. On the other hand, the solution of P in C is invariant under a class of perturbations of P. It is the author's conviction that these properties are related, i.e., that there is a connection between the type of procedures which halt when applied to a problem and the type of perturbations which leave the solution of the problem unperturbed. If this were actually the case, it would provide a powerful tool for the synthesis of efficient iterative procedures.

REFERENCES

[1] J. CULLUM, W. E. DONATH, AND P. WOLFE, *An algorithm for minimizing certain nondifferentiable convex functions*, IBM Res. Rep. RC 4611, IBM, Yorktown Heights, NY, 1973.

[2] A. M. GEOFFRION, *Primal resource-directive approaches for optimizing nonlinear decomposable systems*, Operations Research, 18 (1970), pp. 375–403.

[3] P. HUARD, *Tour d'horizon: Programmation non lineaire*, Rev. Française Rech. Oper., 5 R-1 (1971), pp. 3–48.

[4] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[5] G. G. L. MEYER, *An open loop method of feasible directions for the solution of optimal control problems*, Proc. Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton, NJ, (1972), pp. 679–680.

[6] ———, *A drivable method of feasible directions*, this Journal, 11 (1973), pp. 113–118.

[7] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[8] D. M. TOPKIS AND A. VEINOTT, *On the convergence of some feasible directions algorithms for nonlinear programming*, this Journal, 5 (1967), pp. 268–279.

[9] W. I. ZANGWILL, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[10] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

A NOTE ON THE LACK OF EXACT CONTROLLABILITY FOR MILD SOLUTIONS IN BANACH SPACES*

ROBERTO TRIGGIANI†

Abstract. It is shown that exact controllability in finite time for linear control systems given on an infinite dimensional Banach space in integral form (mild solution) can never arise using locally L_1-controls, if the associated C_0 semigroup is compact for all $t>0$. This includes, in particular, the class of parabolic partial differential equations defined on bounded spatial domains.

Consider the control process described by the following integral model:

$$x(t, x_0, u) = S(t)x_0 + \int_0^t S(t-\tau)Bu(\tau)\,d\tau, \qquad t \geqq 0, \tag{1.1}$$

under the following standard assumptions: $x(t, \cdot, \cdot)$ belongs to a separable Banach space X (state space); $u(t)$ is a U-valued function, locally L_1 (control function), where U (control space) is also a Banach space; $S(t)$, $t \geqq 0$, is a strongly continuous semigroup of bounded operators (of class C_0); B is a bounded operator: $U \to X$; finally $x_0 \in X$. The integral is well defined in the sense of Bochner. Unless otherwise stated, X will be assumed infinite dimensional. Also, (1.1) is (strongly) continuous in t [4, p. 88]. See [4, Chap. III] for the necessary background for vector valued functions.

It is customary to refer to (1.1), for a locally L_1 function $u(t)$, as 'mild solution' of the correspondent differential equation

$$\dot{x} = Ax + Bu, \qquad x(0) = x_0 \in X, \tag{1.1'}$$

where A is the infinitesimal generator of $S(t)$.

Let $\mathscr{A}_t$ be the set of attainability from the origin of the system (1.1), corresponding to L_1-control functions over $[0, t]$ i.e.,

$$\mathscr{A}_t = \{x \in X;\ x = x(t, 0, u),\ u(\cdot) \in L_1[0, t]\}.$$

We then say that (1.1) is exactly controllable on $[0, T]$, $T>0$, (respectively, in finite time) in case: $\mathscr{A}_T = X$ (respectively, $\bigcup_{0 \leqq t < \infty} \mathscr{A}_t = X$).

Notice that the strict solution $x(t, x_0, u)$ of the differential model (1.1')—for $x_0 \in D(A)$ (domain of A) and, say, a C^1-*control* $u(t)$—always lies in $D(A)$, which, when the closed operator A is unbounded, is only dense in X, by the closed graph theorem. In other words, exact controllability of the strict solution when the generator A is unbounded is out of the question.

As for the mild solution, recently the author has shown

THEOREM 1.1 (see [9]). *Let X be infinite dimensional. Then, the system* (1.1) *is never exactly controllable in finite time using locally L_1-controls (in symbols: $\bigcup_{0 \leqq t < \infty} \mathscr{A}_t \subsetneq X$) if the operator B: $U \to X$ is compact.*

Under the additional assumption that X has a Schauder basis, a simpler proof of the above result was previously given by the author in [10, § 3, see Remark 3.3.2].

* Received by the editors April 5, 1976, and in revised form July 26, 1976.
† Mathematics Department, Iowa State University, Ames, Iowa 50011.

For some cases where exact controllability is achieved, see [10, § 3] (e.g. B is onto and $S(t)$ is a group).

In the present note we shall explicitly assume, in view of the above theorem, that the operator B is bounded but *not compact.* We shall then transfer the assumption of compactness from B to the semigroup $S(t)$ for $t>0$ and conclude with a result establishing the lack of exact controllability in finite time of (1.1) analogous to Theorem 1.1 above.

Remark 1.1. The assumption that the operator $S(t)$ is compact on X for all $t>0$ (semigroup of compact operators) is met by large classes of dynamical systems of physical interest. In fact [6], [7] $S(t)$ is compact for all $t>0$ if and only if (i) $S(t)$ is continuous in the uniform operator topology for $t>0$ and (ii) the resolvent $R(\lambda, A)$ of its generator is compact at some (hence all [1, p. 210]) λ in the resolvent set of A. Assumption (ii) is always satisfied, say, by partial differential equations defined on bounded spatial domains and assumption (i) holds, of course, for the large class of semigroups which are differentiable for all $t>0$. Hence, parabolic partial differential equations defined on bounded spatial domains (whose correspondent semigroups are in fact holomorphic) represent an important subclass of dynamical systems, whose correspondent semigroups are compact for all $t>0$. The forthcoming application-oriented book [1] is in fact mainly concentrated on compact (even Hilbert-Schmidt) semigroups.

We shall prove

THEOREM 1.2. *Let X be infinite dimensional. Then the system* (1.1) *is never exactly controllable in finite time using locally L_1-controls (in symbols, $\bigcup_{0\leq t<\infty}\mathscr{A}_t \subsetneq X$) if the semigroup $S(t)$ is compact for all $t>0$.*

Proof 1. We shall first prove that the operator Q,

$$Q\tilde{u}=\int_0^T S(T-t)Bu(t)\,dt, \qquad \tilde{u}\in L_1[[0,T],U],$$

from $L_1[[0,T],U]\to X$ is compact. To this end, define, for each $\varepsilon>0$, the operator Q_ε,

$$Q_\varepsilon\tilde{u}=\int_0^{T-\varepsilon} S(T-t)Bu(t)\,dt, \qquad \tilde{u}\in L_1[[0,T],U] \tag{1.2}$$

from $L_1[[0,T],U]\to X$. That Q_ε is compact for all $\varepsilon>0$ follows from writing the right-hand side of (1.2) as

$$S(\varepsilon)\int_0^{T-\varepsilon} S(T-t-\varepsilon)Bu(t)\,dt \tag{1.3}$$

and the fact that as $u(\cdot)$ runs over all vectors in the unit sphere of $L_1[[0,T],U]$, the integral in (1.3) describes a bounded set in X.

Next we have from standard bounds on semigroups

$$\|Q\tilde{u}-Q_\varepsilon\tilde{u}\|=\left\|\int_{T-\varepsilon}^{T} S(T-t)Bu(t)\,dt\right\|\leq \varepsilon M e^{\alpha\varepsilon}\|B\|\|\tilde{u}\|_1\to 0 \quad \text{as } \varepsilon\to 0,$$

and hence Q, being the uniform limit of compact operators, is also compact.

2. Therefore the image $K_n(T)$ under Q of the sphere in $L_1[[0, T], U]$ of radius n, centered at the origin, is a precompact set in X. Let $\bar{K}_n(T)$ be its closure. Since X is infinite dimensional, $\bar{K}_n(T)$ cannot contain spheres [5, p. 269] and hence is nowhere dense in X. Next, observe that exact controllability on $[0, T]$ demands $X = \bigcup_{n=1}^{\infty} K_n(T)$. This is however impossible by the Baire category theorem [8, p. 139]. Consequently $\bigcup_{n=1}^{\infty} \bar{K}_n(T)$ does not fill all of X. Since $\mathscr{A}_T \subset \bigcup_{n=1}^{\infty} \bar{K}_n(T)$, the lack of exact controllability of (1.1) on $[0, T]$ is established.

3. Take the sequence of time intervals $[0, i]$, $i = 1, 2, \cdots$. Then, step 2 says that the subspace $\mathscr{A}_i = \bigcup_{n=1}^{\infty} K_n(i)$ is a set of first category. Consequently, $K = \bigcup_{i=1}^{\infty} \mathscr{A}_i$ is also a set of first category [8, p. 140] and hence K does not fill all of X. But $i-1 < T < i$, $i = 1, 2, \cdots$, implies

$$K_n(i-1) \subset K_n(T) \subset K_n(i), \qquad n = 1, 2, \cdots, \quad i = 1, 2, \cdots.$$

(In fact a point reachable from the origin over $[0, t_1]$ using the control $u(t)$, $0 \leqq t \leqq t_1$, is also reachable from the origin over a larger interval $[0, t_2]$ by applying first the null control over $[0, t_2 - t_1)$ and then the control $\bar{u}(t) = u(t-(t_2-t_1))$ over $[t_2 - t_1, t_2]$.) Hence, from the above inclusions, taking the union over all n, one gets

$$\mathscr{A}_{i-1} \subset \mathscr{A}_T \subset \mathscr{A}_i.$$

Then, taking first the union over all T in $[i-1, i]$ and then over all i, one arrives at

$$K = \bigcup_{i=1}^{\infty} \mathscr{A}_{i-1} \subset \bigcup_{0 \leqq T < \infty} \mathscr{A}_T \subset \bigcup_{i=1}^{\infty} \mathscr{A}_i = K.$$

Hence $\bigcup_{0 \leqq T < \infty} \mathscr{A}_T = K$ and $\bigcup_{0 \leqq T < \infty} \mathscr{A}_T$ does not fill all of X. Actually, even $\bigcup_{0 \leqq T < \infty} [\bigcup_{n=1}^{\infty} \bar{K}_n(T)]$ is not all of X. Q.E.D.

Remark 1.2. Let x_0 be given and let $\bar{x} \notin \mathscr{A}_T$. Then the point x_1 in X defined by $x_1 = S(T)x_0 + \bar{x}$ cannot be reached from x_0 by using $L_1[[0, T], U]$-controls.

Remark 1.3. In the case of functional differential equations of retarded type written as abstract ordinary differential equations, in the usual way [2] in the Hilbert space $X = R^n \times L_2[[-h, 0], R^n]$, the corresponding semigroup is compact only for all t greater than or equal to a *positive* constant h (delay). The assumption of Theorem 1.2 is therefore not satisfied. However, such case is covered by our previous Theorem 1.1, since the corresponding operator B occurring in the model has finite dimensional range in X[2].

Example 1 (parabolic partial differential equations in bounded domains). Let Ω be a bounded domain in R^n with smooth boundary, let $X = L_2(\Omega)$ and let $A(x, D)$ be the partial differential operator of even order $2m$,

$$A(\xi, D) = \sum_{|\alpha| \leqq 2m} a_\alpha(\xi) D^\alpha,$$

$$a_\alpha(\xi) = \text{sufficiently smooth complex-functions of } \xi \text{ in } \bar{\Omega},$$

(written in the usual notation) assumed to be strongly elliptic, i.e., satisfying

$$\operatorname{Re}(-1)^m A'(\xi, \zeta) \geqq C|\zeta|^{2m}$$

for some constant $C>0$, every $\xi \in \bar{\Omega}$ and $\zeta \in R^n$; here $A'(\xi, \cdot)$ is the principal part of $A(\xi, \cdot)$:

$$A'(\xi, D) = \sum_{|\alpha|=2m} a_\alpha(\xi) D^\alpha.$$

Next, define the operator A by

$$Ax = -A(\xi, D)x, \qquad D(A) = H^{2m}(\Omega) \cap H_0^m(\Omega).$$

Then, A is the infinitesimal generator of a holomorphic semigroup $S(t)$ on X [7, Thm. 5.2.9, p. 139] and [3, Chap. XIV 8.1 p. 1767]. Also, its resolvent $R(\cdot, A)$ is compact, since Ω is bounded [3, Chap. XIV. 6.23, pp. 1739–1740]. Therefore, by Remark 1.1, the semigroup $S(t)$ is compact for all $t>0$. See also [1, Chap. 4] and [5, Chap. 7]. As for an operator B of physical significance which is bounded but not compact, besides the obvious example of the identity on $U=X=L_2(\Omega)$, we cite the multiplication operator $(Bu(t))(\xi) = m(\xi)\mu(t, \xi)$, $\xi \in \Omega$, on $L_2(\Omega)$) where $m(\xi)$ is a fixed, nontrivial, bounded, measurable function on Ω [5, p. 382], and $\mu(t, \xi)$ is the distributed control; etc.

Example 2. Let $\{x_j\}$, $j=1, 2, \cdots$ be an orthonormal system in the Hilbert space X and let $\{\lambda_j\}$ be a sequence of isolated complex numbers (with no finite accumulation point) which satisfy (i) $\operatorname{Re}\lambda_j \leqq \omega < \infty$ and (ii) $\operatorname{Re}\lambda_j \to -\infty$ as $j \to \infty$. Define the family of bounded operators $S(t)$ by

$$S(t)x = \sum_{j=1}^{\infty} e^{\lambda_j t}(x, x_j)x_j, \qquad x \in X, \quad t \geqq 0.$$

Then $S(t)$ is a C_0-semigroup which is compact for $t>0$, since $|e^{\lambda_j t}| \to 0$ as $j \to \infty$ for $t>0$ [5, p. 383; 1, p. 90]. However, $S(t)$ need not be holomorphic for $t>0$, which occurs if the $\{\lambda_j\}$ are chosen so that they are not contained in any triangular sector of the type

$$\{\lambda : \operatorname{Re}\lambda < a - b|\operatorname{Im}\lambda|\}, \quad a, b > 0$$

[7, § 2.2.5]. The infinitesimal generator is a normal operator and is given by

$$Ax = \sum_{j=1}^{\infty} \lambda_j(x, x_j)x_j, \qquad D(A) = \{x \in X : \sum_{j=1}^{\infty} |\lambda_j(x, x_j)|^2 < \infty\}.$$

The spectrum of A is only point spectrum and it consists precisely of all the numbers λ_j. For all $\lambda \neq \lambda_j$ the resolvent $R(\lambda, A)$ of A is normal and given by

$$R(\lambda, A)x = \sum_{j=1}^{\infty} \frac{1}{\lambda - \lambda_j}(x, x_j)x_j, \qquad x \in X,$$

and is compact since $1/|\lambda - \lambda_j| \to 0$ as $j \to \infty$. For a related example see [1, Example 4.6.5].

For the approximate controllability problem, $\overline{\cup_{0 \leqq t < \infty} \mathscr{A}_t} = X$ in the spirit of the classical finite dimensional theory see [11].

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.
[2] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of affine hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
[3] N. DUNFORD AND J. T. SCHWARTZ, *Linear operators*, part 2, Interscience, New York, 1963.
[4] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, American Mathematical Society, Providence, RI, 1957.
[5] T. H. NAYLOR AND G. R. SELL, *Linear Operators in Engineering and Science*, Holt, Rinehart and Winston, New York, 1971.
[6] A. PAZY, *On the differentiability and compactness of semi-groups of linear operators*, J. Math. Mech., 17 (1968), pp. 1131–1141.
[7] ———, *Semigroups of linear operators and applications to partial differential equations*, Lecture Note 10, Dept. of Mathematics, Univ. of Maryland, College Park, 1974.
[8] H. L. ROYDEN, *Real Analysis*, 2nd ed, Macmillan, New York, 1968.
[9] R. TRIGGIANI, *On the lack of exact controllability for mild solutions in Banach spaces*, J. Math. Anal. Appl., 50 (1975), pp. 438–446.
[10] ———, *Controllability and observability in Banach space with bounded operators*, this Journal, 13 (1975), pp. 462–491.
[11] ———, *Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators*, this Journal, 14 (1976), pp. 236–250.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

OBSERVATION AND PREDICTION FOR THE HEAT EQUATION. IV: PATCH OBSERVABILITY AND CONTROLLABILITY*

THOMAS I. SEIDMAN†

Abstract. It is shown first that the problem of predicting the state at $t = T$ of a solution of the heat equation with homogeneous boundary conditions from observation on a patch of complementary boundary data is well-posed in the presence of an *a priori* bound on initial data. Next, it is shown that Russell's star-complementarity sufficient condition for exact boundary controllability is far from necessary by demonstrating controllability from the inner face of an annular sector in $\mathbb{R}^n$. Finally, some results are obtained as to the dependence of the optimal null-control on the patch considered and on T.

1. Introduction. We consider problems of observation and control for the heat equation

$$\dot{u} = \Delta u \tag{1}$$

in a bounded domain Ω in $\mathbb{R}^n$ in which interaction with the process—for either observation or control—is limited to a boundary patch Γ_1 (i.e., a relatively open subset of the boundary $\Gamma = \partial\Omega$). Some of the results presented here were announced in [18] (note, in particular, Theorems 4.2, 4.4 and 5.2 there).

The author would like to express his thanks to David Russell both for the stimulation of his work in this area and for several valuable discussions. The direct stimulus to this work on problems of patch controllability and observability was some work in progress by William Chewning at the time of his death; as with the corresponding section of [18], this paper is dedicated to his memory.

In [13], [17] it was shown that *null-controllability* (i.e., existence, for each initial u_o in $L_2(\Omega)$, of boundary data φ carrying u_0 to zero terminal state at $t = T > 0$) is dual to *well-posedness of observation/prediction* (briefly: *observability* = continuity of the map from observed complementary data to terminal state) and that these equivalent properties hold for quite general regions Ω if Γ_1 is all of Γ. Although not so stated, the extension principle used in the proof of Theorem 3 of [17] gives this result for nonautonomous boundary conditions (cf. [18]). The results of [19] give this for a cylinder $\Omega = \mathscr{D} \times (0, 1)$ in $\mathbb{R}^n$ with Γ_1 a *base* of Ω ($\Gamma_1 = \mathscr{D} \times \{0\}$) and the results of [14] give this for *star-complemented* settings (see Theorem 3 below).

The argument of [14] proceeds from corresponding results for the wave equation ($\ddot{u} = \Delta u$) in Ω for which it is known (see [11]) that the control patch "cannot be too small"—e.g., existence in Ω of a closed polygonal path with reflection at vertices in $\Gamma_2 = \Gamma \backslash \Gamma_1$ implies (see [12]) nonexistence of a decay rate (in the sense of scattering theory) and so noncontrollability: "trapped waves" in Ω cannot be exactly canceled by any admissible boundary control. On the other hand, it was shown in [6], [7] (see, also: [3, Chap. III, § 10.2]) that *approximate* patch controllability (given a patch Γ_1, $\varepsilon > 0$ and u_0 there exists boundary data φ, supported in the patch, carrying u_0 to a terminal state with norm less than ε) holds

* Received by the editors February 26, 1976.

† Division of Mathematics and Physics, University of Maryland, Baltimore County, Baltimore, Maryland 21228. This research was supported in part by U.S. Army Grant DAA-G-29-77-G-0061.

under quite general conditions—it is equivalent to uniqueness for the Cauchy problem: that $u_\nu = 0$ on $(0, T)\times\Gamma$ and $u = 0$ on $(0, T)\times\Gamma_1$ implies $u \equiv 0$. It has been conjectured that the properties of observability and controllability hold for an arbitrarily small patch in the case of the heat equation but this problem remains open. In this paper we obtain a weaker form of patch observability – adequate for certain applications—and also show, by some new examples, that the necessary conditions for wave equation controllability are far from necessary for the heat equation.

2. Constrained observability. We consider the observation/prediction problem in the presence of an a priori constraint on u—specifically, we assume given an a priori bound on the norm of the initial state u_0. Note that in an important class of applications for diffusion processes one has an interpretation of u as a *concentration*, in which case one knows automatically that $0 \leqq u \leqq 1$ pointwise so that the result applies.

We let Ω be a bounded region in $\mathbb{R}^n$ with "smooth" boundary $\Gamma = \partial\Omega$ and divide Γ into an interaction patch Γ_1 and a residue (passive boundary) $\Gamma_2 = \Gamma\backslash\Gamma_1$; let $T > 0$ be given and set $\mathcal{Q} = (0, T)\times\Omega$, $\Sigma = (0, T)\times\Gamma$, $\Sigma_j = (0, T)\times\Gamma_j (j = 1, 2)$. Denote by u an arbitrary solution of (1) in $\mathcal{Q}$ for which $u(0, \cdot) = u_0$ is in $L_2(\Omega)$. We suppose a pair of (smooth) functions (α, β) is given on Σ with $\alpha^2 + \beta^2 = 1$ and consider the boundary data

$$\alpha u + \beta u_\nu = \varphi \quad \text{on } \Sigma. \tag{2}$$

We also consider the *complementary data*

$$\beta u - \alpha u_\nu = \psi \quad \text{on } \Sigma. \tag{3}$$

Let φ_j, ψ_j denote the restrictions of φ, ψ to Σ_j $(j = 1, 2)$ and assume φ_j in $L_2(\Sigma_j)$. By a *setting* $\mathfrak{S} = [\mathcal{Q}, \Sigma_1, (\alpha, \beta)]$ we mean the specification of the region in which (1) holds, the portion of $\partial\Omega$ used for interaction and the form of the data (more generally, $\mathfrak{S} = [\mathbf{L}, \mathcal{Q}, \Sigma_1, (\alpha, \beta)]$ if (1) is replaced by a more general diffusion equation: $\dot{u} = \mathbf{L}u$); this is slightly different from the notation of [18] although we only consider time-independent geometries, as described above, for which $\mathcal{Q}, \Sigma_1$ are determined by Ω, Γ_1. We assume that the initial-boundary value problem given by (1), (2) and specification of $u(0, \cdot) = u_0$ is well-posed and let $\mathcal{U}_M$ be the set of solutions of this with $\|u_0\| \leqq M$; correspondingly, let $\mathcal{V}_M$ be the set of pairs $[\varphi, \psi_1]$ in $L_2(\Sigma)\times L_2(\Sigma_1)$ defined by (2), (3) for u in $\mathcal{U}_M$. The *constrained observation/prediction problem* which we consider concerns the mapping $\mathbf{P}_M$: $\mathcal{V}_M \to L_2(\Omega)$: $[\varphi, \psi_1] \mapsto u(T, \cdot)$.

THEOREM 1. *Let $\mathfrak{S}$ be a setting for which specification of $[\varphi, \psi_1]$ uniquely determines $u(T, \cdot)$. We assume, for convenience, that $\beta > 0$ on Σ so the boundary operator given by* (2) *has constant order* 1. *Then, for any $M > 0$, the estimation mapping $\mathbf{P}_M$: $\mathcal{V}_M \to L_2(\Omega)$ is continuous.*

Proof. We have $u(T, \cdot) = \mathbf{S}u_0 + \mathbf{B}\varphi$ with $\mathbf{S}$ compact from $L_2(\Omega)$ and $\mathbf{B}$ continuous from $L_2(\Sigma)$ into $L_2(\Omega)$; $\mathbf{S}$ is the solution operator for the pure initial value problem (its compactness follows from that of the resolvent of $\mathbf{\Delta}$ for each of the sets of boundary conditions (2) for $0 \leqq t \leqq T$) while the continuity of $\mathbf{B}$ follows from [5, p. 78(iii)], taking, e.g., $s = -1/3$ which actually gives continuity from

$H^{-1/6,-1/12}(\Sigma)$ to $H^{1/6}(\Omega)$. Similarly, $\psi_1 = \mathbf{S}_1 u_0 + \mathbf{B}_1 \varphi$ with $\mathbf{S}_1$ continuous from $L_2(\Omega)$ and $\mathbf{B}_1$ continuous from $L_2(\Sigma)$ into, e.g., $H^{-7/6}(\Sigma_1)$ on taking $s = -2/3$ in [5, p. 78 (iii)]. Now select a sequence $(y_1, y_2, \cdots)$ in $H^{7/6}(\Sigma_1)$ which is total for $H^{-7/6}(\Sigma_1)$, i.e.,

$$\langle \psi_1, y_j \rangle = \int_{\Sigma_1} y_j \psi_1 = 0 \quad (j = 1, 2, \cdots) \quad \text{implies} \quad \psi_1 = 0. \tag{4}$$

Note that each scalar product appearing in (4) is continuously dependent on ψ_1 in $H^{-7/6}(\Sigma_1)$ and so on u_0 in $L_2(\Omega)$ and on φ in $L_2(\Sigma)$. Suppose, now, $([\varphi_k, \psi_k])$ is any sequence in $\mathcal{V}_M$ with $\varphi_k \to \varphi$, $\psi_k \to \psi_1$. Let (u_k) be the corresponding sequence of solutions of (1) with corresponding sequences of initial data (v_k) and of terminal data (w_k), i.e., $v_k = u_k(0, \cdot)$ in $\mathcal{U}_M$ and $w_k = u_k(T, \cdot) = \mathbf{P}_M[\varphi_k, \psi_k]$ in $L_2(\Omega)$ so

$$w_k = \mathbf{S} v_k + \mathbf{B} \varphi_k, \qquad \psi_k = \mathbf{S}_1 v_k + \mathbf{B}_1 \varphi_k. \tag{5}$$

By the weak sequential compactness of $\mathcal{U}_M$, any subsequence of (v_k) contains a weakly convergent sub-subsequence $v_{k(i)} \rightharpoonup v_*$ to which (1), (2) associates a solution u_*. Now, for each j, $\langle \psi_{k(i)}, y_j \rangle \to \langle \psi_1, y_j \rangle$ but

$$\begin{aligned} \langle \psi_{k(i)}, y_j \rangle &= \langle \mathbf{S}_1 v_{k(i)}, y_j \rangle + \langle \mathbf{B}_1 \varphi_{k(i)}, y_j \rangle \\ &= \langle v_{k(i)}, \mathbf{S}_1^* y_j \rangle + \langle \mathbf{B}_1 \varphi_{k(i)}, y_j \rangle \\ &\to \langle v_*, \mathbf{S}_1^* y_j \rangle + \langle \mathbf{B}_1 \varphi, y_j \rangle \end{aligned}$$

whence, by (4), $\psi_1 = \mathbf{S}_1 v_* + \mathbf{B}_1 \varphi$ (observe that this shows that $\mathcal{V}_M$ is closed in $L_2(\Sigma) \times L_2(\Sigma_1)$). At the same time, the compactness of $\mathbf{S}$ means that $v_{k(i)} \rightharpoonup v_*$ implies $\mathbf{S} v_{k(i)} \to \mathbf{S} v_*$ so $w_{k(i)} \to [\mathbf{S} v_* + \mathbf{B} \varphi] = w_*$. Thus, w_* is the terminal data for u_* associated with $[\varphi, \psi_1]$ in $\mathcal{V}_M$ and, by assumption, this is uniquely determined by $[\varphi, \psi_1]$ so $w_* = \mathbf{P}_M[\varphi, \psi_1]$. Since *every* subsequence of $([\varphi_k, \psi_k])$ contains a sub-subsequence for which $w_{k(i)} \to w_*$, it follows that $w_k \to w_*$—i.e., $\mathbf{P}_M[\varphi_k, \psi_k] \to \mathbf{P}_M[\varphi, \psi_1]$ proving continuity of $\mathbf{P}_M$. □

Remark 1. It is clear from this argument that the observation of ψ_1 may be topologized in any $H^s(\Sigma_1)$ (e.g., any negative s) since the sequence $(y_1, \cdots)$ can be selected in $C^\infty(\Sigma_1)$. On the other hand, the condition $\beta > 0$ in the theorem is related to the use of the $L_2(\Sigma)$ topology for φ: the results of [5] give $u(T, \cdot) \in L_2(\Omega)$ for (1), (2) with $\beta \neq 0$, $\varphi \in L_2(\Sigma)$ and $u_0 \in L_2(\Omega)$. If $\beta \equiv 0$ (Dirichlet problem), then the order of the boundary operator given by (2) is zero and an argument almost identical to the above gives the same conclusion provided $\mathcal{V}_M$ is now topologized so that observation of φ is in $H^{2s,s}(\Sigma)$ with $s > 1/4$ (e.g., in $H^1(\Sigma)$); one might expect similar results if $\beta = 0$ on *part* of Σ, although in such a case the results of [5] no longer apply (assuming, of course, that the relevant initial-boundary value problem is well-posed).

Remark 2. If the pair $[\varphi_1, \psi_1]$ is sufficient to determine $u(T, \cdot)$ uniquely, then in the presence of a priori bounds on both u_0 in $L_2(\Omega)$ and ψ_2 in $L_2(\Sigma_2)$, the analogous mapping: $[\varphi_1, \psi_1] \mapsto u(T, \cdot)$ is continuous since the maps $\mathbf{B}$ and $\mathbf{B}_1$ are continuous from $H^{-1/6,-1/2}(\Sigma)$ to $H^{1/6}(\Omega)$ and $H^{-7/6}(\Sigma_1)$, respectively, while the embedding of $L_2(\Sigma_1) = H^0(\Sigma_1)$ into $H^{-1/6,-1/12}(\Sigma_1)$ is compact (cf., e.g., [5, p. 99])

so the analogous argument (with ψ_1 and $u(T, \cdot)$ now dependent on the unknown *pair* $[\varphi_0, u_0]$ rather than just on u_0) works. In applications one is quite likely to be considering *homogeneous* boundary conditions ((2) with $\varphi = 0$) and in that case (cf., Theorem 4.4 of [18]) one similarly might obtain continuity of the mapping: $\psi_1 \mapsto u(T, \cdot)$ without regard for the requirement that the order of the boundary operator be constant (i.e., one may then permit $\beta = 0$ on part of Σ). We state these results below, without further proof, as Theorem 2.

Remark 3. If, e.g., Ω has an analytic boundary Γ, then [10] the uniqueness hypothesis of the theorem is satisfied for any relative open subset Γ_1. One would expect, however, that this would hold under much weaker conditions than global analyticity of Γ.

THEOREM 2. (i) *Let* $\mathfrak{S}$ *be a setting with* β *nonvanishing for which* $[\varphi, \psi_1] = [0, 0]$ *implies* $u(T, \cdot) = 0$. *and let* $\mathbf{P}_M: [\varphi, \psi_1] \mapsto u(T, \cdot)$ *with* $[\varphi, \psi_1] \in L_2(\Sigma) \times H^{-s}(\Sigma_1)$ *associated with* $u_0 \in L_2(\Omega)$ *such that* $\|u_0\| \leqq M$, *or*

(ii) *let* $\mathfrak{S}$ *be a setting with* $\beta \equiv 0$ *for which* $[\varphi, \psi_1] = [0, 0]$ *implies* $u(T, \cdot) = 0$ *and let* $\mathbf{P}_M: [\varphi, \psi_1] \to u(T, \cdot)$ *with* $[\varphi, \psi_1] \in H^1(\Sigma) \times H^{-s}(\Sigma_1)$ *associated with* $u_0 \in L_2(\Omega)$ *such that* $\|u_0\| \leqq M$, *or*

(iii) *let* $\mathfrak{S}$ *be a setting with* β *nonvanishing for which* $[\varphi_1, \psi_1] = [0, 0]$ *implies* $u(T, \cdot) = 0$ *and let* $\mathbf{P}_M: [\varphi_1, \psi_1] \to u(T, \cdot)$ *with* $[\varphi_1, \psi_1] \in L_2(\Sigma_1) \times H^{-s}(\Sigma_1)$ *associated with* $u_0 \in L_2(\Omega)$, $\varphi_2 \in L_2(\Sigma_2)$ *such that* $\|u_0\|, \|\varphi_2\| \leqq M$, *or*

(iv) *let* $\mathfrak{S}$ *be any setting for which* $[\varphi, \psi_1] = [0, 0]$ *implies* $u(T, \cdot) = 0$ *and let* $\mathbf{P}_M: \psi_1 \mapsto u(T, \cdot)$ *with* $\psi_1 \in H^{-s}(\Sigma_1)$ *associated with* $\varphi = 0$ *and* $u_0 \in L_2(\Omega)$ *such that* $\|u_0\| \leqq M$.

Then, in each case and for any $M > 0$, $\mathbf{P}_M$ *is well-defined and continuous to* $L_2(\Omega)$.

The standard 'state identification' problem of finite dimensional control theory is the determination of u_0 from observations over $0 < t < T$. It is clear that the argument above gives this (with observation of ψ_1 on Σ_1) in the presence of an a priori bound on $u(t_0, \cdot)$ for some $t_0 < 0$ and assuming $\varphi \equiv 0$ for $t_0 < t < T$. In this case it is clear that such an a priori bound is actually necessary but for the cases (i), (ii), (iv) considered in Theorem 2 one might plausibly conjecture that (with $s = 0$: $\psi_1 \in L_2(\Sigma_1)$) the imposed bound is unnecessary—indeed, for a variety of situations (see Theorems 3, 4, 6, below) this is known.

3. Some new geometries. The duality between boundary observation/prediction and control problems (stated below as Theorem 3; cf., [17], [18]) shows that the continuity of an *unconstrained* estimation map ($\psi_1 \mapsto u(T, \cdot)$ for $\varphi \equiv 0$) is equivalent to a null-controllability result but the results of Theorem 2, above, seem to have no implication for patch controllability (other than the *approximate* controllability already implied by the uniqueness hypothesis). For convenience, we state the Duality Theorem (cf., [17], [18]) in a form suitable for our present uses.

THEOREM 3. *Let* $\mathfrak{S}$ *be an autonomous setting. Then null-controllability and well-posedness of the estimation (observation/prediction) problem are equivalent, i.e., the following are equivalent*:

(a) *For every* u_0 *in* $L_2(\Omega)$ *there exists* φ *in* $L_2(\Sigma_1)$ *for which the solution* u *of* (1), (2) *starting at* u_0 *has* $u(T, \cdot) = 0$.

(b) *There is a constant K such that $\|u(T,\cdot)\| \leqq K\|\psi\|$ for all solutions of* (1), (2) *with $\varphi = 0$, and $u(0,\cdot)$ in $L_2(\Omega)$, where ψ is defined on Σ_1 by* (3) *and normed in $L_2(\Sigma_1)$.*

The optimum (minimum $L_2(\Sigma_1)$-norm) control φ, when this exists, is in the closure $\bar{\mathcal{M}}$ of the set $\mathcal{M}$ of functions ψ in $L_2(\Sigma_1)$ defined by (3) *for solutions of $-\dot{u} = \Delta u$ with $u(T,\cdot)$ in $L_2(\Omega)$ and satisfying homogeneous boundary conditions* (2).

The most general available results on patch controllability for (1) were obtained by D. Russell [13], [14], paraphrased below.

THEOREM 4. *Let $\mathfrak{S}$ be an* autonomous *setting for which one has exact null-controllability from Σ_1 for the* wave *equation (i.e., for some $T_w > 0$ and every $(u_0, \dot{u}_0)$ in $H^2(\Omega)\times H^1(\Omega)$ there exists φ_1 in $H^{1/2}(\Sigma_1)$ for which the solution of $\ddot{u} = \Delta u$ with* (2) *has $u = \dot{u} = 0$ at $t = T_w$). Then the heat equation* (1) *is null-controllable by φ_1 in $L_2(\Sigma_1)$ (i.e., $\varphi_0 \equiv 0$) for every $T > 0$, every u_0 in $L_2(\Omega)$. In particular, this holds if $\alpha \equiv 1$ on Σ_2 and $\mathfrak{S}$ is star-complemented.*

Here $\mathfrak{S}$ is called *star-complemented* if there is a star-shaped body Ω_* exterior to Ω for which $\Gamma_2 \subset \partial\Omega_*$—essentially, this means the existence of a point ω_* exterior to Ω from which all of Γ_2 is "visible" (i.e., the segment $\omega_2 - \omega_*$ lies outside Ω for each ω_2 in Γ_2). Using the extension principle ([18, Thm. 3.5], stated below as Theorem 5), the autonomy condition on $\mathfrak{S}$ can be relaxed—although still with $\alpha \equiv 1$ on Σ_2.

THEOREM 5. *Let the setting $\mathfrak{S} = [\mathcal{Q}, \Sigma_1, (\alpha, \beta)]$ have an extension $\mathfrak{S}^+ = [\mathcal{Q}^+, \Sigma_1^+, (\alpha, \beta)^+]$ (i.e., $\mathcal{Q} \subset \mathcal{Q}^+$, $T^+ = T$, $\Sigma_2 \subset \Sigma_2^+$, $(\alpha, \beta)^+$ matching (α, β) on Σ_2) which is null-controllable. Then $\mathfrak{S}$ is itself null-controllable.*

We note that the argument for this in [17], [18] requires *uniqueness* for the solution of (1), (2) with $u_0 = 0$ (existence, where relevant, is automatic) and *we henceforth assume this implicitly.*

It is known ([9], using duality, or [1]) that a rectangle (e.g., with $\alpha \equiv 1$) is null-controllable from one face for (1) while this is false [11] for the wave equation. Thus, the implication of Theorem 4 is not reversible. On the other hand, the rectangle is a "perturbation" of a star-complemented setting: "tilt very slightly" the parallel faces adjacent to the control face. See Fig. 1. In general, the star-complementarity condition on the geometry seems reasonable for the wave

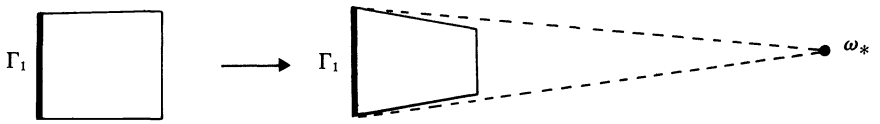


FIG. 1

equation but rather unnatural for the heat equation, suggesting that *any* non-trivial face should suffice for control or observation. While this remains conjectural, the examples adduced below certainly indicate the substantial gap between what suffices for the heat equation and what is needed for the wave equation.

Basically, the argument shows that, in certain separable geometries, null-controllability from one 'face' implies null-controllability from the 'opposite face', after which Theorems 4 and 5 can be applied; specifically, we consider an n-dimensional annulus or annular sector.

Before proceeding we state and prove a result concerning ordinary differential equations. As this seems of interest in its own right, it is presented in considerably greater generality than is needed later.

THEOREM 6. *Let* $\sigma>0$, $a>1$ *and set* $c^2=(\sigma+1)[\max\{1,\sigma/a\}]^{1-1/a}$ $(c=c(\sigma,a)>0)$. *Let* S *be any monotone function on* $[t_0,t_1]$ *with* $|S|\leqq\sigma$ *and let* y *be any nontrivial solution of*

$$\ddot{y}+(S+1)y=0 \tag{6}$$

on $[t_0,t_1]$ *satisfying the boundary conditions*

$$\alpha_0 y(t_0)+\beta_0\dot{y}(t_0)=0=\alpha_1 y(t_1)+\beta_1\dot{y}(t_1) \tag{7}$$

$(\alpha_0^2+\beta_0^2=1=\alpha_1^2+\beta_1^2;\ \beta_0\leqq 0)$. *Then*

$$|\eta_0/\eta_1|,\ |\eta_1/\eta_0|\leqq c\,\exp[(t_1-t_0)\sqrt{a\sigma}], \tag{8}$$

where η_0, η_1 *are the complementary boundary data*:

$$\eta_0=\beta_0 y(t_0)-\alpha_0\dot{y}(t_0),\qquad \eta_1=\beta_1 y(t_1)-\alpha_1\dot{y}(t_1).$$

Proof. With no loss of generality we may take $t_0=0$, $t_1=T$, $y(0)\geqq 0$ and S nondecreasing on $[0,T]$. Taking $0\leqq\theta_0<\pi$ such that $\alpha_0=\cos\theta_0$, $\beta_0=-\sin\theta_0$ we define $r=r(t)>0$ and $\theta=\theta(t)$ on $[0,T]$ by the Prüfer substitution: $\dot{y}=r\cos\theta$, $y=r\sin\theta$ with $\theta(0)=\theta_0$; let $\theta_1=\theta(T)$, noting that (7) implies $\alpha_1=\cos\theta_1$, $\beta_1=\sin\theta_1$. Clearly, $r(0)=|\eta_0|$ and $r(T)=|\eta_1|$. From (6) we have

$$\dot{r}/r=-S\sin\theta\cos\theta,\qquad \dot{\theta}=1+S\sin^2\theta \tag{9}$$

so that

$$\log|\eta_1/\eta_0|=\int_0^T(\dot{r}/r)\,dt=-\int_0^T S\sin\theta\cos\theta\,dt=\mathscr{J}. \tag{10}$$

We distinguish the subintervals

$$\mathscr{A}=[T_*,T]=\{t\in[0,T]:S(t)\geqq 0\},$$
$$\mathscr{B}=[0,T_*]=\{t\in[0,T]:S(t)\leqq 0\},$$

and separately estimate the integrals $\mathscr{J}_{\mathscr{A}}$, $\mathscr{J}_{\mathscr{B}}$ of $(\dot{r}/r)$ over each of these.

On $\mathscr{A}$, $\dot{\theta}>0$ and we rewrite $\mathscr{J}_{\mathscr{A}}$ as an integral $d\theta$. Let $\theta_*=\theta(T_*)$ satisfy $J\pi/2\leqq\theta_*<(J+1)\pi/2$ and let $K\pi/2<\theta_1\leqq(K+1)\pi/2$ (this defines J, K).Then

$$-\mathscr{J}_{\mathscr{A}}=\int_{\theta_*}^{\theta_1}\frac{S\sin\theta\cos\theta}{1+S\sin^2\theta}\,d\theta=\sum_{k=J}^{K}(-1)^k\mathscr{J}_{(k)}, \tag{11}$$

where

$$\mathscr{J}_{(k)}=\int_{k\pi/2}^{(k+1)\pi/2}\frac{|\sin\theta\cos\theta|}{\sin^2\theta+u_k}\,d\theta$$

with

$$u_k=u_k(\theta)=\begin{cases}1/S, & \theta_*<\theta\leqq\theta_1;\quad k\pi/2\leqq\theta\leqq(k+1)\pi/2,\\ \infty, & \text{otherwise.}\end{cases}$$

Since S is nondecreasing, the alternating series in (11) is bounded by the larger of $\mathcal{I}_{(K-1)}$, $\mathcal{I}_{(K)}$ or (as $u_k \geqq 1/\sigma$)more conveniently by

$$\mathcal{I}^*_{(K)} = \int_{K\pi/2}^{(K+1)\pi/2} \frac{|\sin\theta\cos\theta|}{\sin^2\theta + 1/\sigma}\,d\theta$$

$$= \frac{1}{2}\int_0^1 \frac{dz}{z+1/\sigma}, \qquad z = \sin^2\theta,$$

so

$$|\mathcal{I}_{\mathcal{A}}| \leqq \tfrac{1}{2}\log(\sigma+1). \tag{12}$$

The interval $\mathcal{B}$ is now further subdivided into the sets

$$\mathcal{C} = \{t \in B: -S\sin^2\theta \leqq a\},$$

$$\mathcal{D} = \{t \in B: -S\sin^2\theta \geqq a\}.$$

On $\mathcal{C}$ one has

$$|\dot{r}/r| = |S\sin\theta\cos\theta| \leqq |S|^{1/2}|S\sin^2\theta|^{1/2} \leqq \sqrt{a\sigma}$$

so

$$|\mathcal{I}_{\mathcal{C}}| \leqq \sqrt{a\sigma}\,\text{meas}\,\mathcal{C} \leqq T\sqrt{a\sigma}. \tag{13}$$

Finally, we observe that $\dot{\theta} < 0$ on $\mathcal{D}$ and that the monotonicity of S ensures that $\mathcal{E} = \{t \in [0, T]: \dot{\theta} < 0\}$ is a single (possibly empty) interval $[0, T_{**}]$: otherwise there would be points τ_1, τ_2 with $\theta(\tau_1) = \theta(\tau_2)$ while $\theta(\tau_1) > 0 > \dot{\theta}(\tau_2)$ and $\tau_1 < \tau_2$ which would contradict $S(\tau_1) \leqq S(\tau_2)$ (alternatively, if S were differentiable with $\dot{S} > 0$, then at a critical point of $\theta(\dot{\theta} = 0)$ one would have $\ddot{\theta} = \dot{S}\sin^2\theta > 0$; the more general case is obtainable from this by a limit argument). Clearly, then, one cannot have $\sin\theta = 0$ in $\mathcal{D}$ so we many subdivide $\mathcal{D}$ into the two (possibly empty) sub-intervals

$$\mathcal{F} = \{t \in \mathcal{D}: 0 < \theta \leqq \pi/2\}, \qquad \mathcal{G} = \{t \in \mathcal{D}: \pi/2 \leqq \theta < \pi\}.$$

Since $-S\sin\theta\cos\theta$ has opposite signs on $\mathcal{F}$, $\mathcal{G}$ we have $|\mathcal{I}_{\mathcal{D}}| \leqq \max\{|\mathcal{I}_{\mathcal{F}}|, |\mathcal{I}_{\mathcal{G}}|\}$ and, setting $z = \sin^2\theta$, we have on either of these

$$1 \geqq z > a/|S| \geqq a/\sigma = z_*, \qquad 1/|S| \leqq \frac{1}{a}\sin^2\theta$$

(if $\sigma < a$, then $\mathcal{D}$ is empty and we may set $z_* = 1$). Over $\mathcal{F}$ or $\mathcal{G}$

$$\left|\int(\dot{r}/r)\right| = \left|\int \frac{S\sin\theta\cos\theta}{S\sin^2\theta + 1}\,d\theta\right|$$

$$= \frac{1}{2}\int \frac{2|\sin\theta\cos\theta|\,d\theta}{\sin^2\theta - 1/S}$$

$$\leqq \frac{1}{2}\int \frac{2|\sin\theta\cos\theta|\,d\theta}{\sin^2\theta - (1/a)\sin^2\theta}$$

$$\leqq \frac{1/2}{1-1/a}\int_{z_*}^1 \left(\frac{dz}{z}\right)$$

so

$$|\mathscr{I}_{\mathscr{D}}| \leqq \frac{1/2}{1-1/a} \log 1/z_* = \frac{1/2}{1-1/a} \log (\max\{1, \sigma/a\}). \tag{14}$$

Combining (10), (12), (13), (14), noting that $[t_0\ t_1] = \mathscr{A} \cup \mathscr{C} \cup \mathscr{D}$, and taking the exponential gives the desired estimate (8). □

Remark 4. The theorem gives $|\eta_1/\eta_0| = \exp[\mathcal{O}(\sigma^{1/2})]$ and this is easily seen to be sharp ($\sigma^{1/2}$ cannot be replaced by σ^γ with $\gamma < 1/2$); an example can be constructed with, e.g., $\alpha_0 = \alpha_1 = 1$ and $S \equiv -\sigma$ on $[0, T_\sigma]$, $S \equiv \sigma$ on $[T_\sigma, T]$ where T_σ ($\approx T \gg 1$) is determined to satisfy the boundary conditions: $\theta(0) = \pi/2 = \theta(T)$.

Remark 5. Observe that any boundary data at $t = t_0, t_1$ other than η_0, η_1 can be obtained directly from these:

$$\beta y(t_j) - \alpha \dot{y}(t_j) = [\beta_j \beta + \alpha_j \alpha]\eta_j, \qquad j = 0, 1, \tag{15}$$

so for $\hat{\eta}_j = \hat{\beta}_j y(t_j) - \hat{\alpha}_j \dot{y}(t_j)$ $(j = 0, 1)$,

$$|\hat{\eta}_1/\hat{\eta}_0| = \left|\frac{\hat{\beta}_1\beta_1 + \hat{\alpha}_1\alpha_1}{\hat{\beta}_0\beta_0 + \hat{\alpha}_0\alpha_0}\right| |\eta_1/\eta_0|, \tag{16}$$

and Theorem 5 can be applied.

COROLLARY. *Let S be any function on $[t_0, t_1]$ with $|S| \leqq \sigma$ and having at most m "changes of direction" (i.e., there exist $\{\tau_j\}$ with $t_0 < \tau_1 < \cdots < \tau_m < t_1$, partitioning $[t_0, t_1]$ so that S is monotone on each subinterval); let y satisfy* (6), (7). *Then with σ, a, $c = c(\sigma, a)$, η_0, η_1 as in Theorem* 5,

$$|\eta_0/\eta_1|, |\eta_1/\eta_0| \leqq c^{m+1} \exp[(t_1 - t_0)\sqrt{a\sigma}]. \tag{17}$$

Proof. Let $\tau_0 = t_0$, $\tau_{m+1} = t_1$ and set $\hat{\alpha}_j = \gamma_j \dot{y}(\tau_j)$, $\hat{\beta}_j = \gamma_j y(\tau_j)$ with $\gamma_j = [y^2(\tau_j) + \dot{y}^2(\tau_j)]^{1/2}$ so $\hat{\alpha}_j y(\tau_j) + \hat{\beta}_j \dot{y}(\tau_j) = 0$; set $\hat{\eta}_j = \hat{\beta}_j y(\tau_j) - \hat{\alpha}_j \dot{y}(\tau_j)$. Theorem 6 then applies to each of the $(m+1)$ intervals $[\tau_j, \tau_{j+1}]$, giving

$$|\hat{\eta}_j/\hat{\eta}_{j+1}|, |\hat{\eta}_{j+1}/\hat{\eta}_j| \leqq C \exp[(\tau_{j+1} - \tau)\sqrt{a\sigma}]$$

for $j = 0, \cdots, m$ whence (17) follows by multiplication on noting that $|\eta_0| = |\hat{\eta}_0|$, $|\eta_1| = |\hat{\eta}_{m+1}|$. □

LEMMA 1. *Let w, p, q be positive on $[x_0, x_1]$ and suppose $\lambda \geqq 1$ is such that there is a nontrivial solution of*

$$\begin{gathered} w(py')' - qy = -\lambda y, \\ \alpha_0 y(x_0) + \beta_0 y'(x_0) = 0 = \alpha_1 y(x_1) + \beta_1 y'(x_1). \end{gathered} \tag{18}$$

Let $S = (\lambda - q)p/w - 1$ and suppose $\alpha_0\beta_0 \leqq 0 \leqq \alpha_1\beta_1$. Then

$$|S| \leqq C\lambda, \tag{19}$$

where $C = [1 + C_1(q_1/q_0 - 1)]$ with $C_1 = \max\{p/w\}$, $q_1 = \max\{q\}$ and $q_0 = \min\{q\} > 0$.

Proof. The differential equation (18) gives: $(\lambda - q)y^2/w = -(py')'y$. Integrating over $[x_0, x_1]$ (integrating by parts on the right), gives

$$(\lambda - q_0)\int y^2/w \geqq \int (\lambda - q)y^2/w = -py'y|_{x_0}^{x_1} + \int p(y')^2. \tag{20}$$

Since $\alpha_1\beta_1 \geqq 0$, the condition at x_1 in (18) implies $y'(x_1)y(x_1) \leqq 0$ and, similarly, $\alpha_0\beta_0 \leqq 0$ implies $y'(x_0)y(x_0) \geqq 0$; thus the right hand side of (20) is nonnegative so $\lambda \geqq q_0$. Clearly $S \leqq \lambda p/w \leqq C_1\lambda$ while

$$-S \leqq 1 + (q_1 - \lambda)p/w \leqq \lambda + (q_1/q_0 - 1)\lambda p/w$$

(since $1 \leqq \lambda$, $q_1 \leqq (q_1/q_0)\lambda$) which proves (19). □

Returning, now to our control-theoretic considerations, we let Ω_0 be any simply connected domain (assume $\partial\Omega_0$ smooth, if not empty) on the unit $(n-1)$-sphere in $\mathbb{R}^n$ (unit circle for $n = 2$) and let Ω be the region in $\mathbb{R}^n$ given in "polar coordinates" by

$$\Omega = \{\mathbf{x} = r\boldsymbol{\omega}\colon r_0 < r < r_1,\ \boldsymbol{\omega} \in \Omega_0\} \tag{21}$$

with $0 < r_0 < r_1$. The boundary $\Gamma = \partial\Omega$ consists of the *inner face* $\Gamma_1 = \{r_0\boldsymbol{\omega}\colon \boldsymbol{\omega} \in \Omega_0\}$, the *outer face* $\Gamma_0 = \{r_1\boldsymbol{\omega}\colon \boldsymbol{\omega} \in \Omega_0\}$, and (if Ω is not the entire annulus) the *lateral boundary* $\Gamma_3 = \{r\boldsymbol{\omega}\colon r_0 \leqq r \leqq r_1,\ \boldsymbol{\omega} \in \partial\Omega_0\}$; set $\Gamma_2 = \Gamma_0 \cup \Gamma_3$.

Geometrically, Ω is the intersection of the annulus $\{\mathbf{x}\colon r_0 < |\mathbf{x}| < r_1\}$ with the "cone" $\{r\boldsymbol{\omega}\colon \boldsymbol{\omega} \in \Omega_0\}$. It is not hard to see that a setting with control from the outer face Γ_2 is star-complemented so that Theorem 3 will apply (if $\alpha \equiv 1$ on $\Gamma_1 \cup \Gamma_3$) to give controllability. See Fig. 2. Our aim will be to show controllability from the inner face Γ_1 and also to weaken, somewhat, the condition on (α, β) required by Theorem 4.

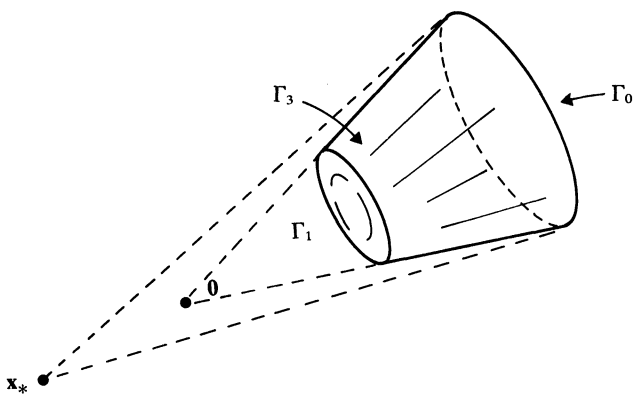


FIG. 2

It is easily seen that such a region Ω contains trapped waves (in the context of the wave equation) if the "open" (interactive) portion of the boundary is limited to the inner face. For example, in the case of an annulus in $\mathbb{R}^2$ one can always inscribe a regular polygon (having "sufficiently many" sides) in the outer circle $r = r_1$ so as not to touch the inner circle $r = r_0$. Thus, controllability from Γ_1 cannot be

obtained for the heat equation—using Russell's argument [13] directly—from such controllability for the wave equation since [11] the wave equation setting is not controllable. Clearly, this geometry can in no sense be considered a perturbation of star-complemented situations and it would appear that any "nearby" setting must similarly be noncontrollable for the wave equation.

THEOREM 7. *Let Ω be given as in* (21), $\mathcal{Q} = (0, T) \times \Omega$, $\Sigma_j = (0, T) \times \Gamma_j$ *and let* $\mathfrak{S} = [\mathcal{Q}, \Sigma_1, (\alpha, \beta)]$ *with* $\alpha \equiv 1$ *on* Σ_3 *and* (α, β) *constant on* Σ_0 *with* $\alpha\beta \geqq 0$. *Then* $\mathfrak{S}$ *is controllable for the heat equation. Similarly,* $\mathfrak{S}^\dagger = [\mathcal{Q}, \Sigma_0, (\alpha, \beta)^\dagger]$ *is controllable if* $\alpha^\dagger \equiv 1$ *on* Σ_3 *and* $(\alpha, \beta)^\dagger$ *is constant on* Σ_1 *with* $\alpha^\dagger\beta^\dagger \geqq 0$.

Proof. Let $0 < r_0^* < r_0 < r_1 < r_1^*$ and $0 < T_* = T - \varepsilon < T$; let $\Omega^0 = \{r\boldsymbol{\omega}: r_0^* < r < r_1^*, \boldsymbol{\omega} \in \Omega_0\}$ and $\mathfrak{S}^0 = [(0, T_*) \times \Omega^0, \Sigma_0^0, (1, 0)]$. The setting $\mathfrak{S}^0$ is star-complemented and Theorem 4 applies to give controllability. Now $\mathfrak{S}^0$ is an extension of $\mathfrak{S}^1 = [(0, T_*) \times \Omega^1, \Sigma_0^1, (\alpha, \beta)^1]$ where $\Omega^1 = \{r\boldsymbol{\omega}: r_0^* < r < r_1, \boldsymbol{\omega} \in \Omega_0\}$ and $(\alpha, \beta)^1$ has $\alpha^1 \equiv 1$ on $\Sigma_1^1 \cup \Sigma_3^1$ and $(\alpha, \beta)^1 = (\alpha, \beta)$ on Σ_0^1 so, by Theorem 5, $\mathfrak{S}^1$ is controllable. By Theorem 3, this is equivalent to the existence of a constant K such that

$$\|u(T_*, \cdot)\| \leqq K\|\psi_0^1\| \tag{22}$$

for all solutions u of (1) in $\mathcal{Q}^1 = (0, T_*) \times \Omega^1$ satisfying *homogeneous* boundary conditions (2) on Σ^1 (using $(\alpha, \beta)^1$) for u_0 in $L_2(\Omega)$ and with the complementary boundary data ψ_0^1 given on Σ_0^1 by (3).

The Laplace operator separates in this geometry

$$\begin{aligned} \boldsymbol{\Delta} &= \frac{\partial^2}{\partial r^2} + \frac{n-1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\mathbf{L} \\ \mathbf{L} &= \sum_{j,k=1}^{n} [\delta_{j,k} - \omega_j\omega_k]\frac{\partial^2}{\partial\omega_j\partial\omega_k} - (n-1)\sum_j \omega_j\frac{\partial}{\partial\omega_j} \end{aligned} \tag{23}$$

(note that $\mathbf{L}$ is "purely angular") so we have the expansion

$$u(t, r\boldsymbol{\omega}) = \sum_{j,k=1}^{\infty} a_{j,k} f_k(\boldsymbol{\omega}) g_{j,k}(r) \exp[-\lambda_{j,k} t] \tag{24}$$

where the $\{f_k\}$ are eigenfunctions of $\mathbf{L}$ (i.e., $\mathbf{L}f_k = -\nu_k f_k$ with $f_k = 0$ on $\partial\Omega_0$; note that $0 \leqq \nu_k \to \infty$) and the $\{g_{j,k}, \lambda_{j,k}\}$ are determined by the eigenvalue problem for Bessel's equation:

$$\begin{aligned} &r^{1-n}(r^{n-1}g')' - (\nu_k/r^2)g = -\lambda g, \\ &g(r_0) = 0, \qquad \alpha_1 g(r_1) - \beta_1 g'(r_1) = 0, \end{aligned} \tag{25}$$

where (α_1, β_1) are the constant values of $(\alpha, \beta)^1$ (i.e., of (α, β)) on Σ_0^1. From (24), assuming suitable normalization of $\{f_k\}$ and $\{g_{j,k}\}$,

$$\begin{aligned} \|u(T_*, \cdot)\|^2 &= \sum_{j,k} |a_{j,k}|^2 \exp[-2\lambda_{j,k}(T - \varepsilon)], \\ \|\psi_2^1\|^2 &= \sum_k \int_0^{T_*} |\sum_j \eta_{j,k} a_{j,k} \exp[-\lambda_{j,k} t]|^2 \, dt \end{aligned} \tag{26}$$

with $\eta_{j,k}$ the observable complementary boundary data (cf., (25)) for $g_{j,k}$ at $r = r_1$

(after "integrating out" f_k):

$$\eta_{j,k} = \beta_1 g_{j,k}(r_1) + \alpha_1 g'_{j,k}(r_1).$$

Combining (26) with (22) gives

$$\sum_{j,k} |a_{j,k} \exp[\varepsilon\lambda_{j,k}]|^2 \exp[-2\lambda_{j,k}T] \leqq K^2 \sum_k \int_0^{T*} \left| \sum_j \eta_{j,k} a_{jk} \exp[-\lambda_{j,k}t] \right|^2 dt \tag{27}$$

for any set of coefficients $\{a_{j,k}\}$ for which the right hand side is finite.

If one sets

$$w = r^{1-n}, \quad p = r^{n-1}, \quad q = \nu_k r^{-2}$$

on $[r_0^*, r_1]$, then the differential equation (25) has the form (8), while after the substitution

$$t = t(r) = [(r_0^*)^{-n} - r^{-n}]/(n-1), \tag{28}$$

it takes the form (6) on $[0, T](T = t(r_1))$ with

$$S = S(r) = (\lambda - q)p/w - 1 = \lambda r^{2n-2} - \nu_k r^{2n-4}.$$

Note that $S' = 2[(n-1)\lambda r^2 - (n-2)\nu_k]r^{2n-5}$ can change sign at most once in $[r_0^*, r_1]$—i.e., at

$$r_* = \left[\frac{n-2}{n-1} \cdot \frac{\nu_k}{\lambda}\right]^{1/2} < r_1$$

if $r_* > r_0^*$ (note that $r_* = 0 < r_0^*$ if $n = 2$)—so the Corollary to Theorem 6 applies with $m = 1$. We also apply the lemma to bound $|S|$ by $C\lambda$, for $\lambda = \lambda_{j,k} \geqq 1$, noting that C is independent of j, k (specifically, of ν_k) since

$$q_1/q_0 = [\nu_k (r_0^*)^{-2}]/[\nu_k r_1^{-2}] = (r_1/r_0^*)^2.$$

The complementary boundary data we will wish to observe for $g_{j,k}$ at r_0^* will be (note that the outward normal on Γ_1 gives $\partial/\partial\nu = -\partial/\partial r$)

$$\eta^*_{j,k} = -g'_{j,k}(r_0^*)$$

and, allowing for the substitution (30) and noting (16), we now have

$$\left|\frac{\eta_{j,k}}{\eta^*_{j,k}}\right| \leqq \left|\frac{\alpha_1^2 r_1^{n-1} + \beta_1^2}{(r_0^*)^{n-1}}\right| c^2(C\lambda, a)\, e^{T\sqrt{aC\lambda}}$$

for $\lambda \geqq 1$, from which it follows that, for any $\varepsilon > 0$,

$$|\eta_{j,k}/\eta^*_{j,k}| \leqq C(\varepsilon) \exp[\varepsilon\lambda_{j,k}] \quad \text{for all } j, k \tag{29}$$

with $C(\varepsilon)$ independent of j, k (for large enough $\lambda_{j,k}$ this is immediate—e.g., with $C(\varepsilon) = 1$—and there are only finitely many eigenvalues with $\lambda_{j,k} < M$ for any given M).

Let $\mathfrak{S}^2 = [\mathcal{Q}^2, \Sigma_1^2, (\alpha, \beta)^2]$ with $\mathcal{Q}^2 = (0, T) \times \Omega^1$, $\Sigma_j^2 = (0, T) \times \Gamma_j^2$, and $(\alpha, \beta)^2$ such that $\alpha^2 \equiv 1$ on $\Sigma_1^2 \cup \Sigma_3^2$ and matching (α, β) on $\Sigma_0^2 = \Sigma_0$ (note that this makes $(\alpha, \beta)^2$ constant on Σ_0^2, extending $(\alpha, \beta)^1$). To prove observability for $\mathfrak{S}^2$, we need an estimate $\|u(T, \cdot)\| \leqq K_2 \|\psi_1^2\|$, corresponding to (22), where, of course, ψ_1^2 is the complementary boundary data $\partial u / \partial \nu$ to be observed on Σ_1^2. We again use the expansion (24), now on $\mathcal{Q}^2$ and writing the coefficients as $\{b_{j,k}\}$ rather than $\{a_{j,k}\}$, permitting us to set

$$a_{j,k} = (\eta_{j,k}^* / \eta_{j,k}) b_{j,k}, \qquad |b_{j,k}| \leqq C(\varepsilon) \exp[\varepsilon \lambda_{j,k}] |a_{j,k}|,$$

for substitution into (27). Then

$$\begin{aligned} \|u(T, \cdot)\|^2 &= \sum_{j,k} |b_{j,k}|^2 \exp[-2\lambda_{j,k} T] \\ &\leqq \sum_{j,k} |C(\varepsilon) \exp[\varepsilon \lambda_{j,k}] a_{j,k}|^2 \exp[-2\lambda_{j,k} T] \\ &\leqq C^2(\varepsilon) K^2 \sum_k \int_0^{T_*} \left| \sum_j \eta_{j,k} a_{j,k} \exp[-\lambda_{j,k} t] \right|^2 dt \qquad (30) \\ &\leqq C^2(\varepsilon) K^2 \sum_k \int_0^{T} \left| \sum_j \eta_{j,k}^* b_{j,k} \exp[-\lambda_{j,k} t] \right|^2 dt \\ &= C^2(\varepsilon) K^2 \|\psi_1^2\|^2. \end{aligned}$$

This gives observability for $\mathfrak{S}^2$ and so, reversing the duality argument used above for $\mathfrak{S}^1$, $\mathfrak{S}^2$ is controllable. We observe that the setting $\mathfrak{S}^2$ is an extension of $\mathfrak{S}$ and so, by Theorem 5, $\mathfrak{S}$ is controllable.

To obtain controllability for $\mathfrak{S}^\dagger$ one would first apply the result proved so far to $\mathfrak{S}^{\dagger\dagger} = [\mathcal{Q}^{\dagger\dagger}, \Sigma_1, (\alpha, \beta)^{\dagger\dagger}]$ (with $\mathcal{Q}^{\dagger\dagger} = (0, T_*) \times \Omega^{\dagger\dagger}$, $\Omega^{\dagger\dagger} = \{r\boldsymbol{\omega}: r_0 < r_1^*, \boldsymbol{\omega} \in \Omega_0\}$, $\alpha^{\dagger\dagger} = 1$ on $\Sigma_0^{\dagger\dagger} \cup \Sigma_3^{\dagger\dagger}$ and $(\alpha, \beta)^{\dagger\dagger}$ constant, matching $(\alpha, \beta)^\dagger$, on $\Sigma_1^{\dagger\dagger} \subset \Sigma_1^\dagger$) and then use the same argument as was used in going from $\mathfrak{S}^1$ to $\mathfrak{S}^2$, above, to obtain controllability on the outer face ($r = r_1^*$, $0 \leqq t \leqq T$) with the condition given by $(\alpha, \beta)^\dagger$ on the inner face, after which application of Theorem 5 gives controllability for $\mathfrak{S}^\dagger$. □

Remark 6. The requirement that $\alpha\beta \geqq 0$ on the passive face was needed, in the proof, only to obtain the positivity of the right hand side of (20). On the other hand it seems entirely natural—"physically" it has the interpretation that heat is radiated *away* from a warm body—and is quite possibly a genuine requirement rather than a mere artifact of the proof. Indeed, it seems plausible that boundary conditions restricted in this fashion might appear in scattering theory to give the controllability for the wave equation with such a boundary condition (more generally, one might have a variable, albeit autonomous, dissipative boundary condition on the passive portion of Γ) and so provide the controllability for $\mathfrak{S}^\dagger$ via Theorem 3.

Remark 7. Strictly speaking, controllability for $\mathfrak{S}^0$ in the proof above does not follow from Russell's result [14] since Γ has "corners" (at $r = r_0^*$, r_1^* for $\boldsymbol{\omega}$ in $\partial\Omega_0$—unless Ω is the entire annulus so $\partial\Omega_0$ is empty) and so does not satisfy the C^∞ smoothness condition imposed in [14]. That condition, however, appears in [13],

[14] to permit exploitation of results in [4], [5] and the condition can be weakened considerably: in particular, it seems clear (especially in view of the separability of the geometry) that the results used remain valid for domains such as Ω, provided $\partial\Omega_0$ is smooth enough.

Remark 8. While the statement of the theorem concludes with controllability on a spherical face, application of the extension principle shows that one may control from a more irregular (even time-dependent) boundary surface. If we consider, as Ω, a ball with an internal cavity and wish to control (observe) at the interior surface, the "wall" of the cavity, with the outer sphere passive, we see from the above that this is possible—provided the center of the ball lies in the interior of the cavity so Theorem 4 can be applied after using Theorem 6 to obtain controllability from a small internal spherical surface. It remains open, then, as to whether such control would be possible if the cavity were placed elsewhere in the ball.

Remark 9. Similar considerations apply to various other geometries in which the Laplacian is separable (cf., e.g., [8]) and those could presumably be handled by essentially the same method (only the argument, following (28), to show a uniform bound on the number of possible sign changes of S' seems special) but we shall not pursue this.

4. Continuous dependence. We consider, now, the effect on the optimal control of varying the boundary patch employed. Letting the region Ω, the boundary conditions (α, β), and the initial state u_0 be fixed, we consider a sequence of nested boundary patches $\Gamma_j (\Gamma_1 \supset \Gamma_2 \supset \cdots)$ shrinking to a boundary patch $\Gamma_\infty = \cap \Gamma_j$. Assume that for each $\Sigma_j = [0, T] \times \Gamma_j$ there exists a null-control, hence an optimal null-control φ_j, in $L_2(\Sigma_j)$; we show that $\varphi_j \to \varphi_\infty$.

THEOREM 8. *Let $\mathfrak{S}_j = [\mathcal{Q}, \Sigma_j, (\alpha, \beta)]$ for $j = 1, 2, \cdots, \infty$ as above and assume there exist optimal null-controls φ_j associated with each of these and a fixed u_0. Then $\varphi_j \to \varphi_\infty$ in $L_2(\Sigma)$.*

Proof. Each φ_j is in $L_2(\Sigma_j) \subset L_2(\Sigma)$ so, for $j^* > j$, we have $\Sigma_{j^*} \subset \Sigma_j$ so φ_{j^*} is an admissible null-control for the setting $\mathfrak{S}_j$. Hence, by minimality, $\|\varphi_{j^*}\| \geqq \|\varphi_j\|$ (indeed, by the uniqueness of the optimal control, $\|\varphi_{j^*}\| > \|\varphi_j\|$ unless $\varphi_j = \varphi_{j^*}$, which would mean $\varphi_j = 0$ on $\Sigma_j \backslash \Sigma_{j^*}$). Similarly, $\|\varphi_j\| \leqq \|\varphi_\infty\|$ for $j = 1, 2, \cdots$ so the sequence $\{\varphi_j\}$ is bounded; extract any weakly convergent subsequence $\varphi_{j(k)} \rightharpoonup \varphi_*$. Let $\mathbf{A}$ be the bounded linear map from boundary data φ in (2) to terminal state $u(T, \cdot)$ for solutions of (1) with zero initial state so φ is a null-control for the state u_0 if and only if $\mathbf{A}\varphi = -u_T$ where u_T is the terminal state reached from u_0 with homogeneous boundary conditions. For any v in $L_2(\Omega)$,

$$\langle -u_T, v \rangle = \langle \mathbf{A}\varphi_{j(k)}, v \rangle$$

$$= \langle \varphi_{j(k)}, \mathbf{A}^* v \rangle \to \langle \varphi_*, \mathbf{A}^* v \rangle = \langle \mathbf{A}\varphi_*, v \rangle$$

so $\mathbf{A}\varphi_* = -u_T$ and φ_* is a null-control. If ψ is any element of $L_2(\Sigma)$ with support in the interior of $\Sigma \backslash \Sigma_\infty$, we have $\langle \varphi_{j(k)}, \psi \rangle = 0$ for k large enough that $\Sigma_{j(k)}$ is disjoint from the support of ψ; thus, $\langle \varphi_*, \psi \rangle = 0$ for each such ψ so φ_* has support in $\Sigma_\infty (\varphi_*$ in $L_2(\Sigma_\infty))$. This shows that φ_* is an admissible null-control for $\mathfrak{S}_\infty$ and, as $\|\varphi_*\| \leqq \liminf \|\varphi_{j(k)}\| \leqq \|\varphi_\infty\|$, we have, by the definition of φ_∞, that $\varphi_* = \varphi_\infty$ and $\|\varphi_{j(k)}\| \to \|\varphi_\infty\|$ so (see, e.g., [19, p. 124]) $\varphi_{j(k)} \to \varphi_\infty$ in $L_2(\Sigma)$. Since this holds for some sub-subsequence of every subsequence of $\{\varphi_j\}$, we have $\varphi_j \to \varphi_\infty$. □

A quite similar argument can be used to investigate the dependence of the optimal null-control on the length of the time interval $[0, T]$. Here, to make comparisons possible, we embed each of the spaces $L_2([0, T]\times\Gamma_1)$ for varying $T>0$ in $L_2([0, \infty)\times\Gamma_1)$. For a fixed boundary patch Γ_1 and fixed initial state u_0, it is convenient to introduce $N(T)=N(T; u_0)=\|\varphi(T)\|$ where $\varphi(T)$ is the optimal null-control in $L_2([0, T]\times\Gamma_1)\subset L_2([0, \infty)\times\Gamma_1)$ carrying u_0 to zero at time T.

THEOREM 9. *Let* $\mathfrak{S}_T=[\mathcal{Q}_T, \Sigma_{1,T}, (\alpha, \beta)_T]$, *where* $(\alpha, \beta)_T$ *is the restriction to* $\Sigma_T=[0, T]\times\Gamma$ *of* (α, β) *on* $\Sigma=(0, \infty)\times\Gamma$ *and where* $\mathcal{Q}_T=[0, T]\times\Omega$, $\Sigma_{1,T}=[0, T]\times\Gamma_1\subset\Sigma_1=[0, \infty)\times\Gamma_1$. *As above, let* $\varphi(T)$ *in* $L_2(\Sigma_{1,T})\subset L_2(\Sigma_1)$ *be the optimal null-control, assuming this exists for each* $T>0$, *associated with the fixed initial state* u_0 *and* $N(T)=\|\varphi(T)\|$. *Then* $N(\cdot)$ *is monotone decreasing (hence continuous except at possibly a countable number of values of* T*) and right continuous, with* $\varphi(0+)=\infty$ *if* $u_0\neq 0$. *The null-control* $\varphi(T)$ *depends continuously on* T *at continuity points of* N *and is right continuous everywhere.*

Proof. As embedded in $L_2(\Sigma_1)$, each $\varphi(T)$ is defined for x in Γ_1 and all t, vanishing for $t>T$. Clearly, the solution (on $[0, \infty)\times\Omega$) determined by $\varphi(T)$ is identically zero for $t\geqq T$ so $\varphi(T)$ is also an admissible null-control for $\mathfrak{S}_{T^*}$ with $T^*>T$; thus $\|\varphi(T)\|\geqq\|\varphi(T^*)\|$ by the optimality of $\varphi(T^*)$. (Indeed, $\|\varphi(T)\|=\|\varphi(T^*)\|$ would imply $\varphi(T)=\varphi(T^*)$ so $\varphi(T^*)$ would vanish for $t>T$; see Remark 11, below.) This shows N is decreasing. Now let $T_j\to T$ and set $\varphi_j=\varphi(T_j)$. By the monotonicity of N, $\|\varphi_j\|$ is bounded by $N(T)$ for all but finitely many j; extract a weakly convergent subsequence $\varphi_{j(k)}\rightharpoonup\varphi_*$. Clearly φ_* vanishes for $t>T$ since, for any $\hat{T}>T$ one has $\varphi_{j(k)}$ vanishing for $t>\hat{T}$ for almost all k. Now choose any $\hat{T}>T$ and let $\mathbf{A}$ be the bounded linear map from boundary data φ, defined for $0\leqq t\leqq\hat{T}$ (note that almost all $\varphi_{j(k)}$ vanish for $t\geqq\hat{T}$—with no loss of generality, we assume this is true for all k—and so may be considered as elements of $L_2(\Sigma_{1,\hat{T}})$), to "terminal" state $u(\hat{T}, \cdot)$ for solutions of (1) with zero initial state; let $\hat{u}$ be the state at $t=\hat{T}$ reached from u_0 with homogeneous boundary conditions so $\mathbf{A}\varphi_{j(k)}=-\hat{u}$ for each k. For any v in $L_2(\Omega)$.

$$\begin{aligned}\langle -\hat{u}, v\rangle &= \langle \mathbf{A}\varphi_{j(k)}, v\rangle \\ &= \langle \varphi_{j(k)}, \mathbf{A}^*v\rangle \to \langle \varphi_*, \mathbf{A}^*v\rangle = \langle \mathbf{A}\varphi_*, v\rangle\end{aligned}$$

so the solution u_* of (1) with φ_* used in (2) and initial state u_0 will vanish at $t=\hat{T}$. Considering u_* on $[T, \hat{T}]\times\Omega$, we note that it satisfies homogeneous boundary conditions on $[T, \hat{T}]\times\Gamma$ so, by uniqueness for the backward heat equation, $u_*\equiv 0$ on $[T, \hat{T}]\times\Omega$. In particular, u_* vanishes at $t=T$ so φ_* is a null-control for the setting $\mathfrak{S}_T$. We have $\|\varphi_*\|\leqq\liminf\|\varphi_{j(k)}\|=\liminf N(T_{j(k)})$. If $T_j\to T+$ then $\limsup N(T_{j(k)})\leqq N(T)=\|\varphi(T)\|$ and if T is a continuity point of $N(\cdot)$ then $\lim N(T_{j(k)})=N(T)=\|\varphi(T)\|$; in either case $\|\varphi_*\|\leqq\|\varphi(T)\|$ which, by the definition of $\varphi(T)$, implies $\|\varphi_*\|=\|\varphi(T)\|$ and $\varphi_*=\varphi(T)$. Since $\|\varphi_{j(k)}\|\to\|\varphi(T)\|$ we have, as before, $\varphi_{j(k)}\to\varphi(T)$ in $L_2(\Sigma_1)$. Since this holds for a sub-subsequence of every subsequence of $\{\varphi_j\}$, one has $\varphi(T_j)\to\varphi(T)$ as $T_j\to T+$ or as $T_j\to T$ with T a continuity point of N. It is clear from this that N is right continuous. Taking $T=0(T_j\to 0+)$, we see that if a bounded sequence $N(T_j)$ were to exist then the same argument would provide a null-control φ_* vanishing for $t>0$ which is impossible if $u_0\neq 0$. □

Remark 10. The uniqueness for the backward (homogeneous) problem on $T \leqq t \leqq T$ is used in the proof. For (α, β) autonomous, this is easy to obtain: letting $\{(e_j, -\lambda_j)\}$ be the eigenpairs for the Laplacian on Ω with homogeneous conditions (2), solutions of (1) with homogeneous boundary conditions have the representation

$$u(t, x) = \Sigma_j c_j \exp[-\lambda_j t]\, e_j(x), \qquad T \leqq t \leqq \hat{T}, \quad x \in \Omega.$$

Since $\{e_j\}$ may be assumed orthonormal, $u(\hat{T}, \cdot) = 0$ implies that each coefficient $c_j = 0$ so also $u(T, \cdot) = 0$. For nonautonomous (α, β) this backward uniqueness must be obtained by another argument (or assumed).

Remark 11. As noted in the proof above, N will be *strictly* decreasing if it can be shown that the optimal control $\varphi(T^*)$ cannot vanish on any nontrivial interval (T, T^*). Using the characterization given by Theorem 3, that an optimal control $\varphi = \varphi(T^*)$ is an element of the subspace $\bar{\mathcal{M}}(= \bar{\mathcal{M}}(T^*))$ in $L_2(\Sigma_{1,T^*})$, we see in the autonomous *one-dimensional* case (Ω an interval in $\mathbb{R}$ so Γ_1 is, e.g., a single endpoint) that φ is a real analytic function of t on $[0, T^*)$ (elements of $\bar{\mathcal{M}}$ are here representable as Dirichlet series—cf., e.g., [16]—and we may apply results of [15]). In n-dimensional situations, however, we see that for (α, β) autonomous the solution semigroup for (1) is analytic (see [2]) so elements of $\mathcal{M}$ are analytic in t but it is not certain that this property holds for the closure $\bar{\mathcal{M}}$. One would like to know that if φ is the limit in $L_2(\Sigma_{1,T})$ of complementary boundary data (given by (3)) of solutions in $[0, T] \times \Omega$ of the adjoint equation

$$-\dot{u} = \Delta u, \tag{31}$$

then there is a solution u of (31) on $[0, T) \times \Omega$ (i.e., on $[0, \hat{T}] \times \Omega$ for every $\hat{T} < T$) whose complementary boundary data is just φ; this would, of course, imply the real analyticity in t of controls, guaranteeing that $\varphi(T^*)$ could not vanish for $T < t < T^*$ unless it were identically zero (impossible unless $u_0 = 0$) so that N would have to be strictly decreasing.

Remark 12. For an autonomous problem, let $\mathbf{S}(\cdot)$ be the solution semigroup associated with homogeneous boundary conditions (2). It is well known that $\mathbf{S}$ is a contraction semigroup with $\|\mathbf{S}(t)\| \leqq e^{-ct}$ (some $c > 0$) if $\alpha \not\equiv 0$ (i.e., other than the Neumann problem). Suppose the setting is null-controllable for every $T > 0$, so for any fixed T there is a constant C_T such that the norm of the null-control φ_* carrying a state u_* at time T_1 to zero at time $(T_1 + T)$ is bounded by $C_1 \|u_*\|$ (as the problem is autonomous, C_T may depend on T but not on T_1). Given any initial state u_0, we may let the problem proceed uncontrolled to time T_1 (i.e., $\varphi = 0$ for $0 < t < T_1$) obtaining the state $u_* = \mathbf{S}(T_1)u_0$ and then control optimally to zero for $T_1 < t \leqq T_1 + T$ by φ_*; thus, the null-control φ (for $0 \leqq t \leqq T_1 + T$) is given by

$$\varphi(t, \cdot) = \begin{cases} 0, & 0 \leqq t < t, \\ \varphi_*(t - T_1, \cdot), & T_1 \leqq t \leqq T_1 + T, \end{cases}$$

and

$$\|\varphi\| = \|\varphi_*\| \leqq C_T \|\mathbf{S}(T_1)u_0\| \leqq C_T e^{-cT_1} \|u_0\|.$$

Since the *optimal* control on $[0, T_1 + T]$ has norm less than that of *this* φ, we see that in this case the norm of the optimal null-control must decay (at least) exponentially as the control interval increases.

REFERENCES

[1] H. O. FATTORINI, *Boundary control of temperature distributions in a parellelopipedon*, this Journal, 13 (1975), pp. 1–13.
[2] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.
[3] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
[4] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, vol. I, Springer-Verlag, New York, 1972.
[5] ———, *Non-Homogeneous Boundary Value Problems and Applications*, vol. II, Springer-Verlag, New York, 1972.
[6] R. C. MACCAMY, V. J. MIZEL AND T. I. SEIDMAN, *Approximate boundary controllability for the heat equation, I*, J. Math. Anal. and Appl., 23 (1968), pp. 699–703.
[7] ———, *Approximate boundary controllability for the heat equation, II*, J. Math. Anal. Appl., 28 (1969), pp. 482–492.
[8] H. MARGENAU AND G. M. MURPHY, *The Mathematics of Physics and Chemistry*, Van Nostrand, New York, 1943.
[9] V. J. MIZEL AND T. I. SEIDMAN, *Observation and prediction for the heat equation*, J. Math. Anal. Appl., 28 (1969), pp. 303–312.
[10] S. MIZOHATA, *Unicité du prolongement des solutions pour quelques opérateurs différentiels paraboliques*, Mem. Coll. Sc. Univ. Kyoto, Sér. A, 31 (1958), pp. 219–239.
[11] J. P. QUINN AND D. L. RUSSELL, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Tech. Sum. Rep. 1575, MRC-Univ. Wisconsin, Madison, 1975.
[12] J. V. RALSTON, *Solutions of the wave equation with localized energy*, Comm. Pure Appl. Math., 22 (1969), pp. 807–823.
[13] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., 52 (1973), pp. 189–211.
[14] ———, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, Differential Games and Control Theory, Marcel Dekker, New York, 1974.
[15] L. SCHWARTZ, *Étude des Sommes d'Exponentielles*, 2nd ed., Hermann, Paris, 1959.
[16] T. I. SEIDMAN, *Observation and prediction for one-dimensional diffusion equations*, J. Math. Anal. Appl., 51 (1975), pp. 165–175.
[17] ———, *Observation and prediction for the heat equation, III*, J. Differential Equations, 20 (1976), pp. 18–27.
[18] ———, *Boundary control and observation for the heat equation*, Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976.
[19] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1965.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

STATIONARY PROBABILITY DISTRIBUTIONS FOR LINEAR TIME-INVARIANT SYSTEMS*

JAKOV SNYDERS†

Abstract. Existence and properties of stationary probability distributions for the output vector of linear time-invariant systems perturbed by white noise are examined. It is shown that a necessary and sufficient condition for existence of such probability distribution is unobservability of the noise controlled unstable modes of the system. In particular, there exists a Gaussian stationary distribution for the output process under the above condition. This is the unique stationary probability distribution if and only if, in addition to the previous condition, all modes corresponding to zero or an imaginary characteristic value are unobservable. Convergence in distribution of the output process is examined, and equivalence of a given system and its dual (transposed system) with respect to existence of stationary probability distributions for their output processes is demonstrated.

1. Introduction. Consider the state vector x, taking values in an n-space $\mathscr{X}_n$, and an output m-vector y satisfying

$$\begin{aligned} dx(t) &= Ax(t)\,dt + B\,dw(t), \qquad x(0) = x_0, \\ y(t) &= Cx(t), \end{aligned} \tag{1}$$

where w is the standard q-dimensional Brownian motion, and x_0 is independent of $\{w(t);\, t \geqq 0\}$. Thus, the output process y is determined by the matrices C, A and B that represent a linear time-invariant system, and by the probability measure $P_{x_0}(\cdot)$ induced on $\mathscr{X}_n$ by the initial state value x_0. This paper deals with stationarity of the probability measure induced on the m-space $\mathscr{X}_m$ by y, i.e. a probability measure satisfying $P_{y_t}(dy) = P_{y_0}(dy)$ for all $t > 0$, where $P_{y_t}(\cdot)$ is the measure induced on $\mathscr{X}_m$ by $y(t)$. Necessary and sufficient conditions for existence of such measure are derived and its properties are investigated.

Stationary probability distributions for a state process were extensively investigated; their existence is often regarded as a kind of stability for the system considered [1], [2], [3], [10]. Results pertinent to the linear system [1] and particularly to the presentation here appeared in [4], [5], and [6], where stationary probability distributions for x or, equivalently, for y under the restriction that C is a nonsingular square matrix, were explored. Dym [4] actually attacked the case of an nth order scalar valued differential equation using state representation, and his results were utilized in [6] for treating the problem that corresponds to any pair (A, B) of matrices. Here we deal with the general case represented by (1).

It is proved in § 3 that a stationary probability distribution for y exists if and only if the noise-perturbed unstable modes of the system are unobservable, and several other equivalent conditions are given. The covariance matrix of such distribution, provided it is finite, is shown to satisfy certain equations that can be reduced, by appropriate transformation of coordinates, to an algebraic Lyapunov equation. In § 4 The structure of possible stationary measures is examined, and in particular conditions for uniqueness are obtained. It is also shown that the system (C, A, B) and its dual (B', A', C'), where a prime denotes transposition, are either

* Received by the editors February 12, 1976, and in revised form July 21, 1976.
† School of Engineering, Tel-Aviv University, Tel-Aviv, Israel.

both stable or both unstable in the sense of existence of a stationary probability distribution for their output processes.

2. Preliminaries. Let F be a map (linear transformation) or any of its matrix representations. The image of F is written $\operatorname{im} F$, and $\ker F = \{x; Fx = 0\}$. The restriction of F to an F-invariant subspace $\mathcal{S}$ is denoted $F|\mathcal{S}$. If $\mathcal{S}$ is some subspace, $\mathcal{S}^{\perp}$ will stand for its orthogonal compliment. Consider now the maps A, B and C in (1). The controllable subspace $\langle A|B\rangle$ and the unobservable subspace $\mathcal{N}(C, A)$ of $\mathcal{X}_n$, associated with (A, B) and (C, A), respectively, are defined by $\langle A|B\rangle = \sum_{j=1}^{n} A^{j-1}(\operatorname{im} B)$ and $\mathcal{N}(C, A) = \bigcap_{j=1}^{n} \ker (CA^{j-1})$. Let ϕ^+, ϕ^0 and ϕ^- be factors of the minimal polynomial of A having roots exclusively with positive, zero and negative real part, respectively. Then $\mathcal{X}_n$ is decomposable into the direct sum $\mathcal{X}_n = \mathcal{X}^+(A) \oplus \mathcal{X}^0(A) \oplus \mathcal{X}^-(A)$, where $\mathcal{X}^x(A) = \ker \phi^x(A)$ with ^{x} standing for any superscript. We write $\mathcal{X}^{0+}(A)$ for $\mathcal{X}^0(A) \oplus \mathcal{X}^+(A)$. Note that in the notations introduced above for various subspaces of $\mathcal{X}_n$ there is no need to indicate explicitly (e.g. by a subscript n) the dimension of the space being considered. $\operatorname{Re}(\lambda)$ and $\operatorname{Im}(\lambda)$ stand, respectively, for the real and imaginary parts of a number λ.

Direct application of Itô's formula to (1) verifies that

$$y(t) = Ce^{At}x_0 + \int_0^t Ce^{A(t-s)}B\, dw(s). \tag{2}$$

The two terms on the right hand side of (2) are independent, and the second one is Gaussian with zero mean and covariance function $K(\cdot)$ given by

$$K(t) = \int_0^t Ce^{As}BB'e^{A's}C'\, ds, \tag{3}$$

where a prime denotes transposition. Let $f(\cdot)$ be the characteristic function of $P_{x_0}(\cdot)$ defined by

$$f(v) = E\{\exp(iv'x)\} = \int_{\mathcal{X}_n} \exp(iv'x) P_{x_0}(dx).$$

Then by (2),

$$E\{\exp(iu'y(t))\} = E\{\exp(iu'Ce^{At}x_0)\} \cdot \exp\{-\tfrac{1}{2}u'K(t)u\} \tag{4}$$

for any $u \in \mathcal{X}_m$. Assuming that the probability measure induced on $\mathcal{X}_m$ by y is stationary, we have

$$E\{\exp(iu'y(t))\} = E\{\exp(iu'Cx_0)\} = f(C'u),$$

and consequently,

$$f(C'u) = f(e^{A't}C'u) \cdot \exp\left\{-\frac{1}{2}u'\left[\int_0^t Ce^{As}BB'e^{A's}C'\, ds\right]u\right\}, \qquad t \geqq 0, \tag{5}$$

for all $u \in \mathcal{X}_m$. Stated otherwise, the characteristic function $g(\cdot)$ of any stationary probability measure induced on $\mathcal{X}_m$ by y may be represented in the form $g(u) = f(C'u)$, where $f(\cdot)$ is some characteristic function in n variables that

satisfies

$$f(v)=f(e^{A't}v)\cdot\exp\left\{-\frac{1}{2}v'\left[\int_0^t e^{As}BB'e^{A's}\,ds\right]v\right\},\qquad t\geqq 0, \tag{6}$$

for all $v\in\operatorname{im}C'$.

Assuming further that x_0 has a (finite) covariance matrix D_0, we see that it follows from (2), that

$$E\{[y(t)-Ey(t)][y(t)-Ey(t)]'\}=Ce^{At}D_0e^{A't}C'+K(t);$$

hence

$$CD_0C'=Ce^{At}D_0e^{A't}C'+\int_0^t Ce^{As}BB'e^{A's}C'\,ds,\qquad t\geqq 0. \tag{7}$$

By differentiation we obtain

$$Ce^{At}(AD_0+D_0A'+BB)e^{A't}C'=0,\qquad t\geqq 0, \tag{8}$$

and, conversely, (8) implies (7). Consequently, (7) is equivalent to the familiar Lyapunov equation

$$AD_0+D_0A'+BB'=0 \tag{9}$$

if and only if $\mathcal{N}(C,A)=0$. Thus (9) is applicable for obtaining any possible covariance matrix of a stationary probability distribution for x, and in turn for y, whenever $\mathcal{N}(C,A)=0$. There exists a nonnegative definite solution D_0 to (9) if and only if $\mathscr{R}^-(A)\supset\langle A|B\rangle$, and uniqueness of this solution is assured if and only if, in addition to the previous condition, A has no characteristic value with zero real part [7]. As a general approach for handling (7), one perhaps tends to replace (9) by

$$C(AD_0+D_0A'+BB')C'=0.$$

However, this equation is not equivalent to (7) unless $\ker C=0$ and, moreover, it may possess a nonnegative definite solution even if there is no nonnegative definite solution to (7), as seen by Lemma 1 of the next section and the example $C=(1\ \ 1)$, $A=\operatorname{diag}(-1,1)$, $B=(1\ \ 1)'$, $D_0=\operatorname{diag}(2,0)$.

It follows that for the general case a time-independent counterpart of (9), stated explicitly in terms of A, B and C is not available, excluding, of course, some set of n equations derived from (8). An alternative that may be useful is the following: if $D(\cdot)$ is a function satisfying

$$\frac{d}{dt}D(t)=AD(t)+D(t)A'+BB',\qquad D(0)=D_0, \tag{10}$$

$$\frac{d}{dt}CD(t)C'=0, \tag{11}$$

then D_0 solves (7) and, conversely, for any solution D_0 to (7) there exists a function $D(\cdot)$ satisfying (10) and (11). Indeed, by (10),

$$\frac{d}{ds}\{e^{A(t-s)}D(s)e^{A'(t-s)}\}=e^{A(t-s)}BB'e^{A'(t-s)};$$

integration yields

$$D(t)=e^{At}D_0e^{A't}+\int_0^t e^{A(t-s)}BB'e^{A'(t-s)}\,ds, \tag{12}$$

and (7) follows according to (11). Conversely, if D_0 solves (7) then $D(\cdot)$ defined by (12) satisfies (10) and (11). Obviously, if in (12) D_0 stands for the covariance matrix of x_0, then $D(t)$ is the covariance matrix of $x(t)$, $t>0$.

3. Existence of a stationary measure. It is straightforward to check that if D_0 is a nonnegative definite solution to (7), then a Gaussian random vector x_0 with zero mean and covariance D_0 induces, through the resulting output process y, a stationary probability measure on $\mathscr{X}_m$. Also, existence of a nonnegative definite D_0 is obviously necessary for existence of a finite-covariance stationary probability measure that is induced on $\mathscr{X}_m$ by y. The last statement is strengthened by Lemma 1 and Theorem 1 below, namely, the condition applies to all stationary probability measures for y. Hence, as anticipated by linearity of the system and the Gaussian nature of the input, existence of any stationary probability distribution for y implies the existence of a Gaussian one. It is appropriate to mention that, in general, there is no need to set $Ex_0=0$ for obtaining a constant-mean (e.g. Gaussian) process: $Ey(t)$ is constant if and only if $Ex_0\in\mathscr{N}(CA, A)$ i.e. if and only if $A(Ex_0)\in\mathscr{N}(C, A)$. In particular this condition is satisfied if Ex_0 is unobservable.

LEMMA 1. *The following conditions are equivalent*:

(a) *There exists a nonnegative definite matrix* D_0 *satisfying* (7).

(b) $\lim_{t\to\infty}\int_0^t Ce^{As}BB'^{A's}C'ds$ *exists.*

(c) $\lim_{t\to\infty} Ce^{At}B=0$.

(d) $\ker C\supset\mathscr{X}^{0+}(A)\cap\langle A|B\rangle$.

Proof. Trivially (a) implies (b) and (b) implies (c). Now suppose that (d) does not hold. Since $\mathscr{X}^{0+}(A)\cap\langle A|B\rangle$ is an A invariant subspace of $\mathscr{X}_n$, there exists $x\in\langle A|B\rangle$ such that $Ax=\lambda x$ where $\mathrm{Re}(\lambda)\geqq 0$ and $Cx\neq 0$. Thus $Cx^{At}x=e^{\lambda t}Cx$ does not converge to zero as $t\to\infty$. However, we also have $x=\sum_{j=1}^n A^{j-1}Bz_j$ for some $z_j\in\mathscr{X}_q$, $j=1, 2, \cdots, n$, yielding

$$Ce^{At}x=\sum_{j=0}^{n-1}\frac{d^j}{dt^j}(Ce^{At}B)z_{j+1}.$$

Assuming that (c) is satisfied, each entry of a time derivative of $Ce^{At}B$ is a linear combination of decreasing exponentials with polynomial coefficients, and consequently $\lim_{t\to\infty} Ce^{At}x=0$. For completing the proof it remains to show that (d) implies (a). Selecting a basis for $\mathscr{X}_n$ such that

$$C=(0\quad C_2),\qquad A=\begin{pmatrix}A_{11} & A_{12}\\ 0 & A_{22}\end{pmatrix},\qquad B=\begin{pmatrix}B_1\\ B_2\end{pmatrix}, \tag{13}$$

where $A_{11}=A|\mathscr{X}^{0+}(A)\cap\langle A|B\rangle$, and setting

$$D_0=\begin{pmatrix}D_{11} & D_{12}\\ D_{12}' & D_{22}\end{pmatrix} \tag{14}$$

into (7) we see that

$$(15)\quad C_2D_{22}C_2' = C_2e^{A_{22}t}D_{22}e^{A_{22}'t}C_2' + \int_0^t C_2e^{A_{22}s}B_2B_2'e^{A_{22}'s}C_2'\,ds, \qquad t \geqq 0.$$

This equation is solved by $D_{22} = \int_0^\infty e^{A_{22}s}B_2B_2'^{A_{22}'s}\,ds$, where the integral converges since $A_{22}|\langle A_{22}|B_2\rangle$ is stable according to (d). D_{11} and D_{12} in (14) are set equal to zero.

As demonstrated by the proof, in general (7) possesses many nonnegative definite solutions: D_{11} and D_{12} may be arbitrarily selected as long as nonnegative definiteness of D_0 is maintained, and (15) may possess many solutions, partly because ker C_{22} may be different from zero. For a given D_0 equations (10) and (11) have a solution, obviously unique, if and only if (d) holds. Moreover, it is easily seen that under this condition (10) and (11) can be reduced, by appropriate selection of coordinates, to a time-independent Lyapunov equation that is applicable for obtaining any possible covariance matrix of a stationary probability measure for y.

THEOREM 1. *The process defined by* (1) *possesses a stationary probability measure if and only if*

$$(16)\qquad \ker C \supset \mathscr{X}^{0+}(A) \cap \langle A\,|\,B\rangle$$

or, equivalently, if and only if

$$(17)\qquad \operatorname{im} B \subset \mathscr{X}^{-}(A) + \mathscr{N}(C, A).$$

Proof. Assume that the measure induced by y on $\mathscr{X}_m$ is stationary and (16) is violated. Then by Lemma 1 there exists $z \in \mathscr{X}_m$ such that $z'(\int_0^t Ce^{As}BB'e^{A's}C'\,ds)z$ does not remain bounded as $t \to \infty$. Setting $u = \alpha z$ into (5) where α is a real nonzero number we thus conclude $f(\alpha C'z) = 0$. Hence by continuity of characteristic functions in the real variable [8, p. 194] it follows that $f(0) = 0$, in contradiction of $f(0) = 1$. Thus, the necessity of (16) is proved. Sufficiency of (16) follows by Lemma 1 and the discussion preceding it. For showing the equivalence of (16) and (17), we observe first that ker C and im B may be replaced, respectively, by $\mathscr{N}(C, A)$ and $\langle A|B\rangle$. Assuming (16) we obtain the modified version of (17) as follows:

$$(18)\qquad \langle A|B\rangle = \mathscr{X}^{-}(A) \cap \langle A|B\rangle \oplus \mathscr{X}^{0+}(A) \cap \langle A|B\rangle \subset \mathscr{X}^{-}(A) + \mathscr{N}(C, A),$$

whereas if (17) holds then

$$\mathscr{X}^{0+}(A) \cap \langle A|B\rangle \subset \mathscr{X}^{0+}(A) \cap [\mathscr{X}^{-}(A) + \mathscr{N}(C, A)] = \mathscr{X}^{0+}(A) \cap \mathscr{N}(C, A),$$

and consequently

$$(19)\qquad \mathscr{X}^{0+}(A) \cap \langle A|B\rangle \subset \mathscr{N}(C, A).$$

Interpretation of Theorem 1 is straightforward, and may be stated in view of (18) and (19) as follows: there exists a stationary probability distribution for the output process of a linear system if and only if every controllable vector is the sum of a stable vector and an unobservable vector or, equivalently, every vector that is both controllable and unstable is unobservable. It should be noted that these

conditions are also necessary and sufficient for existence of a second-order stationary output process [9].

4. Some properties. Since $\lim_{t\to\infty} Ce^{At}B = 0$ if and only if $\lim_{t\to\infty} B'e^{A't}C' = 0$ it follows by Lemma 1 that the system (C, A, B) and its "dual", i.e. transposed system (B', A', C') have the following common stability property.

THEOREM 2. *Let the process $\tilde{y}$ be given by*

$$\begin{aligned} d\tilde{x}(t) &= A'\tilde{x}(t)\,dt + C'\,d\tilde{w}(t), \qquad \tilde{x}(0) = \tilde{x}_0, \\ \tilde{y}(t) &= B'\tilde{x}(t), \end{aligned} \tag{20}$$

where A, B, C are as in (1) *and $\tilde{w}$ is the standard q-dimensional Brownian motion such that $\tilde{x}_0$ is independent of $\{\tilde{w}(t);\, t \geqq 0\}$. Either both y of* (1) *and $\tilde{y}$ of* (20) *or none of these processes possess a stationary probability measure.*

Note that equivalence of (16) and (17) is related to the above result. By Lemma 1, (16) is satisfied if and only if

$$\ker B' \supset \mathscr{X}^{0+}(A') \cap \langle A' | C' \rangle,$$

and since

$$(\ker B')^{\perp} = \operatorname{im} B, \qquad [\mathscr{X}^{0+}(A') \cap \langle A' | C' \rangle]^{\perp} = \mathscr{X}^{-}(A) + \mathscr{N}(C, A),$$

the conclusion follows.

Let $\mathscr{S} = \mathscr{N}(CA, A) + \sum_{\lambda>0} \mathscr{N}(C(A^2 + \lambda^2 I), A)$ where I is the identity matrix. Then $\mathscr{S}$ is an A-invariant subspace of $\mathscr{X}_n$ and $\mathscr{S} = \mathscr{X}^0(A) \cap \mathscr{S} + \mathscr{N}(C, A)$. Hence $C(\mathscr{S}) = C(\mathscr{X}^0(A) \cap \mathscr{S}) \subset C(\mathscr{X}^0(A))$.

THEOREM 3. *Let g be the characteristic function of a stationary probability measure for y of* (1). *Then $g(u) = g_1(u)g_2(u)$, where*

$$g_1(u) = \exp\left\{-\frac{1}{2}u\left(\int_0^{\infty} Ce^{At}BB'e^{A't}C'\,dt\right)u\right\} \tag{21}$$

and g_2 is the characteristic function of a stationary probability measure for y of (1) *with $B = 0$. Conversely, if g_1 and g_2 are as above, then their product g_1g_2 is the characteristic function of a stationary probability measure for y of* (1). *Furthermore, the support of the measure corresponding to g_1 is $C(\langle A | B \rangle)$, and the support of the measure corresponding to g_2 is included in $C(\mathscr{S})$.*

Proof. Assuming existence of a stationary probability measure, write $g(u) = f(C'u)$ where f is a suitable characteristic function. According to Lemma 1 and Theorem 1 the integral in (21) converges, and by (5),

$$f(C'u) = g_1(u) \cdot \lim_{t\to\infty} f(e^{A't}C'u).$$

Set $g_2(u) = f_2(C'u)$ where $f_2(C'u) = \lim_{t\to\infty} f(e^{A't}C'u)$. Then g_2 is a characteristic function in view of the continuity theorem [8, p. 191], and f_2 obviously satisfies the following version of (6):

$$f_2(v) = f_2(e^{A't}v), \qquad t \geqq 0, \tag{22}$$

for all $v \in \operatorname{im} C'$. Turning to prove the converse statement let $g_2(u) = f_2(C'u)$

where f_2 is a suitably selected characteristic function satisfying (22). Then $g(u)=f(C'u)$ where f is a characteristic function, and

$$f(e^{A't}C'u)=g_2(u)\cdot\exp\left\{-\frac{1}{2}u'\left(\int_t^\infty Ce^{As}BB'e^{A's}C'\,ds\right)u\right\}$$

$$=g_2(u)g_1(u)\cdot\exp\left\{\frac{1}{2}u'\left(\int_0^t Ce^{As}BB'e^{A's}C'\,ds\right)u\right\}.$$

Consequently f satisfies (6).

We shall examine the support of the measure corresponding to g_1 by selecting a basis for $\mathscr{X}_n$ such that

$$C=(0\quad C_2\quad C_3),\qquad A=\begin{pmatrix}A_{11}&0&A_{13}\\0&A_{22}&A_{23}\\0&0&A_{33}\end{pmatrix},\qquad B=\begin{pmatrix}B_1\\B_2\\0\end{pmatrix},$$

where

$$A_{11}=A|\mathscr{X}^{0+}(A)\cap\langle A|B\rangle,\qquad A_{22}=A|\mathscr{X}^{-}(A)\cap\langle A|B\rangle.$$

Then $g_1(u)=\exp\{-\frac{1}{2}u'C_2LC_2'u\}$ where $L=\int_0^\infty e^{A_{22}t}B_2B_2'e^{A_{22}'t}\,dt$ is positive definite. The characteristic function $\psi(v)=\exp\{-\frac{1}{2}v'Lv\}$ corresponds to a zero mean Gaussian measure with support $\mathscr{X}^{-}(A)\cap\langle A|B\rangle$; hence the conclusion follows according to

$$C(\langle A|B\rangle)=C(\mathscr{X}^{-}(A)\cap\langle A|B\rangle\oplus\mathscr{X}^{0+}(A)\cap\langle A|B\rangle)$$

$$=C(\mathscr{X}^{-}(A)\cap\langle A|B\rangle).$$

Assume now $B=0$ and observe that (22) implies

$$f_2(v)=f_2(e^{A't}v) \tag{23}$$

for every real t and $v\in\langle A'|C'\rangle$. Setting $v\in\mathscr{X}^{-}(A')\cap\langle A'|C'\rangle$ and $v\in\mathscr{X}^{+}(A')\cap\langle A'|C'\rangle$ and evaluating the limit as $t\to\infty$ and $t\to-\infty$, respectively, it follows in both cases that $f_2(v)=f_2(0)=1$. Hence the support of the measure corresponding to g_2 is included in

$$C(\{[\mathscr{X}^{-}(A')\oplus\mathscr{X}^{+}(A')]\cap\langle A'|C'\rangle\}^{\perp})=C(\mathscr{X}^{0}(A)+\mathscr{N}(C,A))=C(\mathscr{X}^{0}(A)).$$

For proving the stronger statement claimed in the theorem, consider a chain of generalized characteristic vectors p_j; $j=1,2,\cdots,s$, of A', i.e.,

$$A'p_1=i\lambda p_1,$$

$$A'p_j=i\lambda p_j+p_{j-1},\qquad j=2,3,\cdots,s,$$

where λ is real. For a member p_k of this chain

$$e^{A't}e^{-i\lambda t}p_k=p_k+tp_{k-1}+\frac{t^2}{2}p_{k-2}+\cdots+\frac{t^{k-2}}{(k-2)!}p_2+\frac{t^{k-1}}{(k-1)!}p_1,$$

and if $\mathrm{Re}\,(p_k),\mathrm{Im}\,(p_k)\in\langle A'|C'\rangle$ then $\mathrm{Re}\,(p_j),\mathrm{Im}\,(p_j)\in\langle A'|A'C'\rangle$ provided $\lambda=0$, whereas otherwise $\mathrm{Re}\,(p_j),\mathrm{Im}\,(p_j)\in\langle A'|(A^2+\lambda^2 I)'C'\rangle$, for $j=1,2,\cdots,k-1$.

According to (23),

$$f_2\Big(\frac{(k-1)!}{t^{k-1}}\operatorname{Re}\{e^{-i\lambda t}p_k\}\Big)=f_2\Big(\operatorname{Re}\Big\{\frac{(k-1)!}{t^{k-1}}p_k+\frac{(k-1)!}{t^{k-2}}p_{k-1}+\cdots+\frac{k-1}{t}p_2+p_1\Big\}\Big),$$

and by letting $t\to\infty$ this equation yields

$$f_2(\operatorname{Re}(p_1))=f_2(0)=1. \tag{24}$$

Similarly

$$\begin{aligned}f_2\Big(\frac{(k-2)!}{t^{k-2}}\operatorname{Re}\{e^{-i\lambda t}p_k\}\Big)&=f_2\Big(\operatorname{Re}\Big\{\frac{(k-2)!}{t^{k-2}}p_k+\frac{(k-2)!}{t^{k-3}}p_{k-1}+\cdots+p_2+\frac{t}{k-1}p_1\Big\}\Big)\\&=f_2\Big(\operatorname{Re}\Big\{\frac{(k-2)!}{t^{k-2}}p_k+\frac{(k-2)!}{t^{k-3}}p_{k-1}+\cdots+p_2\Big\}\Big),\end{aligned}$$

where the second equality is due to (24), and by letting $t\to\infty$ it follows that $f_2(\operatorname{Re}(p_2))=1$. This procedure shows that $f_2(\operatorname{Re}(p_j))=1$, $j=1,2,\cdots,k-1$, and the same technique yields $f_2(\operatorname{Im}(p_j))=1$, $j=1,2,\cdots,k-1$. Now $\langle A'|(A^2+\lambda^2 I)'C'\rangle^{\perp}=\mathcal{N}(C(A^2+\lambda^2 I),A)+\ker\phi(\lambda;A)$ if $\lambda\neq 0$, and a similar equality holds for $\langle A'|A'C'\rangle^{\perp}$, where $\phi(\lambda;\cdot)$ is the minimal polynomial of A divided by its factors having roots $\pm i\lambda$; hence the conclusion follows.

For obtaining further details about the support of the measure corresponding to g_2 we use a representation

$$C=(0\quad C_2),\qquad A=\begin{pmatrix}A_{11}&A_{12}\\0&A_{22}\end{pmatrix},$$

where $A_{11}=A|\mathcal{N}(C,A)$. The measure's covariance matrix, provided it exists, is given by C_2DC_2', where D is a nonnegative definite solution to the equation

$$A_{22}D+DA_{22}'=0, \tag{25}$$

and the support of the measure is contained in a hyperplane of $\mathcal{X}_m$ with dimension rank (C_2DC_2'). Investigation of (25) reveals [7] that

$$\operatorname{rank}(C_2DC_2')=\operatorname{rank}(D)=\alpha+2\beta,$$

where α and β are integers satisfying

$$0\leqq\alpha\leqq\dim(\ker A_{22}),$$

$$0\leqq 2\beta\leqq\dim\Big(\sum_{\lambda>0}\ker(A_{22}^2+\lambda^2 I)\Big),$$

and evidently

$$\dim(\ker A_{22})=\dim(\mathcal{N}(CA,A))-\dim(\mathcal{N}(C,A)),$$

$$\dim\Big(\sum_{\lambda>0}\ker(A_{22}^2+\lambda^2 I)\Big)=\dim(\mathcal{S})-\dim(\mathcal{N}(CA,A)).$$

We also have the following result, obtainable by considering Gaussian measures with covariance matrices that satisfy (25).

THEOREM 4. *Let y be given by* (1) *with $B=0$, let $\mathscr{P}$ and $\mathscr{Q}(\lambda)$, $\lambda>0$ be subspaces of $\mathscr{N}(CA, A)$ and $\mathscr{N}(C(A^2+\lambda^2 I), A)$, respectively, and let $z \in C(\mathscr{N}(CA, A))$. There exists a stationary probability measure η for y such that the support of η shifted by z is $C(\mathscr{P}+\sum_{\lambda>0}\mathscr{Q}(\lambda))$. In particular, there exists a stationary probability measure for y with support $C(\mathscr{S})$.*

By Theorem 3 and Theorem 4, there exists a *unique* stationary probability distribution for y of (1) if and only if, in addition to (16), $C(\mathscr{S})=0$.

COROLLARY 1. *There exists a unique stationary probability distribution for y of* (1) *if and only if*

$$\ker C \supset \mathscr{X}^0(A) \oplus [\mathscr{X}^+(A) \cap \langle A|B\rangle].$$

Furthermore, this distribution is Gaussian with zero mean and covariance $K(\infty)$, where K is given by (3).

Let $g_{t,x_0}(\cdot)$ be the characteristic function of the transition probability for y, i.e., $g_{t,x_0}(u) = E\{\exp iu'y(t) | x(0)=x_0\}$. Then

$$g_{t,x_0}(u) = \exp\left(iu'Ce^{At}x_0 - \tfrac{1}{2}u'K(t)u\right),$$

where $K(t)$ is given by (3). Hence the probability distribution of $y(t)$ converges as $t\to\infty$ for all $x_0 \in \mathscr{X}_n$ if and only if both $\mathscr{N}(CA, A) \supset \mathscr{X}^{0+}(A)$ and (16) are satisfied. This yields the following result.

THEOREM 5. *$y(t)$ given by* (1) *converges in distribution as $t\to\infty$ for any $y(0) \in \operatorname{im} C$ if and only if*

$$\ker C \supset \mathscr{X}^+(A) \oplus [\mathscr{X}^0(A) \cap \langle A|B\rangle] \tag{26}$$

and

$$\ker CA \supset \mathscr{X}^0(A). \tag{27}$$

COROLLARY 2. *There exists a probability distribution μ such that $y(t)$ given by* (1), *starting at any $y(0) \in \operatorname{im} C$, converges in distribution to μ as $t\to\infty$ if and only if* $\ker C \supset \mathscr{X}^{0+}(A)$.

Note that $\mathscr{X}^0(A)$ and $\mathscr{X}^+(A)$ in (26) and (27) may be replaced, respectively, by $\ker A^j$ and $\mathscr{X}^c(A)$, where j is the order of the zero factor in the minimal polynomial of A, and $\mathscr{X}^{0+}(A) = \mathscr{X}^c(A) \oplus \ker A^j$. Of course, (26) and (27) are *not* necessary for convergence in distribution of $y(t)$ as $t\to\infty$ for all $x(0) \in \mathscr{M}$, where $\mathscr{M}$ is a set of initial state values such that $\operatorname{im} C = C(\mathscr{M})$. Instead, it is necessary and sufficient to have (16) and

$$\operatorname{im} C = C(\mathscr{X}^-(A) + \mathscr{N}(CA, A)). \tag{28}$$

Obviously, (28) is a considerably milder requirement than $C(\mathscr{S})=0$.

Acknowledgment. The author wishes to thank one of the reviewers for suggesting several improvements, and for bringing reference [10] to his attention.

REFERENCES

[1] W. M. WONHAM, *Lyapunov criteria for weak stochastic stability*, J. Differential Equations, 2 (1966), pp. 195–207.

[2] H. J. KUSHNER, *The Cauchy problem for a class of degenerate parabolic equations and asymptotic properties of the related diffusion processes*, Ibid., 6 (1969), pp. 209–231.

[3] M. ZAKAI, *A Lyapunov criterion for the existence of stationary probability distributions for systems perturbed by noise*, this Journal, 3 (1969), pp. 390–397.
[4] H. DYM, *Stationary measures for the flow of a linear differential equation driven by white noise*, Trans. Amer. Math. Soc., 123 (1966), pp. 130–164.
[5] M. ZAKAI AND J. SNYDERS, *Stationary probability measures for linear differential equations driven by white noise*, J. Differential Equations, 1 (1970), pp. 27–33.
[6] R. V. ERIKSON, *Constant coefficient linear differential equations driven by white noise*, Ann. Math. Statis., 2 (1971), pp. 820–823.
[7] J. SNYDERS, *On nonnegative solutions of the equation* $AD + DA' = -C$, SIAM J. Appl. Math., 3 (1970), pp. 704–713.
[8] M. LOÈVE, *Probability Theory*, Van Nostrand, Princeton, NJ, 1963.
[9] J. SNYDERS, *Error expressions for optimal stationary state estimation*, this Journal, 5 (1975), pp. 1093–1102.
[10] Y. MIYAHARA, *Invariant measures of ultimately bounded stochastic processes*, Nagoya Math. J., 49 (1973), pp. 149–153.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

MONOTONE MAPPINGS WITH APPLICATION IN DYNAMIC PROGRAMMING*

DIMITRI P. BERTSEKAS†

Abstract. The structure of many sequential optimization problems over a finite or infinite horizon can be summarized in the mapping that defines the related dynamic programming algorithm. In this paper we take as a starting point this mapping and obtain results that are applicable to a broad class of problems. This approach has also been taken earlier by Denardo under contraction assumptions. The analysis here is carried out without contraction assumptions and thus the results obtained can be applied, for example, to the positive and negative dynamic programming models of Blackwell and Strauch. Most of the existing results for these models are generalized and several new results are obtained relating mostly to the convergence of the dynamic programming algorithm and the existence of optimal stationary policies.

1. Introduction. It is well known that dynamic programming (D.P. for short) is the principal method for analysis of sequential optimization problems. It is also known that it is possible to describe each iteration of a D.P. algorithm by means of a certain mapping which maps the set of extended real-valued functions defined on the state space into itself. In problems with a finite, say N, number of stages, after N successive applications of this mapping (i.e. after N steps of the D.P. algorithm) one obtains the optimal value function of the problem. In problems with an infinite number of stages one hopes that the sequence of functions generated by successive application of the D.P. iteration converges in some sense to the optimal value function of the problem. Furthermore it is possible to define the optimization problem itself in terms of the underlying mapping.

To illustrate this viewpoint let us consider formally the deterministic optimal control problem of finding a control law, i.e. a finite sequence of control functions, $\pi = \{\mu_0, \mu_1, \cdots, \mu_{N-1}\}$ which minimizes

$$J_\pi(x_0) = \sum_{k=0}^{N-1} g[x_k, \mu_k(x_k)] \tag{1}$$

subject to the system equation constraint

$$x_{k+1} = f[x_k, \mu_k(x_k)], \qquad k = 0, 1, \cdots, N-1. \tag{2}$$

The states x_k belong to a state space S and the controls $\mu_k(x_k)$ are elements of a control space C. The initial state x_0 is known and f, g are given functions. The D.P. algorithm for this problem is given by

$$J_0(x) = 0, \tag{3}$$

$$J_{k+1}(x) = \inf_u \{g(x, u) + J_k[f(x, u)]\}, \qquad k = 0, \cdots, N-1, \tag{4}$$

and the optimal value of the problem $J^*(x_0)$ is obtained at the Nth step of the D.P. algorithm

$$J^*(x_0) = \inf_\pi J_\pi(x_0) = J_N(x_0).$$

* Received by the editors December 3, 1974, and in revised form July 1, 1976.

† Department of Electrical Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801. This work was carried out at the Coordinated Science Laboratory and was supported by the National Science Foundation under Grant ENG 74-19332.

One may also obtain the value $J_\pi(x_0)$ corresponding to any $\pi = \{\mu_0, \mu_1, \cdots, \mu_{N-1}\}$ by means of the algorithm

$$J_{0,\pi}(x) = 0, \tag{5}$$

$$J_{k+1,\pi}(x) = g[x, \mu_k(x)] + J_{k,\pi}[f(x, \mu_k(x))], \quad k = 0, \cdots, N-1, \tag{6}$$

$$J_\pi(x_0) = J_{N,\pi}(x_0).$$

Now it is possible to formulate the problem above as well as to describe the D.P. algorithm (3), (4) by means of the mapping H given by

$$H(x, u, J) = g(x, u) + J[f(x, u)]. \tag{7}$$

Let us define the mapping T by

$$T(J)(x) = \inf_u H(x, u, J), \tag{8}$$

and for any function $\mu: S \to C$ the mapping T_μ by

$$T_\mu(J)(x) = H[x, \mu(x), J]. \tag{9}$$

Both T and T_μ map the set of real-valued (or perhaps extended real-valued) functions on S into itself. Then in view of (5), (6) we may write the cost functional $J_\pi(x_0)$ of (1) as

$$J_\pi(x_0) = (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}})(J_0)(x_0), \tag{10}$$

where J_0 is the zero function on $S(J_0(x) = 0, \forall x \in S)$, and $(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}})$ denotes the composition of the mappings $T_{\mu_0}, T_{\mu_1}, \cdots, T_{\mu_{N-1}}$. Similarly the D.P. algorithm (3), (4) may be described by

$$J_{k+1}(x) = T(J_k)(x), \qquad k = 0, 1, \cdots, N-1, \tag{11}$$

and we have

$$J^*(x_0) = \inf_\pi J_\pi(x_0) = T^N(J_0)(x_0), \tag{12}$$

where T^N is the composition of T with itself N times.

One may consider also an infinite horizon version of the deterministic problem above whereby we seek a sequence $\pi = \{\mu_0, \mu_1, \cdots\}$ that minimizes

$$J_\pi(x_0) = \lim_{N \to \infty} \sum_{k=0}^{N-1} g[x_k, \mu_k(x_k)] \tag{13}$$

subject to the system equation constraint (2). In this case one needs, of course, to make assumptions which ensure that the limit in (13) is well defined for each π and x_0. A primary question of interest is whether the optimal value function J^* satisfies Bellman's functional equation

$$J^*(x) = \inf_u \{g(x, u) + J^*[f(x, u)]\}$$

or equivalently whether

$$J^*(x) = T(J^*)(x) \quad \forall x \in S,$$

and J^* is a fixed point of the mapping T. This question has been answered in the affirmative for broad classes of problems [1], [3], [6], [11]. Other questions relate to the existence and characterization of optimal policies. It is also of both computational and analytical interest to know whether

$$J^*(x) = \lim_{N\to\infty} T^N(J_0)(x) \quad \forall x \in S. \tag{14}$$

When (14) holds, the D.P. algorithm yields in the limit the optimal value function of the problem. While (14) holds in discounted and positive dynamic programming models [1], [11], it has been proved only under restrictive finiteness assumptions for the negative model of Strauch (see [11, Thm. 9.1]). In fact for such models (14) may easily fail to hold as the following example shows:

Example. Let $S = [0, \infty)$, $C = (0, \infty)$ be the state and control spaces respectively. Let the system equation be

$$x_{k+1} = 2x_k + u_k, \qquad k = 0, 1, \cdots,$$

and let the cost per stage be defined by

$$g(x, u) = x.$$

Then it can be easily verified that

$$J^*(x) = \inf_{\pi} J_\pi(x) = +\infty \quad \forall x \in S$$

while

$$T^N(J_0)(0) = 0 \quad \forall N = 1, 2, \cdots.$$

The deterministic optimization problem described above is representative of a plethora of sequential optimization problems of practical interest which may be formulated in terms of mappings similar to the mapping H of (7). A class of such problems has been formulated and analyzed by Denardo [4]. His framework however is restricted by contraction and boundedness assumptions which preclude the use of his results in many types of problems including the positive and negative models of Blackwell [3] and Strauch [11]. The purpose of this paper is to provide a broader framework than the one of Denardo which includes in particular positive and negative models. Questions such as those described above for the deterministic problem are analyzed in this broader setting. Most of the existing results on positive and negative models are generalized. Some entirely new results are also obtained, most notably a necessary and sufficient condition for convergence of the D.P. algorithm (Proposition 11). This result yields in turn powerful a priori verifiable sufficient conditions for convergence of the D.P. algorithm as well as for existence of an optimal stationary policy (Proposition 12). Since under our assumptions we cannot rely on contraction properties, the line of analysis is entirely different from the one of Denardo and utilizes primarily the monotonicity of the mappings involved.

2. Notation and assumptions. The following notational conventions will be used throughout the paper:

1. S and C are two given nonempty sets referred to as the *state space* and *control space* respectively.

2. For each $x \in S$ there is given a nonempty subset $U(x)$ of C referred to as the *control constraint set at x.*

3. We denote by M the set of all functions $\mu: S \to C$ such that $\mu(x) \in U(x)$ for all $x \in S$. We denote by Π the set of all sequences $\pi = \{\mu_0, \mu_1, \cdots\}$ such that $\mu_k \in M$ for all k. Elements of Π are referred to as *policies.* Elements of Π of the form $\pi = \{\mu, \mu, \cdots\}$ where $\mu \in M$ are referred to as *stationary policies.*

4. We denote

F: The set of all extended real valued functions $J: S \to [-\infty, \infty]$.

B: The Banach space of all bounded real-valued functions $J: S \to (-\infty, \infty)$ with the sup norm $\|\cdot\|$ defined by

$$\|J\| = \sup_{x \in S} |J(x)| \quad \forall J \in B.$$

The unit function in F will be denoted e $[e(x) = 1, \forall x \in S]$.

5. For all $J, J' \in F$ we write

$$J = J' \quad \text{if } J(x) = J'(x) \quad \forall x \in S,$$

$$J \leqq J' \quad \text{if } J(x) \leqq J'(x) \quad \forall x \in S.$$

6. For any sequence $\{J_k\}$ with $J_k \in F$ for all k we denote by $\lim_{k\to\infty} J_k$ the pointwise limit of $\{J_k\}$ (assuming it is well defined as an extended real-valued function), and by $\liminf_{k\to\infty} J_k$ the pointwise limit inferior of $\{J_k\}$. Throughout the paper the convergence analysis is carried out within the set of extended real numbers, i.e. $+\infty$ or $-\infty$ are allowed as limits of sequences of extended real numbers. For any collection $\{J_a | a \in A\} \subset F$ parameterized by the elements of a set A we denote $\inf_{\alpha\in A} J_a$ the function taking value $\inf_{\alpha\in A} J_\alpha(x)$ at each $x \in S$.

7. We are given a mapping $H: S \times C \times F \to [-\infty, +\infty]$ and we define for each $\mu \in M$ the mapping $T_\mu: F \to F$ by

$$T_\mu(J)(x) = H[x, \mu(x), J] \quad \forall x \in S. \tag{15}$$

We define also the mapping $T: F \to F$ by

$$T(J)(x) = \inf_{u \in U(x)} H(x, u, J) \quad \forall x \in S. \tag{16}$$

We denote by T^k, $k = 1, 2, \cdots$, the composition of T with itself k times. For convenience we also define $T^0(J) = J$ for all $J \in F$. For any $\pi = \{\mu_0, \mu_1, \cdots\} \in \Pi$ we denote by $(T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k})$ the composition of the mappings $T_{\mu_0}, \cdots, T_{\mu_k}$, $k = 0, 1, \cdots$.

The following assumption will be in effect throughout the paper.

Monotonicity assumption. There holds for every $x \in S$, $u \in U(x)$, $J, J' \in F$,

$$H(x, u, J) \leqq H(x, u, J') \quad \text{if } J \leqq J'.$$

Notice that the monotonicity assumption implies the following relations:

$$J \leqq J' \Rightarrow T(J) \leqq T(J') \qquad \forall J, J' \in F,$$

$$J \leqq J' \Rightarrow T_\mu(J) \leqq T_\mu(J') \quad \forall J, J' \in F, \quad \mu \in M.$$

We shall make frequent use of these relations.

3. Problem formulation. We are given a function $\bar{J} \in F$ satisfying

$$\bar{J}(x) > -\infty \quad \forall x \in S$$

and we consider for every $\pi = \{\mu_0, \mu_1, \cdots\} \in \Pi$ the function $J_\pi \in F$ defined by

$$J_\pi(x) = \lim_{N \to \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}})(\bar{J})(x) \quad \forall x \in S. \tag{17}$$

We refer to J_π as the *value function of* π. Under the assumptions that we will introduce shortly J_π is well defined. Throughout the paper we will be concerned with the optimization problem

$$\text{minimize } J_\pi(x) \quad \text{subject to } \pi \in \Pi. \tag{18}$$

The optimal value of this problem for a fixed $x \in S$ is denoted by $J^*(x)$,

$$J^*(x) = \inf_{\pi \in \Pi} J_\pi(x). \tag{19}$$

We refer to the function $J^* \in F$ as the *optimal value function*. We say that a policy $\pi^* \in \Pi$ is *optimal at* $x \in S$ if $J_{\pi^*}(x) = J^*(x)$ and we say that a policy $\pi^* \in \Pi$ is *optimal* if $J_{\pi^*} = J^*$. For any stationary policy $\pi = (\mu, \mu, \cdots) \in \Pi$ we write $J_\pi = J_\mu$. Thus a stationary policy $\pi^* = \{\mu^*, \mu^*, \cdots\}$ is optimal if $J^* = J_{\mu^*}$.

For every result to be shown *one of the following three assumptions will be in effect.*

Assumption C (Contraction assumption). The functions $\bar{J}$, $T(J)$, and $T_\mu(J)$ belong to B for all $\mu \in M$ and $J \in B$, and for every $\pi = \{\mu_0, \mu_1, \cdots\} \in \Pi$ the limit

$$\lim_{N \to \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}})(\bar{J})(x)$$

exists and is a real number for each $x \in S$. Furthermore there exist a positive integer m, and scalars ρ, α with $0 < \rho < 1$, $0 < \alpha$ such that for all $J, J' \in B$ there holds

$$\|T_\mu(J) - T_\mu(J')\| \leqq \alpha \|J - J'\| \quad \forall \mu \in M, \tag{20}$$

$$\begin{aligned} \|(T_{\mu_0}, T_{\mu_1} \cdots T_{\mu_{m-1}})(J) - (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{m-1}})(J')\| \\ \leqq \rho \|J - J'\| \quad \forall \mu_0, \cdots, \mu_{m-1} \in M. \end{aligned} \tag{21}$$

Assumption I (Uniform increase assumption). There holds

$$\bar{J}(x) \leqq H(x, u, \bar{J}) \quad \forall x \in S, \quad u \in U(x). \tag{22}$$

Assumption D (Uniform decrease assumption). There holds

$$\bar{J}(x) \geqq H(x, u, \bar{J}) \quad \forall x \in S, \quad u \in U(x). \tag{23}$$

It is easy to see that under each of these assumptions the limit in (17) is well defined as a real number or $\pm\infty$. Indeed in the case of Assumption I we have using (22)

$$\bar{J} \leqq T_{\mu_0}(\bar{J}) \leqq (T_{\mu_0}T_{\mu_1})(\bar{J}) \leqq \cdots \leqq (T_{\mu_0}T_{\mu_1}\cdots T_{\mu_{N-1}})(\bar{J}) \leqq \cdots \tag{24}$$

while in the case of Assumption D we have using (23)

$$\bar{J} \geqq T_{\mu_0}(\bar{J}) \geqq (T_{\mu_0}T_{\mu_1})(\bar{J}) \geqq \cdots \geqq (T_{\mu_0}T_{\mu_1}\cdots T_{\mu_{N-1}})(\bar{J}) \geqq \cdots \tag{25}$$

In both cases the limit in (17) clearly exists for each $x \in S$.

A large number of sequential optimization problems which are of interest in practice may be viewed as special cases of the abstract problem formulated above. We provide below some examples. Several other examples can be found in the author's textbook [1], and in the paper by Denardo [4] who considered a somewhat different problem under assumptions similar to Assumption C.

1. *Deterministic optimal control problems with additive cost functional.*

$$\text{minimize} \lim_{N\to\infty} \sum_{k=0}^{N-1} \alpha^k g[x_k, \mu_k(x_k)] \tag{26}$$

subject to

$$x_{k+1} = f[x_k, \mu_k(x_k)], \quad \mu_k \in M, \quad k = 0, 1, \cdots.$$

If we define

$$H(x, u, J) = g(x, u) + \alpha J[f(x, u)], \tag{27}$$

then problem (26) is equivalent to our abstract problem (18) for $\bar{J}(x) = 0$, $\forall x \in S$. Assumption C is satisfied if $0 < \alpha < 1$ and g is uniformly bounded, i.e., there exists a scalar $b > 0$ such that

$$|g(x, u)| \leqq b \quad \forall x \in S, \quad u \in U(x). \tag{28}$$

This case corresponds to a discounted problem and is examined in [1, §§ 6.1–6.3]. Assumption I is satisfied if $0 < \alpha$ and

$$g(x, u) \geqq 0 \quad \forall x \in S, \quad u \in U(x) \tag{29}$$

while Assumption D is satisfied if $0 < \alpha$ and

$$g(x, u) \leqq 0 \quad \forall x \in S, \quad u \in U(x). \tag{30}$$

These cases are covered in [1, §§ 6.4, 7.1]. If g is extended real valued some care must be exercised in the definition of the mapping H in (27) so that the forbidden sum $(+\infty, -\infty)$ does not arise. This can be done by defining under Assumption I [c.f. (29)] $H(x, u, J) \equiv -\infty$ if $J(x) = -\infty$ for some $x \in S$, and by defining under Assumption D [c.f. (30)] $H(x, u, J) = +\infty$ if $J(x) = +\infty$ for some $x \in S$. We mention that state constraints of the form $x_k \in X$, $\forall k = 0, 1, \cdots$, can be incorporated under I in the cost functional by defining $g(x, u) = +\infty$ whenever $x \notin X$. Note that the deterministic versions of Blackwell's positive D.P. model [3] and Strauch's negative D.P. model [11] are covered under Assumption D and Assumption I respectively.

Deterministic optimal control problems with nonstationary cost per stage and system equation (including finite horizon problems) may be reformulated into the form of problem (26) (see [1, § 6.7]). A generalization of problem (26) is obtained if the scalar α is replaced by a function $\alpha(x, u)$ in (27) and the discount factor depends on the state x and the control u. Then Assumption C is satisfied if the assumption $0<\alpha<1$ is replaced by

$$0 \leqq \inf\{\alpha(x, u) | x \in S, u \in U(x)\} \leqq \sup\{\alpha(x, u) | x \in S, u \in U(x)\} < 1$$

and (28) holds. If $0 \leqq \alpha(x, u)$ for all $x \in S$, $u \in U(x)$, then Assumption I or D is satisfied if (29) or (30) holds respectively.

2. *Stochastic optimal control with additive cost functional.* This problem is obtained from problem (18) when $\bar{J} \equiv 0$ and

$$H(x, u, J) = \underset{w}{E}\{g(x, u, w) + \alpha J[f(x, u, w)] | x, u\},$$

where w is an uncertain parameter element of a *countable* set W with given probability distribution depending on x and u. Such problems are examined in [1, Chaps. 6 and 7] and include a large variety of Markovian decision problems with countable state space. Assumption C holds if $0<\alpha<1$ and $|g(x, u, w)| \leqq b$ for some $b>0$ and all $x \in S$, $u \in U(x)$, $w \in W$. Assumptions I and D hold if $a>0$ and $g(x, u, w) \geqq 0$ or $g(x, u, w) \leqq 0$ respectively for all x, u, w. A generalized version is obtained when α is replaced by a function $\alpha(x, u, w)$ satisfying similar assumptions as the corresponding functions in the previous example. This case covers certain discounted semi-Markov decision problems.

When the set W is not countable then matters are complicated by the need to impose a measurable space structure on S, C, and W and to require that the functions $\mu \in M$ be measurable (in the works of Blackwell, Strauch, and Hinderer [3], [11], [6], S, C, and W are Borel subsets of complete separable metric spaces and μ is required to be Borel measurable). Because of these restrictions the reformulation of the stochastic control problem into the form of the abstract problem (18) is not straightforward. Recent work of S. Shreve and the author [10] has demonstrated however that the framework of this paper is applicable in its entirety as well as convenient once the stochastic control problem is converted to a deterministic control problem (such as the one of the previous example) for which the state space is the set of all probability measures on S. For a detailed treatment we refer to the thesis of Shreve [12].

3. *Minimax control problem with additive cost functional.* This problem is obtained from problem (18) when $\bar{J} \equiv 0$ and

$$H(x, u, J) = \sup_{w \in W(x,u)} \{g(x, u, w) + \alpha J[f(x, u, w)]\}.$$

Here again w is an uncertain parameter belonging to a set W, and $W(x, u)$ is a given subset of W for each $x \in S$, $u \in U(x)$. Under assumptions analogous to those of the previous two examples, Assumptions C, I, or D can be shown to hold. The problem of reachability over an infinite horizon examined by the author in [2] can be shown to be a special case of this problem.

4. *Stochastic optimal control problems with exponential cost functional.* Under similar assumptions for w as in Example 2 consider

$$H(x, u, J) = \underset{w}{E}\{J[f(x, u, w)]\, e^{g(x,u,w)} | x, u\}.$$

This problem corresponds to minimization of the exponential cost functional

$$J_\pi(x) = \lim_{N\to\infty} \underset{w_k}{E}\left\{\exp\left(\sum_{k=0}^{N-1} g[x_k, \mu_k(x_k), w_k]\right)\right\}$$

subject to the system equation $x_{k+1} = f[x_k, \mu_k(x_k), w_k]$. An example of a finite horizon version of this problem has been considered in [7]. Here we take $\bar{J}(x) = 1$, $\forall x \in S$. If $g(x, u, w) \geqq 0$ for all (x, u, w) then Assumption I is satisfied, while if $g(x, u, w) \leqq 0$ for all (x, u, w) then Assumption D is satisfied.

4. Results under Assumption C. As mentioned earlier, a variation of our problem under Assumption C has been analyzed by Denardo. We shall restrict ourselves to providing some results which yield the connection between Denardo's framework and the one considered here.

PROPOSITION 1. *Let Assumption* C *hold. Then*:

(a) *For every* $J \in B$, $\pi \in \Pi$ *and* $x \in S$ *there holds*

$$J_\pi(x) = \lim_{N\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}})(\bar{J})(x) = \lim_{N\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_{N-1}})(J)(x).$$

(b) *The function* J^* *belongs to* B *and is the unique fixed point of* T *within* B, *i.e.*, $J^* = T(J^*)$ *and if* $J' \in B$, $J' = T(J')$, *then* $J' = J^*$. *Furthermore if* $J' \in B$ *is such that* $T(J') \leqq J'$ *then* $J^* \leqq J'$ *while if* $J' \leqq T(J')$ *then* $J' \leqq J^*$.

(c) *For every* $\mu \in M$ *the function* J_μ *belongs to* B *and is the unique fixed point of* T_μ *within* B.

(d) *There holds*

$$\lim_{N\to\infty} \|T^N(J) - J^*\| = 0 \quad \forall J \in B,$$

$$\lim_{N\to\infty} \|T_\mu^N(J) - J_\mu\| = 0 \quad \forall J \in B, \quad \mu \in M.$$

(e) *A stationary policy* $\pi^* = \{\mu^*, \mu^*, \cdots\} \in \Pi$ *is optimal if and only if*

$$T_{\mu^*}(J^*) = T(J^*).$$

Equivalently π^* *is optimal if and only if*

$$T_{\mu^*}(J_{\mu^*}) = T(J_{\mu^*}).$$

(f) *If there exists an optimal policy, there exists an optimal stationary policy.*

(g) *For any* $\varepsilon > 0$ *there exists a stationary policy* $\pi_\varepsilon = \{\mu_\varepsilon, \mu_\varepsilon, \cdots\}$ *such that*

$$\|J^* - J_{\mu_\varepsilon}\| \leqq \varepsilon.$$

Proof. Since the proof uses similar arguments as those in [4] (see also [1, Chap. 6, Prob. 4]) it will be abbreviated.

(a) For any integer $k \geqq 0$ write $k = nm + q$ where q, n are nonnegative integers and $0 \leqq q < m$. Then for any $J, J' \in B$ using (20), (21) we obtain

$$\|(T_{\mu_0} \cdots T_{\mu_{k-1}})(J) - (T_{\mu_0} \cdots T_{\mu_{k-1}})(J')\| \leqq \rho^n \alpha^q \|J - J'\|$$

from which

$$\lim_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_{k-1}})(\bar{J}) = \lim_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_{k-1}})(J) \quad \forall J \in B.$$

(b), (c), (d) Relation (20) can be used to show (compare with the proof of Proposition 3) that

$$T^N(\bar{J})(x) = \inf_{\pi\in\Pi} (T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J})(x) \quad \forall x \in S, \quad N = 1, 2, \cdots,$$

and it follows from (21) that T and T_μ, $\mu \in M$ are m-stage contraction mappings, i.e., $\|T^m(J) - T^m(J')\| \leqq \bar{\rho}\|J - J'\|$ and $\|T_\mu^m(J) - T_\mu^m(J')\| \leqq \bar{\rho}_\mu \|J - J'\|$ for some $\bar{\rho} \in (0, 1)$, $\bar{\rho}_\mu \in (0, 1)$ and all $J, J' \in B$. Hence T and T_μ have unique fixed points in B. The fixed point of T_μ is clearly J_μ and hence part (c) is proved. Let $\tilde{J}^*$ be the unique fixed point of T. We have $\tilde{J}^* = T(\tilde{J}^*)$. For any $\bar{\varepsilon} > 0$ take $\bar{\mu} \in M$ such that

$$T_{\bar{\mu}}(\tilde{J}^*) \leqq \tilde{J}^* + \bar{\varepsilon} e.$$

Using (20) it follows that $T_{\bar{\mu}}^2(\tilde{J}^*) \leqq T_{\bar{\mu}}(\tilde{J}^*) + \alpha\bar{\varepsilon} e \leqq \tilde{J}^* + (1+\alpha)\bar{\varepsilon} e$. Continuing in the same manner we obtain

$$T_{\bar{\mu}}^m(\tilde{J}^*) \leqq \tilde{J}^* + (1 + \alpha + \cdots + \alpha^{m-1})\bar{\varepsilon} e.$$

Using (21) we obtain

$$\begin{aligned} T_{\bar{\mu}}^{2m}(\tilde{J}^*) &\leqq T_{\bar{\mu}}^m(\tilde{J}^*) + \rho(1 + \alpha + \cdots + \alpha^{m-1})\bar{\varepsilon} e \\ &\leqq \tilde{J}^* + (1+\rho)(1 + \alpha + \cdots + \alpha^{m-1})\bar{\varepsilon} e. \end{aligned}$$

Proceeding similarly we obtain for all $k \geqq 1$,

$$T_{\bar{\mu}}^{km}(\tilde{J}^*) \leqq \tilde{J}^* + (1 + \rho + \cdots + \rho^{k-1})(1 + \alpha + \cdots + \alpha^{m-1})\bar{\varepsilon} e.$$

Taking the limit as $k \to \infty$ and using the fact $J_{\bar{\mu}} = \lim_{k\to\infty} T_{\bar{\mu}}^{km}(\tilde{J}^*)$ we obtain

$$J_{\bar{\mu}} \leqq \tilde{J}^* + \frac{1}{1-\rho}(1 + \alpha + \cdots + \alpha^{m-1})\bar{\varepsilon} e.$$

Taking $\bar{\varepsilon} = (1-\rho)(1 + \alpha + \cdots + \alpha^{m-1})^{-1}\varepsilon$ we obtain

$$J_{\bar{\mu}} \leqq \tilde{J}^* + \varepsilon e. \tag{31}$$

Since $J^* \leqq J_{\bar{\mu}}$ and $\varepsilon > 0$ is arbitrary we obtain $J^* \leqq \tilde{J}^*$. We also have

$$J^* = \inf_{\pi\in\Pi} \lim_{N\to\infty} (T_{\mu_0} \cdots T_{\mu_{N-1}})(\tilde{J}^*) \geqq \lim_{N\to\infty} T^N(\tilde{J}^*) = \tilde{J}^*.$$

Hence $J^* = \tilde{J}^*$ and J^* is the unique fixed point of T. Thus part (b) is proved. Part (d) follows immediately from the contraction property of T and T_μ.

(e) If π^* is optimal then $J_{\mu^*} = J^*$ and the result follows from part (b) and (c). If $T_{\mu^*}(J^*) = T(J^*)$ then $T_{\mu^*}(J^*) = J^*$ and hence $J_{\mu^*} = J^*$ by part (c). If $T_{\mu^*}(J_{\mu^*}) = T(J_{\mu^*})$ then $J_{\mu^*} = T(J_{\mu^*})$ and $J_{\mu^*} = J^*$ by part (b).

(f) Let $\pi^* = \{\mu_0^*, \mu_1^*, \cdots\}$ be an optimal policy. Then using parts (a) and (b)

$$J^* = J_{\pi^*} = \lim_{k\to\infty} (T_{\mu_0^*} \cdots T_{\mu_k^*})(\bar{J})$$

$$= \lim_{k\to\infty} (T_{\mu_0^*} \cdots T_{\mu_k^*})(J^*) \geqq \lim_{k\to\infty} (T_{\mu_0^*} T^k)(J^*) = T_{\mu_0^*}(J^*) \geqq T(J^*) = J^*.$$

Hence $T_{\mu_0^*}(J^*) = T(J^*)$ and by part (e) the stationary policy $(\mu_0^*, \mu_0^*, \cdots\}$ is optimal.

(g) This part was proved earlier in the proof of part (b), [cf. (31)]. Q.E.D.

For additional results and computational methods the reader is referred to Denardo's paper [4] and the author's textbook [1, Chap. 6]. Notice that part (a) shows that $\bar{J}$ may be replaced by any function $J \in B$. Thus it is often possible to select $\bar{J}$ in a way that Assumption I or D is satisfied and obtain alternative proofs of parts of Proposition 1 by using the results of the next section.

5. Results under Assumptions I or D. In our analysis under Assumptions I or D we will occasionally need to assume one or more of the following continuity properties for the mapping H. Assumptions I.1 and I.2 will be used in conjunction with Assumption I, while Assumptions D.1 and D.2 will be used in conjunction with Assumption D.

Assumption I.1. If $\{\bar{J}_k\} \subset F$ is a sequence satisfying $\bar{J} \leqq \bar{J}_k \leqq \bar{J}_{k+1}$ for all k, then

$$\lim_{k\to\infty} H(x, u, \bar{J}_k) = H(x, u, \lim_{k\to\infty} \bar{J}_k) \quad \forall x \in S, \quad u \in U(x). \tag{32}$$

Assumption I.2. There exists a scalar $\alpha > 0$ such that for all scalars $r > 0$ and functions $J \in F$ with $\bar{J} \leqq J$ there holds

$$H(x, u, J) \leqq H(x, u, J + re) \leqq H(x, u, J) + \alpha r \quad \forall x \in S, \quad u \in U(x), \tag{33}$$

where e denotes the unit function $[e(x) = 1, \forall x \in S]$.

Assumption D.1. If $\{\bar{J}_k\} \subset F$ is a sequence satisfying $\bar{J}_{k+1} \leqq \bar{J}_k \leqq \bar{J}$ for all k, then

$$\lim_{k\to\infty} H(x, u, \bar{J}_k) = H(x, u, \lim_{k\to\infty} \bar{J}_k) \quad \forall x \in S, \quad u \in U(x). \tag{34}$$

Assumption D.2. There exists a scalar $\alpha > 0$ such that for all scalars $r > 0$ and functions $J \in F$ with $J \leqq \bar{J}$ there holds

$$H(x, u, J) - \alpha r \leqq H(x, u, J - re) \leqq H(x, u, J) \quad \forall x \in S, \quad u \in U(x), \tag{35}$$

where e denotes the unit function $[e(x) = 1, \forall x \in S]$.

Notice that both the deterministic and the stochastic optimal control problems of § 3 satisfy I.1, I.2, D.1, D.2. The minimax control problem of § 3 satisfies I.1, I.2, D.2 while additional assumptions are needed in order that D.1 is satisfied as well. The mapping of Example 4 in § 3 satisfies I.1, D.1, and D.2 while if it is assumed that $|g(x, u, w)| \leqq b$ for some scalar b and all (x, u, w), then I.2 holds as well.

Dynamic programming and the finite horizon version of the problem. It is both interesting and helpful in the analysis that follows to consider the finite horizon version of our problem which involves finding for any positive integer N

$$J_N(x) = \inf_{\pi \in \Pi} (T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J})(x) \tag{36}$$

as well as a policy attaining the infimum above (if one exists). We refer to this problem as the *N-stage problem.* We have the following results:

PROPOSITION 2. *Let* I *and* I.2 *hold. Then* $J_N = T^N(\bar{J})$ *for all* $N = 1, 2, \cdots$.

Proof. For any $\varepsilon > 0$ let $\bar{\mu}_k \in M$, $k = 0, 1, \cdots, N-1$, be such that

$$T_{\bar{\mu}_k}[T^{N-k-1}(\bar{J})] \leqq T^{N-k}(\bar{J}) + \varepsilon e, \qquad k = 0, 1, \cdots, N-1.$$

Such functions exist since $\bar{J}(x) > -\infty$ for all $x \in S$ and $T^{N-k}(\bar{J}) \geqq \bar{J}$ by I. We have using I.2,

$$\begin{aligned}
J_N = \inf_{\pi \in \Pi} (T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J}) &\leqq (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{N-1}})(\bar{J}) \\
&\leqq (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{N-2}})[T(\bar{J}) + \varepsilon e] \\
&\leqq (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{N-3}})[(T_{\bar{\mu}_{N-2}} T)(\bar{J}) + \alpha \varepsilon e] \\
&\quad \cdots \cdot \\
&\leqq (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{N-2}} T)(\bar{J}) + \alpha^{N-1} \varepsilon e \\
&\quad \cdots \cdot \\
&\leqq (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{N-3}} T^2)(\bar{J}) + (\alpha^{N-2} + \alpha^{N-1}) \varepsilon e \\
&\quad \cdots \cdot \\
&\leqq T^N(\bar{J}) + \left(\sum_{k=0}^{N-1} \alpha^k \right) \varepsilon e.
\end{aligned}$$

Since ε is arbitrary we obtain $J_N \leqq T^N(\bar{J})$. On the other hand we have, by the definition of T and J_N, $T^N(\bar{J}) \leqq J_N$. Hence $J_N = T^N(\bar{J})$. Q.E.D.

Proposition 2 may fail to hold in the absence of I.2 even if I.1 holds as the following counterexample shows.

Counterexample 1. Take $S = \{0\}$, $C = U(0) = (0, 1]$, $\bar{J}(0) = 0$, $H(0, u, J) = 1$ if $J(0) > 0$, $H(0, u, J) = u$ if $J(0) \leqq 0$. Then $(T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J})(0) = 1$ for every $\pi \in \Pi$ and $N \geqq 2$ and hence $J_N(0) = 1$ for $N \geqq 2$. On the other hand we have $T^N(\bar{J})(0) = 0$ for all N. Here I and I.1 are satisfied but I.2 is violated.

PROPOSITION 3. *Let* D *hold. Assume that either* D.1 *holds or else* D.2 *holds and* $T^N(\bar{J})(x) > -\infty$ *for all* $x \in S$. *Then* $J_N = T^N(\bar{J})$.

Proof. Let D.1 hold. For each $k = 0, 1, \cdots, N-1$ consider a sequence $\{\mu_k^i\} \subset M$ such that

$$\lim_{i \to \infty} T_{\mu_k^i}[T^{N-k-1}(\bar{J})] = T^{N-k}(\bar{J}), \qquad k = 0, \cdots, N-1.$$

$$T_{\mu_k^i}[T^{N-k-1}(\bar{J})] \geqq T_{\mu_k^{i+1}}[T^{N-k-1}(\bar{J})],$$

$$k = 0, \cdots, N-1, \quad i = 0, 1, \cdots.$$

We have by using D.1,

$$
\begin{aligned}
J_N &\leqq \lim_{\substack{i_0\to\infty \\ \cdots \\ i_{N-1}\to\infty}} (T_{\mu_0^{i_0}} \cdots T_{\mu_{N-1}^{i_{N-1}}})(\bar{J}) \\
&= \lim_{\substack{i_0\to\infty \\ \cdots \\ i_{N-2}\to\infty}} (T_{\mu_0^{i_0}} \cdots T_{\mu_{N-2}^{i_{N-2}}})[\lim_{i_{N-1}\to\infty} T_{\mu_{N-1}^{i_{N-1}}}(\bar{J})] \\
&= \lim_{\substack{i_0\to\infty \\ \cdots \\ i_{N-2}\to\infty}} (T_{\mu_0^{i_0}} \cdots T_{\mu_{N-2}^{i_{N-1}}})[T(\bar{J})] \\
&\quad \cdots\cdots \\
&= T^N(\bar{J}).
\end{aligned}
$$

On the other hand we have clearly $T^N(\bar{J}) \leqq J_N$ and hence $J_N = T^N(\bar{J})$.

Let D.2 hold and assume $T^N(\bar{J})(x) > -\infty \; \forall x \in S$. For any $\varepsilon > 0$ let $\bar{\mu}_k \in M$, $k = 0, 1, \cdots, N-1$, be such that

$$
\begin{aligned}
&T_{\bar{\mu}_{N-1}}(\bar{J}) \leqq T(\bar{J}) + \varepsilon e, \\
&(T_{\bar{\mu}_{N-2}} T_{\bar{\mu}_{N-1}})(\bar{J}) \leqq T[T_{\bar{\mu}_{N-1}}(\bar{J})] + \varepsilon e, \\
&\cdots\cdots \\
&(T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{N-1}})(\bar{J}) \leqq T[(T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_{N-1}})(\bar{J})] + \varepsilon e.
\end{aligned}
$$

The assumption $T^N(\bar{J})(x) > -\infty$, $\forall x \in S$ guarantees that such functions, $\bar{\mu}_k$ exist. We have using D.2,

$$
\begin{aligned}
T^N(\bar{J}) &\geqq T^{N-1}[T_{\bar{\mu}_{N-1}}(\bar{J}) - \varepsilon e] \geqq (T^{N-1} T_{\bar{\mu}_{N-1}})(\bar{J}) - \alpha^{N-1} \varepsilon e \\
&\geqq T^{N-2}[(T_{\bar{\mu}_{N-2}} T_{\bar{\mu}_{N-1}})(\bar{J}) - \varepsilon e] - \alpha^{N-1} \varepsilon e \\
&\geqq (T^{N-2} T_{\bar{\mu}_{N-2}} T_{\bar{\mu}_{N-1}})(\bar{J}) - (\alpha^{N-2} + \alpha^{N-1}) \varepsilon e \\
&\cdots\cdots \\
&\geqq (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{N-1}})(\bar{J}) - \left(\sum_{k=0}^{N-1} \alpha^k\right) \varepsilon e \\
&\geqq J_N - \left(\sum_{k=0}^{N-1} \alpha^k\right) \varepsilon e.
\end{aligned}
$$

Since ε is arbitrary it follows that $T^N(\bar{J}) \geqq J_N$. On the other hand we have clearly $T^N(\bar{J}) \leqq J_N$ and hence $J_N = T^N(\bar{J})$. Q.E.D.

Proposition 3 may fail to hold if its assumptions are slightly relaxed.

Counterexample 2. Take $S = \{0\}$, $C = U(0) = (-1, 0]$, $\bar{J}(0) = 0$, $H(0, u, J) = u$ if $-1 < J(0)$, $H(0, u, J) = J(0) + u$ if $J(0) \leqq -1$. Then $(T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J})(0) = \mu_0(0)$ and $J_N(0) = -1$, while $T^N(\bar{J})(0) = -N$ for every N. Here D and the assumption $T^N(\bar{J})(0) > -\infty$ are satisfied, but D.1 and D.2 are both violated.

Counterexample 3. Take $S = \{0, 1\}$, $C = U(0) = U(1) = (-\infty, 0]$, $\bar{J}(0) = \bar{J}(1) = 0$, $H(0, u, J) = u$ if $J(1) = -\infty$, $H(0, u, J) = 0$ if $J(1) > -\infty$, and $H(1, u, J) =$

u. Then $(T_{\mu_0}\cdots T_{\mu_{N-1}})(\bar{J})(0)=0$, $(T_{\mu_0}\cdots T_{\mu_{N-1}})(\bar{J})(1)=\mu_0(1)$ for all $N\geqq 1$. Hence $J_N(0)=0$, $J_N(1)=-\infty$. On the other hand we have $T^N(\bar{J})(0)=T^N(\bar{J})(1)=-\infty$ for all $N\geqq 2$. Here D and D.2 are satisfied, but D.1 and the assumption $T^N(\bar{J})(x)>-\infty$, $\forall x\in S$ are both violated.

Characterization of the optimal value function. We now consider the question whether Bellman's equation, [i.e. $J^*=T(J^*)$] holds within our generalized setting. We first prove a preliminary result which is of independent interest.

PROPOSITION 4. *Let* I, I.1, *and* I.2 *hold. Then given any* $\varepsilon>0$ *there exists a policy* $\pi_\varepsilon\in\Pi$ *such that*

$$J^*\leqq J_{\pi_\varepsilon}\leqq J^*+\varepsilon e. \tag{37}$$

Furthermore if the scalar α *in* I.2 *satisfies* $\alpha<1$ *the policy* π_ε *can be taken stationary.*

Proof. Let $\{\varepsilon_k\}$ be a sequence such that $\varepsilon_k>0$ for all k, and

$$\sum_{k=0}^{\infty}\alpha^k\varepsilon_k=\varepsilon. \tag{38}$$

For each $x\in S$ consider a sequence of policies $\{\pi_k[x]\}\subset\Pi$ of the form

$$\pi_k[x]=\{\mu_0^k[x],\ \mu_1^k[x],\cdots\}$$

such that for $k=0,1,\cdots$,

$$J_{\pi_k[x]}(x)\leqq J^*(x)+\varepsilon_k\quad\forall x\in S. \tag{39}$$

Such a sequence exists since we have $J^*(x)>-\infty$ under our assumptions.

The (admittedly confusing) notation used above and later in the proof should be interpreted as follows. The policy $\pi_k[x]=\{\mu_0^k[x],\ \mu_1^k[x],\cdots\}$ is associated with x. Thus $\mu_i^k[x]$ denotes, for each $x\in S$ and k, a function in M, while $\mu_i^k[x](z)$ denotes the value of $\mu_i^k[x]$ at an element $z\in S$. In particular $\mu_i^k[x](x)$ denotes the value of $\mu_i^k[x]$ at x.

Consider the functions $\bar{\mu}_k\in M$ defined by

$$\bar{\mu}_k(x)=\mu_0^k[x](x)\quad\forall x\in S, \tag{40}$$

and the functions $\bar{J}_k$ defined by

$$\bar{J}_k(x)=H[x,\bar{\mu}_k(x),\lim_{i\to\infty}(T_{\mu_0^k[x]}\cdots T_{\mu_i^k[x]})(\bar{J})]\quad\forall x\in S,\quad k=0,1,\cdots. \tag{41}$$

By using (39), (40), I, and I.1 we obtain

$$\bar{J}_k(x)=\lim_{i\to\infty}(T_{\mu_0^k[x]}\cdots T_{\mu_i^k[x]})(\bar{J})(x)=J_{\pi_k[x]}(x)\leqq J^*(x)+\varepsilon_k \tag{42}$$

$$\forall x\in S,\quad k=0,1,\cdots$$

We have using (41), (42), and I.2 for all $k=1,2,\cdots$ and $x\in S$,

$$\begin{aligned}T_{\bar{\mu}_{k-1}}(\bar{J}_k)(x)&=H[x,\bar{\mu}_{k-1}(x),\bar{J}_k]\\&\leqq H[x,\bar{\mu}_{k-1}(x),(J^*+\varepsilon_k e)]\leqq H[x,\bar{\mu}_{k-1}(x),J^*]+\alpha\varepsilon_k\\&\leqq H[x,\bar{\mu}_{k-1}(x),\lim_{i\to\infty}(T_{\mu_1^{k-1}[x]}\cdots T_{\mu_i^{k-1}[x]})(\bar{J})]+\alpha\varepsilon_k\\&=\bar{J}_{k-1}(x)+\alpha\varepsilon_k,\end{aligned}$$

and finally

$$T_{\bar{\mu}_{k-1}}(\bar{J}_k) \leqq \bar{J}_{k-1} + \alpha\varepsilon_k e \quad \forall k = 1, 2, \cdots. \tag{43}$$

Using this inequality and I.2 we obtain

$$\begin{aligned} T_{\bar{\mu}_{k-2}}[T_{\bar{\mu}_{k-1}}(\bar{J}_k)] &\leqq T_{\bar{\mu}_{k-2}}(\bar{J}_{k-1} + \alpha\varepsilon_k e) \\ &\leqq T_{\bar{\mu}_{k-2}}(\bar{J}_{k-1}) + \alpha^2\varepsilon_k e \leqq \bar{J}_{k-2} + (\alpha\varepsilon_{k-1} + \alpha^2\varepsilon_k)e. \end{aligned}$$

Continuing in the same manner we obtain for $k = 1, 2, \cdots$,

$$(T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}})(\bar{J}_k) \leqq \bar{J}_0 + (\alpha\varepsilon_1 + \cdots + \alpha^k\varepsilon_k)e \leqq J^* + \left(\sum_{i=0}^{k} \alpha^i\varepsilon_i\right)e.$$

Since $\bar{J} \leqq \bar{J}_k$ it follows that

$$(T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}})(\bar{J}) \leqq J^* + \left(\sum_{i=0}^{k} \alpha^i\varepsilon_i\right)e.$$

Denote $\pi_\varepsilon = \{\bar{\mu}_0, \bar{\mu}_1, \cdots\}$. Then by taking limit above and using (38) we obtain $J_{\pi_\varepsilon} \leqq J^* + \varepsilon e$. If $\alpha < 1$ take $\varepsilon_k = \varepsilon(1-a)$ and $\pi_k[x] = \{\mu_0[x], \mu_1[x], \cdots\}$ for all k. Then the policy $\pi_\varepsilon = \{\bar{\mu}_0, \bar{\mu}_1, \cdots\}$ is stationary. Q.E.D.

By using Proposition 4 we can prove the following.

PROPOSITION 5. *Let* I, I.1, *and* I.2 *hold. Then*

$$J^* = T(J^*).$$

Furthermore if $J' \in F$ *is such that* $J' \geqq \bar{J}$ *and* $J' \geqq T(J')$, *then* $J' \geqq J^*$.

Proof. For every $\pi = \{\mu_0, \mu_1, \cdots\} \in \Pi$ and $x \in S$ we have using I.1

$$\begin{aligned} J_\pi(x) &= \lim_{k\to\infty} (T_{\mu_0}T_{\mu_1} \cdots T_{\mu_k})(\bar{J})(x) \\ &= T_{\mu_0}[\lim_{k\to\infty} (T_{\mu_1} \cdots T_{\mu_k})(\bar{J})](x) \\ &\geqq T_{\mu_0}(J^*)(x) \geqq T(J^*)(x). \end{aligned}$$

By taking the infimum of the left hand side over $\pi \in \Pi$

$$J^* \geqq T(J^*).$$

To prove the reverse inequality let $\varepsilon_1, \varepsilon_2$ be any positive scalars and let $\bar{\pi} = \{\bar{\mu}_0, \bar{\mu}_1, \cdots\}$ be such that

$$\begin{aligned} T_{\bar{\mu}_0}(J^*) &\leqq T(J^*) + \varepsilon_1 e, \\ J_{\bar{\pi}_1} &\leqq J^* + \varepsilon_2 e, \end{aligned}$$

where $\bar{\pi}_1 = \{\bar{\mu}_1, \bar{\mu}_2, \cdots\}$. Such a policy exists by Proposition 4. We have

$$\begin{aligned} J_{\bar{\pi}} &= \lim_{k\to\infty} (T_{\bar{\mu}_0}T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_k})(\bar{J}) \\ &= T_{\bar{\mu}_0}[\lim_{k\to\infty} (T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_k})(\bar{J})] = T_{\bar{\mu}_0}(J_{\bar{\pi}_1}) \\ &\leqq T_{\bar{\mu}_0}(J^*) + \alpha\varepsilon_2 e \leqq T(J^*) + (\varepsilon_1 + \alpha\varepsilon_2)e. \end{aligned}$$

Since $J^* \leqq J_{\bar{\pi}}$ and $\varepsilon_1, \varepsilon_2$ can be taken arbitrarily small it follows that

$$J^* \leqq T(J^*).$$

Hence $J^* = T(J^*)$.

Assume that $J' \in F$ satisfies $J' \geqq \bar{J}$ and $J' \geqq T(J')$. Let $\{\varepsilon_k\}$ be any sequence with $\varepsilon_k > 0$ and consider a policy $\bar{\pi} = \{\bar{\mu}_0, \bar{\mu}_1, \cdots\} \in \Pi$ such that

$$T_{\bar{\mu}_k}(J') \leqq T(J') + \varepsilon_k e, \qquad k = 0, 1, \cdots.$$

We have using I.2,

$$\begin{aligned}
J^* &= \inf_{\pi \in \Pi} \lim_{k \to \infty} (T_{\mu_0} \cdots T_{\mu_k})(\bar{J}) \\
&\leqq \inf_{\pi \in \Pi} \liminf_{k \to \infty} (T_{\mu_0} \cdots T_{\mu_k})(J') \\
&\leqq \liminf_{k \to \infty} (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_k})(J') \\
&\leqq \liminf_{k \to \infty} (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}})[T(J') + \varepsilon_k e] \\
&\leqq \liminf_{k \to \infty} (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-2}})[T_{\bar{\mu}_{k-1}}(J' + \varepsilon_k e)] \\
&\leqq \liminf_{k \to \infty} (T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-2}})[T_{\bar{\mu}_{k-1}}(J') + \alpha \varepsilon_k e] \\
&\leqq \liminf_{k \to \infty} [(T_{\bar{\mu}_0} \cdots T_{\bar{\mu}_{k-1}})(J') + \alpha^k \varepsilon_k e] \\
&\cdots \\
&\leqq \lim_{k \to \infty} \left[T(J') + \left(\sum_{i=0}^{k} \alpha^i \varepsilon_i \right) e \right] \leqq J' + \left(\sum_{i=0}^{\infty} \alpha^i \varepsilon_i \right) e.
\end{aligned}$$

Since we may choose $\sum_{i=0}^{\infty} \alpha^i \varepsilon_i$ as small as desired it follows that $J^* \leqq J'$. Q.E.D.

The following counterexamples show that I.1 and I.2 are essential in order for Proposition 5 to hold.

Counterexample 4. Take $S = \{0, 1\}$, $C = U(0) = U(1) = (-1, 0]$, $\bar{J}(0) = \bar{J}(1) = -1$, $H(0, u, J) = u$ if $J(1) \leqq -1$, $H(0, u, J) = 0$ if $J(1) > -1$, and $H(1, u, J) = u$. Then $(T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J})(0) = 0$ and $(T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J})(1) = \mu_0(1)$ for $N \geqq 1$. Thus $J^*(0) = 0$, $J^*(1) = -1$ while $T(J^*)(0) = -1$, $T(J^*)(1) = -1$ and hence $J^* \neq T(J^*)$. Notice also that $\bar{J}$ is a fixed point of T while $\bar{J} \leqq J^*$ and $\bar{J} \neq J^*$. Here I and I.1 are satisfied but I.2 is violated.

Counterexample 5. Take $S = \{0, 1\}$, $C = U(0) = U(1) = \{0\}$, $\bar{J}(0) = \bar{J}(1) = 0$, $H(0, 0, J) = 0$ if $J(1) < \infty$, $H(0, 0, J) = \infty$ if $J(1) = \infty$, $H(1, 0, J) = J(1) + 1$. Then $(T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J})(0) = 0$ and $(T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J})(1) = N$. Thus $J^*(0) = 0, J^*(1) = \infty$. On the other hand we have $T(J^*)(0) = T(J^*)(1) = \infty$ and $J^* \neq T(J^*)$. Here I and I.2 are satisfied but I.1 is violated.

As a corollary to Proposition 5 we obtain the following:

COROLLARY 5.1. *Let* I, I.1 *and* I.2 *hold. Then for every stationary policy* $\pi = \{\mu, \mu, \cdots\}$ *there holds*

$$J_\mu = T_\mu(J_\mu).$$

Furthermore if $J' \in F$ *is such that* $J' \geqq \bar{J}$, $J' \geqq T_\mu(J')$, *then* $J' \geqq J_\mu$.

Proof. Consider the variation of our problem where the control constraint set is $U_\mu(x) = \{\mu(x)\}\ \forall x \in X$ rather than $U(x)$. Application of Proposition 5 yields the result. Q.E.D.

We now provide the counterpart of Proposition 5 under Assumption D.

PROPOSITION 6. *Let* D *and* D.1 *hold. Then*

$$J^* = T(J^*).$$

Furthermore if $J' \in F$ *is such that* $J' \geqq \bar{J}$, $J' \geqq T(J')$, *then* $J' \geqq J^*$.

Proof. We first show the following lemma:

LEMMA 1. *Let* D *hold. Then*

$$J^* = \lim_{N\to\infty} J_N, \tag{44}$$

where J_N *is the optimal value function of the* N*-stage problem defined by* (36).

Proof. Clearly we have $J^* \leqq J_N$ for all N and hence $J^* \leqq \lim_{N\to\infty} J_N$. Also for all $\pi = \{\mu_0, \mu_1, \cdots) \in \Pi$ we have

$$(T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J}) \geqq J_N,$$

and by taking limit of both sides we obtain $J_\pi \geqq \lim_{N\to\infty} J_N$, and hence $J^* \geqq \lim_{N\to\infty} J_N$. Q.E.D.

We return to the proof of Proposition 6. An entirely similar argument as the one of the proof of Lemma 1 shows that under D we have for all $x \in S$,

$$\lim_{N\to\infty} \inf_{u\in U(x)} H(x, u, J_N) = \inf_{u\in U(x)} \lim_{N\to\infty} H(x, u, J_N). \tag{45}$$

Using D.1 the above equation yields

$$\lim_{N\to\infty} T(J_N) = T(\lim_{N\to\infty} J_N). \tag{46}$$

By Proposition 3 we have $J_N = T^N(\bar{J})$ and hence $T(J_N) = T^{N+1}(\bar{J})$. Combining this relation with (44) and (46) we obtain $J^* = T(J^*)$.

To complete the proof, let $J' \in F$ be such that $J' \leqq \bar{J}$, $J' \leqq T(J')$. Then we have

$$\begin{aligned} J^* &= \inf_{\pi\in\Pi} \lim_{M\to\infty} (T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J}) \\ &\geqq \lim_{N\to\infty} \inf_{\pi\in\Pi} (T_{\mu_0} \cdots T_{\mu_{N-1}})(\bar{J}) \\ &\geqq \lim_{N\to\infty} \inf_{\pi\in\Pi} (T_{\mu_0} \cdots T_{\mu_{N-1}})(J') \\ &\geqq \lim_{N\to\infty} T^N(J') \geqq J'. \end{aligned}$$

Hence $J^* \geqq J'$. Q.E.D.

In Counterexamples 2 and 3, Assumption D.1 does not hold. In both cases we have $J^* \neq T(J^*)$ as the reader can easily verify.

A cursory examination of the proof of Proposition 6 reveals that the only point where we used D.1 was in establishing the relation $\lim_{N\to\infty} T(J_N) = T(\lim_{N\to\infty} J_N)$ [cf. (46)]. Hence if this relation can be established independently then the result of Proposition 6 follows. In this manner we obtain the following corollary.

COROLLARY 6.1. *Let* D *hold and assume that* D.2 *holds, S is a finite set, and* $J^*(x) > -\infty$ *for all* $x \in S$. *Then* $J^* = T(J^*)$. *Furthermore if* $J' \in F$ *is such that* $J' \leqq \bar{J}$, $J' \leqq T(J')$, *then* $J' \leqq J^*$.

Proof. We will show that

$$\lim_{N\to\infty} H(x, u, J_N) = H(x, u, \lim_{N\to\infty} J_N) \quad \forall x \in S, \quad u \in U(x).$$

Then using (45) we obtain (46) and the result follows as in the proof of Proposition 6. Assume the contrary, i.e., that for some $\tilde{x} \in S$, $\tilde{u} \in U(\tilde{x})$, and $\varepsilon > 0$ there holds

$$H(\tilde{x}, \tilde{u}, J_k) - \varepsilon > H(\tilde{x}, \tilde{u}, \lim_{N\to\infty} J_N) \quad \forall k = 1, 2, \cdots.$$

Using the finiteness of S and the fact $J^*(x) = \lim_{N\to\infty} J_N(x) > -\infty$ for all x we obtain that for some positive integer $\bar{k}$ we have

$$J_k - \frac{\varepsilon}{\alpha} e \leqq \lim_{N\to\infty} J_N \quad \forall k \geqq \bar{k}.$$

By using D.2 we obtain for all $k \geqq \bar{k}$,

$$H(\tilde{x}, \tilde{u}, J_k) - \varepsilon \leqq H\left(\tilde{x}, \tilde{u}, J_k - \frac{\varepsilon}{\alpha} e\right) \leqq H(\tilde{x}, \tilde{u}, \lim_{N\to\infty} J_N)$$

which contradicts the earlier inequality. Q.E.D.

Similarly as under I we have the following corollary:

COROLLARY 6.2. *Let* D *and* D.1 *hold. Then for every stationary policy* $\pi = \{\mu, \mu, \cdots\}$ *there holds*

$$J_\mu = T_\mu(J_\mu).$$

Furthermore if $J' \in F$ *is such that* $J' \leqq \bar{J}$, $J' \leqq T_\mu(J')$ *then* $J' \leqq J_\mu$.

It is worth noting that Propositions 5 and 6 may form the basis for computation of J^* when the state space S is a finite set with n elements denoted $x_1, x_2, \cdots, x_n$. It follows from Proposition 5 that, under I, I.1, and I.2, $J^*(x_1), \cdots, J^*(x_n)$ solve the problem

$$\text{minimize} \sum_{i=1}^{n} \lambda_i$$

subject to

$$\lambda_i \geqq \inf_{u\in U(x_i)} H(x_i, u, J_\lambda), \qquad i = 1, \cdots, n,$$

$$\lambda_i \geqq \bar{J}(x_i), \qquad i = 1, \cdots, n,$$

where J_λ is the function taking values $J_\lambda(x_i) = \lambda_i$, $i = 1, \cdots, n$. Under D and D.1,

or D, D.2 and $J^*(x) > -\infty \ \forall x \in S$ the corresponding problem is

$$\text{maximize} \sum_{i=1}^{n} \lambda_i$$

subject to

$$\lambda_i \leqq \inf_{u \in U(x_i)} H(x_i, u, J_\lambda), \qquad i = 1, \cdots, n,$$

$$\lambda_i \leqq \bar{J}(x_i), \qquad i = 1, \cdots, n.$$

When $U(x_i)$ is also a finite set for all i, then the problems above become finite-dimensional nonlinear programming problems.

Characterization of optimal stationary policies. We have the following necessary and sufficient conditions for optimality of a stationary policy.

PROPOSITION 7. *Let* I, I.1, *and* I.2 *hold. Then a stationary policy* $\pi^* = \{\mu^*, \mu^*, \cdots\}$ *is optimal if and only if*

$$T_{\mu^*}(J^*) = T(J^*). \tag{47}$$

Furthermore if there exists an optimal policy there exists an optimal stationary policy.

Proof. If π^* is optimal then $J_{\mu^*} = J^*$ and the result follows from Proposition 5 and Corollary 5.1. Conversely if $T_{\mu^*}(J^*) = T(J^*)$ then $J^* = T(J^*)$ (by Proposition 5) and it follows that $T_{\mu^*}(J^*) = J^*$. Hence by Corollary 5.1, $J_{\mu^*} \leqq J^*$ and it follows that π^* is optimal. If $\bar{\pi} = \{\bar{\mu}_0, \bar{\mu}_1, \cdots\}$ is optimal then we have by using I.1

$$J^* = J_{\bar{\pi}} = \lim_{k \to \infty} (T_{\bar{\mu}_0} T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_k})(\bar{J})$$

$$= T_{\bar{\mu}_0}[\lim_{k \to \infty} (T_{\bar{\mu}_1} \cdots T_{\bar{\mu}_k})](\bar{J}) \geqq T_{\bar{\mu}_0}(J^*) \geqq T(J^*) = J^*.$$

It follows that $T_{\bar{\mu}_0}(J^*) = T(J^*)$ and, by the result just proved, the stationary policy $\{\bar{\mu}_0, \bar{\mu}_0, \cdots\}$ is optimal. Q.E.D.

PROPOSITION 8. *Let* D *and* D.1 *hold. Then a stationary policy* $\pi^* = \{\mu^*, \mu^*, \cdots\}$ *is optimal if and only if*

$$T_{\mu^*}(J_{\mu^*}) = T(J_{\mu^*}). \tag{48}$$

Proof. If π^* is optimal then $J_{\mu^*} = J^*$ and, using Proposition 6, and Corollary 6.2, we have $T_{\mu^*}(J_{\mu^*}) = J_{\mu^*} = J^* = T(J^*) = T(J_{\mu^*})$. Conversely if $T_{\mu^*}(J_{\mu^*}) = T(J_{\mu^*})$ then we obtain from Corollary 6.2, $J_{\mu^*} = T(J_{\mu^*})$, and Proposition 6 yields $J_{\mu^*} \leqq J^*$. Hence π^* is optimal. Q.E.D.

Examples where π^* satisfies (47) or (48) but is not optimal under D or I respectively are given in [1, § 6.4]. It is also easy to modify the proof of Proposition 7 and show the stronger result that if there exists an optimal policy at each $x \in S$ then there exists an optimal stationary policy.

Convergence of the dynamic algorithm—existence of optimal stationary policies. The D.P. algorithm consists of successive generation of the function $T(\bar{J}), T^2(\bar{J}), \cdots$. Under either Assumption I or D the function $J_\infty \in F$ defined by

$$J_\infty(x) = \lim_{N \to \infty} T^N(\bar{J})(x) \quad \forall x \in S \tag{49}$$

is well defined. We would like to investigate the question whether $J_\infty = J^*$. When Assumption D holds, the following proposition shows that we have $J_\infty = J^*$ under mild assumptions.

PROPOSITION 9. *Let* D *hold and assume that either* D.1 *holds or else* $J_N = T^N(\bar{J})$ *for all* N *where* J_N *is the optimal value function of the* N*-stage problem defined by* (36). *Then*

$$J_\infty = J^*.$$

Proof. By Lemma 1 we have that D implies $J^* = \lim_{N\to\infty} J_N$. Proposition 3 shows also that under our assumptions $J_N = T^N(\bar{J})$. Hence $J^* = \lim_{N\to\infty} T^N(\bar{J}) = J_\infty$. Q.E.D.

Under Assumptions I, I.1, and I.2 the equality $J_\infty = J^*$ may easily fail to hold even in very simple deterministic optimal control problems as shown in the example of § 1. This fact has been known since Strauch's work (see [11, p. 880]). Reference [2, p. 608] provides an example where $J_\infty \neq J^*$ even though there exists an optimal stationary policy. The following preliminary result shows that in order to have $J_\infty = J^*$ it is necessary and sufficient to have $J_\infty = T(J_\infty)$.

PROPOSITION 10. *Let* I, I.1, *and* I.2 *hold. Then*

$$J_\infty \leqq T(J_\infty) \leqq T(J^*) = J^*. \tag{50}$$

Furthermore the relation

$$J_\infty = T(J_\infty) = T(J^*) = J^* \tag{51}$$

holds if and only if

$$J_\infty = T(J_\infty). \tag{52}$$

Proof. Clearly we have $J_\infty \leqq J_\pi$ for all $\pi \in \Pi$ and it follows that $J_\infty \leqq J^*$. Furthermore by Proposition 5 we have $T(J^*) = J^*$. Also we have for all $k \geqq 1$,

$$T(J_\infty) = \inf_{u\in U(x)} H(x, u, J_\infty) \geqq \inf_{u\in U(x)} H[x, u, T^k(\bar{J})] = T^{k+1}(\bar{J}).$$

Taking limit of the right side we obtain $T(J_\infty) \geqq J_\infty$ which combined with $J_\infty \leqq J^*$ and $T(J^*) = J^*$ proves (50). If (51) holds then (52) also holds. Conversely let (52) hold. Then, since we have $J_\infty \geqq \bar{J}$, we obtain by Proposition 5, $J_\infty \geqq J^*$ which combined with (50) proves (51). Q.E.D.

In what follows we provide a necessary and sufficient condition for $J_\infty = T(J_\infty)$ (and hence also (51)) to hold under Assumptions I, I.1, and I.2. We subsequently obtain a useful sufficient condition for $J_\infty = T(J_\infty)$ to hold which at the same time guarantees the existence of an optimal stationary policy.

For any $J \in F$ we denote by $E(J)$ the *epigraph* of J, i.e. the subset of $S \times (-\infty, \infty)$ given by

$$E(J) = \{(x, \lambda) \mid J(x) \leqq \lambda\}. \tag{53}$$

Under I we have $T^k(\bar{J}) \leqq T^{k+1}(\bar{J})$ for all k, $J_\infty = \lim_{k\to\infty} T^k(\bar{J})$, and it follows easily that

$$E(J_\infty) = \bigcap_{k=0}^{\infty} E[T^k(\bar{J})]. \tag{54}$$

Consider for each $k \geqq 1$ the subset C_k of $S \times C \times (-\infty, \infty)$ given by

$$C_k = \{(x, u, \lambda) | H[x, u, T^{k-1}(\bar{J})] \leqq \lambda, x \in S, u \in U(x)\}. \tag{55}$$

Denote $P(C_k)$ the projection of C_k on $S \times (-\infty, \infty)$,

$$P(C_k) = \{(x, \lambda) | \exists u \in U(x) \text{ s.t. } (x, u, \lambda) \in C_k\}. \tag{56}$$

In the above relation and later the symbol $\exists$ denotes "there exists" and the initials "s.t." stand for "such that". Consider also the following set:

$$\overline{P(C_k)} = \{(x, \lambda) | \exists \{\lambda_n\} \text{ s.t. } \lambda_n \to \lambda, (x, \lambda_n) \in P(C_k), n = 0, 1, \cdots\}. \tag{57}$$

The set $\overline{P(C_k)}$ is obtained from $P(C_k)$ by adding for each x the point $[x, \bar{\lambda}(x)]$ where $\bar{\lambda}(x)$ is the perhaps missing end point of the half line $\{\lambda | (x, \lambda) \in P(C_k)\}$. We have the following lemma:

LEMMA 2. *Let* I *hold. Then for all* $k \geqq 1$,

$$P(C_k) \subset \overline{P(C_k)} = E[T^k(\bar{J})]. \tag{58}$$

Furthermore we have

$$P(C_k) = \overline{P(C_k)} = E[T^k(\bar{J})] \tag{59}$$

if and only if the infimum is attained for each $x \in S$ *in the relation*

$$T^k(\bar{J})(x) = \inf_{u \in U(x)} H[x, u, T^{k-1}(\bar{J})]. \tag{60}$$

Proof. If $(x, \lambda) \in E[T^k(\bar{J})]$ we have

$$T^k(\bar{J})(x) = \inf_{u \in U(x)} H[x, u, T^{k-1}(\bar{J})] \leqq \lambda.$$

Let $\{\varepsilon_n\}$ be a sequence such that $\varepsilon_n > 0$, $\varepsilon_n \to 0$ and let $\{u_n\}$ be a sequence such that

$$H[x, u_n, T^{k-1}(\bar{J})] \leqq T^k(\bar{J})(x) + \varepsilon_n \leqq \lambda + \varepsilon_n.$$

Then $(x, u_n, \lambda + \varepsilon_n) \in C_k$ and $(x, \lambda + \varepsilon_n) \in P(C_k)$ for all n. Since $\{\lambda + \varepsilon_n\} \to \lambda$ by (57) we obtain $(x, \lambda) \in \overline{P(C_k)}$. Hence

$$E[T^k(\bar{J})] \subset \overline{P(C_k)}. \tag{61}$$

Conversely let $(x, \lambda) \in \overline{P(C_k)}$. Then by (55)–(57) there exists a sequence $\{\lambda_n\}$ with $\lambda_n \to \lambda$ and a corresponding sequence $\{u_n\}$ such that

$$T^k(\bar{J})(x) \leqq H[x, u_n, T^{k-1}(\bar{J})] \leqq \lambda_n.$$

Taking limit as $n \to \infty$ we obtain $T^k(\bar{J})(x) \leqq \lambda$ and $(x, \lambda) \in E[T^k(\bar{J})]$. Hence

$$\overline{P(C_k)} \subset E[T^k(\bar{J})]$$

and using (61), we obtain (58).

To prove that (59) is equivalent to the attainment of the infimum in (60) assume first that the infimum is attained by $\mu^*_{k-1}(x)$ for each $x \in S$. Then for each

$(x, \lambda) \in E[T^k(\bar{J})]$

$$H[x, \mu_{k-1}^*(x), T^{k-1}(\bar{J})] \leqq \lambda$$

implying by (56) that $(x, \lambda) \in P(C_k)$. Hence $E[T^k(\bar{J})] \subset P(C_k)$ and in view of (58) we obtain (59). Conversely if (59) holds we have $[x, T^k(\bar{J})(x)] \in P(C_k)$ for every x for which $T^k(\bar{J})(x) < \infty$. Then by (55), (56) there exists a $\mu_{k-1}^*(x) \in U(x)$ such that

$$H[x, \mu_{k-1}^*(x), T^{k-1}(\bar{J})] \leqq T^k(\bar{J})(x) = \inf_{u \in U(x)} H[x, u, T^{k-1}(\bar{J})].$$

Hence the infimum in (56) is attained for all x such that $T^k(\bar{J})(x) < \infty$. It is also trivially attained by all $u \in U(x)$ whenever $T^k(\bar{J})(x) = \infty$ and the proof is complete. Q.E.D.

Consider now the set $\cap_{k=1}^{\infty} C_k$ and define similarly as in (56), (57) the sets

$$P\Big(\bigcap_{k=1}^{\infty} C_k\Big) = \Big\{(x, \lambda) \mid \exists u \in U(x) \text{ s.t. } (x, u, \lambda) \in \bigcap_{k=1}^{\infty} C_k\Big\}, \tag{62}$$

$$\overline{P\Big(\bigcap_{k=1}^{\infty} C_k\Big)} = \Big\{(x, \lambda) \mid \exists \{\lambda_n\} \text{ s.t. } \lambda_n \to \lambda, (x, \lambda_n) \in P\Big(\bigcap_{k=1}^{\infty} C_k\Big)\Big\}. \tag{63}$$

Using (54) and Lemma 2 it is easy to see that we have

$$P\Big(\bigcap_{k=1}^{\infty} C_k\Big) \subset \bigcap_{k=1}^{\infty} P(C_k) \subset \bigcap_{k=1}^{\infty} \overline{P(C_k)} = \bigcap_{k=1}^{\infty} E[T^k(\bar{J})] = E(J_\infty), \tag{64}$$

$$\overline{P\Big(\bigcap_{k=1}^{\infty} C_k\Big)} \subset \bigcap_{k=1}^{\infty} \overline{P(C_k)} = \bigcap_{k=1}^{\infty} E[T^k(\bar{J})] = E(J_\infty). \tag{65}$$

We have the following proposition:

PROPOSITION 11. *Let* I, I.1, *and* I.2 *hold. Then*:

(a) *There holds* $J_\infty = T(J_\infty)$ (*equivalently* $J_\infty = J^*$) *if and only if*

$$\overline{P\Big(\bigcap_{k=1}^{\infty} C_k\Big)} = \bigcap_{k=1}^{\infty} \overline{P(C_k)}. \tag{66}$$

(b) *There holds* $J_\infty = T(J_\infty)$ (*equivalently* $J_\infty = J^*$) *and in addition the infimum in*

$$J_\infty(x) = \inf_{u \in U(x)} H(x, u, J_\infty) \tag{67}$$

is attained for each $x \in S$ (*equivalently there exists an optimal stationary policy*) *if and only if*

$$P\Big(\bigcap_{k=1}^{\infty} C_k\Big) = \bigcap_{k=1}^{\infty} \overline{P(C_k)}. \tag{68}$$

Proof. (a) Assume $J_\infty = T(J_\infty)$ and let $(x, \lambda) \in E(J_\infty)$, i.e.

$$\inf_{u \in U(x)} H(x, u, J_\infty) = J_\infty(x) \leqq \lambda.$$

Let $\{\varepsilon_n\}$ be any sequence with $\varepsilon_n > 0$, $\varepsilon_n \to 0$. Then there exists a sequence $\{u_n\}$ such that

$$H(x, u_n, J_\infty) \leqq \lambda + \varepsilon_n \quad \forall n = 1, 2, \cdots,$$

and hence

$$H[x, u_n, T^{k-1}(\bar{J})] \leqq \lambda + \varepsilon_n \quad \forall k, n = 1, 2, \cdots.$$

Hence $(x, u_n, \lambda + \varepsilon_n) \in C_k$ for all k, n and $(x, u_n, \lambda + \varepsilon_n) \in \bigcap_{k=1}^{\infty} C_k$ for all n. Hence $(x, \lambda + \varepsilon_n) \in P(\bigcap_{k=1}^{\infty} C_k)$ for all n and since $\lambda + \varepsilon_n \to \lambda$ we obtain $(x, \lambda) \in \overline{P(\bigcap_{k=1}^{\infty} C_k)}$. Therefore

$$E(J_\infty) \subset \overline{P\left(\bigcap_{k=1}^{\infty} C_k\right)}$$

and by (65) we obtain (66).

Conversely let (66) hold. Then we have by (65) $\overline{P(\bigcap_{k=1}^{\infty} C_k)} = E(J_\infty)$. Let $x \in S$ be such that $J_\infty(x) < \infty$. Then $[x, J_\infty(x)] \in \overline{P(\bigcap_{k=1}^{\infty} C_k)}$ and there exists a sequence $\{\lambda_n\}$ with $\lambda_n \to J_\infty(x)$ and a sequence $\{u_n\}$ such that

$$H[x, u_n, T^{k-1}(J^-)] \leqq \lambda_n \quad \forall k, n = 1, 2, \cdots.$$

Taking limit with respect to k and using I.1 we obtain

$$H(x, u_n, J_\infty) \leqq \lambda_n \quad \forall n = 1, 2, \cdots,$$

and since $T(J_\infty)(x) \leqq H(x, u_n, J_\infty)$ it follows that

$$T(J_\infty)(x) \leqq \lambda_n.$$

Taking limit as $n \to \infty$ we obtain

$$T(J_\infty)(x) \leqq J_\infty(x)$$

for all $x \in S$ such that $J_\infty(x) < \infty$. Since the inequality above holds also for all $x \in S$ with $J_\infty(x) = \infty$ we have

$$T(J_\infty) \leqq J_\infty.$$

On the other hand by Proposition 10 we have $J_\infty \leqq T(J_\infty)$. Combining the two inequalities we have $J_\infty = T(J_\infty)$.

(b) Assume $J_\infty = T(J_\infty)$ and that the infimum in (67) is attained for each $x \in S$. Then there exists a function $\mu^* \in M$ such that for all $(x, \lambda) \in E(J_\infty)$

$$H[x, \mu^*(x), J_\infty] \leqq \lambda.$$

Hence $H[x, \mu^*(x), T^{k-1}(\bar{J})] \leqq \lambda$ for all k and $[x, \mu^*(x), \lambda] \in \bigcap_{k=1}^{\infty} C_k$. As a result $(x, \lambda) \in P(\bigcap_{k=1}^{\infty} C_k)$. Hence

$$E(J_\infty) \subset P\left(\bigcap_{k=1}^{\infty} C_k\right)$$

and, by (64), equation (68) follows.

Conversely let (68) hold. We have for all $x \in S$ with $J_\infty(x) < \infty$,

$$[x, J_\infty(x)] \in E(J_\infty) = P\left(\bigcap_{k=1}^{\infty} C_k\right).$$

Hence there exists a $\mu^*(x) \in U(x)$ such that

$$[x, \mu^*(x), J_\infty(x)] \in \bigcap_{k=1}^{\infty} C_k$$

from which

$$H[x, \mu^*(x), T^{k-1}(\bar{J})] \leqq J_\infty(x) \quad \forall k = 0, 1, \cdots.$$

Taking limit and using I.1, we have

$$T(J_\infty)(x) \leqq H[x, \mu^*(x), J_\infty] \leqq J_\infty(x).$$

It follows that $T(J_\infty) \leqq J_\infty$ and since by Proposition 10, $J_\infty \leqq T(J_\infty)$ we finally obtain $J_\infty = T(J_\infty)$. Furthermore the inequality above shows that $\mu^*(x)$ attains the infimum in (67) when $J_\infty(x) < \infty$. When $J_\infty(x) = \infty$ every $u \in U(x)$ attains the infimum and the proof is complete. Q.E.D.

The proposition above states that the equality $J_\infty = T(J_\infty)$, which in view of Proposition 10 is equivalent to the validity of interchanging infimum and limit as shown below:

$$J_\infty = \lim_{k\to\infty} \inf_{\pi\in\Pi} (T_{\mu_0} \cdots T_{\mu_k})(\bar{J}) = \inf_{\pi\in\Pi} \lim_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_k})(\bar{J}) = J^*,$$

is in fact equivalent to the validity of interchanging projection and intersection in the manner of (66) or (68).

The following proposition provides a compactness assumption which guarantees that (68) holds. If C is a topological space (see e.g. [5]) we say that a subset U of C is compact if every collection of open sets that covers U has a finite subcollection that covers U. The empty set in particular is considered to be compact. Any sequence $\{u_n\}$ belonging to a compact set $U \subset C$ has at least one accumulation point $\bar{u} \in U$, i.e., a point $\bar{u} \in U$ every (open) neighborhood of which contains an infinite number of elements of $\{u_n\}$. Furthermore all accumulation points of $\{u_n\}$ belong to U. If $\{U_n\}$ is a sequence of nonempty compact sets of C and $U_n \supset U_{n+1}$ for all n, then the intersection $\bigcap_{n=1}^{\infty} U_n$ is nonempty and compact. This yields the following lemma which will be useful in what follows.

LEMMA 3. *Let C be a topological space, $f: C \to [-\infty, +\infty]$ be a function, and U be a subset of C. Assume that the set $U(\lambda)$ defined by*

$$U(\lambda) = \{u \in U | f(u) \leqq \lambda\}$$

is compact for each $\lambda \in (-\infty, \infty)$. Then f attains a minimum over U.

Proof. If $f(u) = +\infty$ for all $u \in U$ then every $u \in U$ attains the minimum. If $f^* = \inf\{f(u) | u \in U\} < +\infty$ let $\{\lambda_n\}$ be a scalar sequence such that $\lambda_n > \lambda_{n+1}$ for all n and $\lambda_n \to f^*$. Then the sets $U(\lambda_n)$ are nonempty, compact, and satisfy $U(\lambda_n) \supset U(\lambda_{n+1})$ for all n. Hence the intersection $\bigcap_{n=1}^{\infty} U(\lambda_n)$ is nonempty and compact. Let u^* be any point in the intersection. Then $u^* \in U$ and $f(u^*) \leqq \lambda_n$ for all n, and it follows that $f(u^*) \leqq f^*$. Hence u^* attains the minimum of f over U. Q.E.D.

PROPOSITION 12. *Let* I, I.1 *and* I.2 *hold and let the control space C be a topological space. Assume that there exists a nonnegative integer $\bar{k}$ such that for each $x \in S$, $\lambda \in (-\infty, \infty)$ and $k \geqq \bar{k}$ the set*

$$U_k(x, \lambda) = \{u \in U(x) | H[x, u, T^k(\bar{J})] \leqq \lambda\} \tag{69}$$

is compact. Then

$$P\Big(\bigcap_{k=1}^{\infty} C_k\Big) = \bigcap_{k=1}^{\infty} \overline{P(C_k)} \tag{70}$$

and (by Propositions 10 *and* 11) *there holds*

$$J_\infty = T(J_\infty) = T(J^*) = J^*.$$

Furthermore there exists an optimal stationary policy.

Proof. By (64) it will be sufficient to show that

$$P\Big(\bigcap_{k=1}^{\infty} C_k\Big) \supset \bigcap_{k=1}^{\infty} P(C_k), \qquad \bigcap_{k=1}^{\infty} P(C_k) = \bigcap_{k=1}^{\infty} \overline{P(C_k)}. \tag{71}$$

Let $(x, \lambda) \in \bigcap_{k=1}^{\infty} P(C_k)$. Then there exists a sequence $\{u_n\}$ such that

$$H[x, u_n, T^k(\bar{J})] \leqq H[x, u_n, T^n(\bar{J})] \leqq \lambda \quad \forall n \geqq k,$$

or equivalently

$$u_n \in U_k(x, \lambda) \quad \forall n \geqq k.$$

Since $U_k(x, \lambda)$ is compact for $k \geqq \bar{k}$ it follows that the sequence $\{u_n\}$ has an accumulation point $\bar{u}$ and

$$\bar{u} \in U_k(x, \lambda) \quad \forall k \geqq \bar{k}.$$

Hence

$$H[x, \bar{u}, T^k(\bar{J})] \leqq \lambda$$

and $(x, \bar{u}, \lambda) \in \bigcap_{k=1}^{\infty} C_k$. It follows that $(x, \lambda) \in P(\bigcap_{k=1}^{\infty} C_k)$ and

$$P\Big(\bigcap_{k=1}^{\infty} C_k\Big) \supset \bigcap_{k=1}^{\infty} P(C_k).$$

Also by the compactness of $U_k(x, \lambda)$ and the result of Lemma 3 it follows that the infimum in (60) is attained for every $x \in S$ and $k > \bar{k}$. Hence, by Lemma 2, $P(C_k) = \overline{P(C_k)}$ for $k > \bar{k}$ and

$$\bigcap_{k=1}^{\infty} P(C_k) = \bigcap_{k=1}^{\infty} \overline{P(C_k)}.$$

Thus (71) is proved. Q.E.D.

The following proposition shows also that a stationary optimal policy may be obtained in the limit by means of the D.P. algorithm.

PROPOSITION 13. *Let the assumptions of Proposition* 12 *hold. Then*:

(a) *There exists a policy* $\pi^* = \{\mu_0^*, \mu_1^*, \cdots\} \in \Pi$ *attaining the minimum in the D.P. algorithm for all* $k \geqq \bar{k}$, *i.e.*

$$(T_{\mu_k^*} T^k)(\bar{J}) = T^{k+1}(\bar{J}) \quad \forall k \geqq \bar{k}. \tag{72}$$

(b) *For every policy* π^* *satisfying* (72) *the sequence* $\{\mu_k^*(x)\}$ *has at least one accumulation point for each* $x \in S$ *with* $J^*(x) < \infty$.

(c) *If $\mu^*: S \to C$ is such that $\mu^*(x)$ is an accumulation point of $\{\mu_k^*(x)\}$ for all $x \in S$ with $J^*(x) < \infty$, and $\mu^*(x) \in U(x)$ for all $x \in S$ with $J^*(x) = \infty$, then the stationary policy $\{\mu^*, \mu^*, \cdots\}$ is optimal.*

Proof. (a) For an $x \in S$ such that $T^{k+1}(\bar{J})(x) < \infty$ consider a sequence $\{\lambda_n\}$ with $\lambda_n > \lambda_{n+1}$, for all n and $\lambda_n \to T^{k+1}(\bar{J})(x)$. Then the sets $U_k(x, \lambda_n)$ are nonempty and compact and hence their intersection is also nonempty and compact. Any point $\mu_k^*(x)$ in the intersection satisfies $(T_{\mu_k^*} T^k)(\bar{J})(x) = T^{k+1}(\bar{J})(x)$. For an $x \in S$ such that $T^{k+1}(\bar{J})(x) = \infty$ any element of $U(x)$, call it $\mu_k^*(x)$, satisfies $(T_{\mu_k^*} T^k)(\bar{J})(x) = T^{k+1}(\bar{J})(x)$.

(b) For any $\pi^* = \{\mu_0^*, \mu_1^*, \cdots\}$ satisfying (72) and $x \in S$ such that $J^*(x) < \infty$ we have

$$H[x, \mu_n^*(x), T^k(\bar{J})] \leqq H[x, \mu_n^*(x), T^n(\bar{J})] \leqq J^*(x) \quad \forall k \geqq \bar{k}, \quad n \geqq k.$$

Hence we have

$$\mu_n^*(x) \in U_k[x, J^*(x)] \quad \forall k \geqq \bar{k}, \quad n \geqq k.$$

Since $U_k[x, J^*(x)]$ is compact, $\{\mu_n^*(x)\}$ has at least one accumulation point. Furthermore each accumulation point $\mu^*(x)$ of $\{\mu_n^*(x)\}$ belongs to $U(x)$ and satisfies

$$H[x, \mu^*(x), T^k(\bar{J})] \leqq J^*(x) \quad \forall k \geqq \bar{k}. \tag{73}$$

(c) By taking the limit in (73) and using I.1 we obtain

$$H[x, \mu^*(x), J_\infty] = H[x, \mu^*(x), J^*] \leqq J^*(x)$$

for all $x \in S$ with $J^*(x) < \infty$. The relation above holds also trivially for all $x \in S$ with $J^*(x) = \infty$. Hence $T_{\mu^*}(J^*) \leqq J^* = T(J^*)$ which implies $T_{\mu^*}(J^*) = T(J^*)$. It follows, by Proposition 7, that $\{\mu^*, \mu^*, \cdots\}$ is optimal. Q.E.D.

The compactness of the sets $U_k(x, \lambda)$ of (69) may be verified in a number of important special cases. One such case is when $U_k(x, \lambda)$ is a finite set for all k, x, λ. Simply consider the discrete topology on C, i.e. the topology consisting of all subsets of U. In this topology a set is compact if and only if it is finite. For this case the relation $J_\infty = J^*$ for the negative model of Strauch has been shown earlier [11]. There are many other important cases where the compactness of $U_k(x, \lambda)$ can be verified. Several examples have been given in [1, (Chap. 6 and 7)]. It is not our intention to provide an extensive list. Instead we state as an illustration two sets of assumptions which guarantee compactness of $U_k(x, \lambda)$ in the case of the mapping

$$H(x, u, J) = g(x, u) + \alpha(x, u) J[f(x, u)]$$

corresponding to a deterministic optimal control problem.

Assume that $g(x, u) \geqq 0$, $\alpha(x, u) \geqq 0$ for all $x \in S$, $u \in U(x)$ and take $\bar{J}(x) = 0$, $\forall x \in S$. Then compactness of $U_k(x, \lambda)$ is guaranteed if:

(a) $S = R^n$ (n-dimensional Euclidean space), $C = R^m$, $U(x) \equiv C$, f, g, α are continuous in (x, u) and g satisfies $\lim_{n \to \infty} g(x_n, u_n) = \infty$ for every bounded sequence $\{x_n\}$ and every sequence $\{u_n\}$ for which $|u_n| \to \infty$ ($|\cdot|$ is a norm on R^m).

(b) $S = R^n$, $C = R^m$, f, g, and α are continuous, $U(x)$ is compact and nonempty for each $x \in R^n$, and $U(\cdot)$ is a continuous point-to-set mapping from R^n to the set of all compact subsets of R^m.

Aside from the result of Strauch mentioned earlier, other general sufficient conditions which guarantee that an optimal stationary policy exists for special cases of our problem are those of Maitra for discounted problems (see [9] and [6, Thm. 5.11]), and Kushner for free end time problems [8]. In these cases Assumption C is satisfied. In both cases the sufficient conditions or existence of an optimal stationary policy can be shown to follow from Proposition 12.

We finally show that the compactness of the sets $U_k(x, \lambda)$ of (69) guarantees existence of an optimal stationary policy under Assumption C which can be obtained in the limit by means of the D.P. algorithm.

PROPOSITION 14. *The conclusions of Proposition* 13 *hold if Assumptions* I. I.1, *and* I.2 *are replaced by the Contraction Assumption* C.

Proof. (a) The proof of this part is identical to the corresponding proof in Proposition 13.

(b) Let $\pi^* = \{\mu_0^*, \mu_1^*, \cdots\}$ satisfy (72) and define

$$\varepsilon_k = \sup \{\|T^i(\bar{J}) - J^*\| \mid i \geqq k\}, \qquad k = 0, 1, \cdots.$$

We have from (20), (72) and the fact $T(J^*) = J^*$,

$$\|(T_{\mu_n^*}T^n)(\bar{J}) - J^*\| = \|T^{n+1}(\bar{J}) - T^{n+1}(J^*)\|$$

$$\leqq \alpha\|T^n(\bar{J}) - T^n(J^*)\| = \alpha\|T^n(\bar{J}) - J^*\| \quad \forall n \geqq \bar{k},$$

$$\|(T_{\mu_n^*}T^n)(\bar{J}) - (T_{\mu_n^*}T^k)(\bar{J})\| \leqq \alpha\|T^n(\bar{J}) - T^k(\bar{J})\|$$

$$\leqq \alpha\|T^n(\bar{J}) - J^*\| + \alpha\|T^k(\bar{J}) - J^*\|$$

$$\forall n \geqq \bar{k}, \quad k = 0, 1, \cdots.$$

From the above two relations we obtain

$$H[x, \mu_n^*(x), T^k(\bar{J})] \leqq H[x, \mu_n^*(x), T^n(\bar{J})] + 2\alpha\varepsilon_k$$

$$\leqq J^*(x) + 3\alpha\varepsilon_k \quad \forall n \geqq k, \quad k \geqq \bar{k}.$$

It follows that $\mu_n^*(x) \in U_k[x, J^*(x) + 3\alpha\varepsilon_k]$ for all $n \geqq k$ and $k \geqq \bar{k}$, and $\{\mu_n^*(x)\}$ has an accumulation point by the compactness of $U_k[x, J^*(x) + 3\alpha\varepsilon_k]$.

(c) If $\mu^*(x)$ is an accumulation point of $\{\mu_n^*(x)\}$ then $\mu^*(x) \in U_k[x, J^*(x) + 3\alpha\varepsilon_k]$ for all $k \geqq \bar{k}$ or equivalently

$$(T_{\mu^*}T^k)(\bar{J})(x) \leqq J^*(x) + 3\alpha\varepsilon_k \quad \forall x \in S, \quad k \geqq \bar{k}.$$

By using (20) we have for all k

$$\|(T_{\mu^*}T^k)(\bar{J}) - T_{\mu^*}(J^*)\| \leqq \alpha\|T^k(\bar{J}) - J^*\| \leqq \alpha\varepsilon_k.$$

Combining the two inequalities above we obtain

$$T_{\mu^*}(J^*)(x) \leqq J^*(x) + 4\alpha\varepsilon_k \quad \forall x \in S, \quad k \geqq \bar{k}.$$

Since $\varepsilon_k \to 0$ [cf. Prop. 1, part (d)] we obtain $T_{\mu^*}(J^*) \leqq J^*$. Using the fact $J^* = T(J^*) \leqq T_{\mu^*}(J^*)$, we obtain $T_{\mu^*}(J^*) = J^*$ which implies, by Proposition 1, that the stationary policy $\{\mu^*, \mu^*, \cdots\}$ is optimal. Q.E.D.

REFERENCES

[1] D. P. BERTSEKAS, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.

[2] ———, *Infinite time reachability of state space regions by using feedback control*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 604–613.

[3] D. BLACKWELL, *Positive dynamic programming*, Proc. 5th Berkeley Symposium on Mathematics, Statistics, and Probability, vol. 1, 1965, pp. 415–418.

[4] E. V. DENARDO, *Contraction mappings in the theory underlying dynamic programming*, SIAM Rev., 9 (1967), pp. 165–177.

[5] J. DUGUNDJI, *Topology*, Allyn and Bacon, Boston, 1968.

[6] K. HINDERER, *Foundations of Non-Stationary Dynamic Programming with Discrete-Time Parameter*, Springer-Verlag, New York, 1970.

[7] D. H. JACOBSON, *Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 124–131.

[8] H. J. KUSHNER, *Introduction to Stochastic Control*, Holt, Rinehart, and Winston, New York, 1971.

[9] A. MAITRA, *Discounted dynamic programming on compact metric spaces*, Sankhyā, Ser. A, 30 (1968), pp. 211–216.

[10] S. SHREVE AND D. P. BERTSEKAS, *Equivalent deterministic and stochastic optimal control problems*, Proc. 1976 IEEE Conference on Decision and Control, Clearwater Beach, Fla., 1976.

[11] R. E. STRAUCH, *Negative dynamic programming*, Ann. Math. Statist., 37 (1966), pp. 871–890.

[12] S. SHREVE, *Dynamic programming in complete separable metric spaces*, Ph.D. thesis, Dept. of Mathematics, Univ. of Illinois, Urbana, Jan. 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

FILTER STABILITY FOR STOCHASTIC EVOLUTION EQUATIONS*

RICHARD B. VINTER†

Abstract. It is established that the Kalman filter associated with signal and observation processes defined through stochastic evolution equations is stable under very weak hypotheses; namely when appropriate stabilizability/detectability criteria hold. Thus in this general setting we obtain results as sharp as are available for processes taking values in finite dimensional linear spaces. The conditions are shown to be directly verifiable in certain important situations.

1. Introduction. This paper treats stability of the error process associated with the filtering of signals defined through linear stochastic evolution equations.

The viewpoint taken is as follows. A filter is judged to have desirable asymptotic properties if the error process, that is the difference between the signal and the signal conditioned on the observations up to the present time, is stable. Further, since in applications choice of the initial distribution, x_0, of the signal process is often nominal, we require that the asymptotic behavior of the filter be insensitive to modeling of x_0. Accordingly, we term a filter which gives rise to a stable error process, even in the presence of mismodeling of x_0, a *stable filter* and seek conditions assuring this property.

Stability of the error process e_t will be understood in the sense of convergence of the distributions induced on the range space X of the signal process by e_t as t tends to infinity. When X is a finite dimensional linear space, developing sufficient conditions for filter stability is little more than an adjunct to the study of the asymptotic behavior of solutions to the differential Riccati equation. With X a separable Hilbert space however, the probabilistic aspects of the problem are not quite so trivial, for in this case there are a number of possible choices of topologies on the space of distributions on X and some thought needs to be given to the precise manner in which we stipulate that the distributions converge. The core of the paper is § 4 where conditions are given for filter stability analogous to the known results for the case that X is finite dimensional; the results are then interpreted for signals defined through stochastic differential delay equations.

For X finite dimensional, the earliest available sufficient conditions for filter stability [13] are given in terms of controllability/observability hypotheses for time-invariant systems (uniform controllability/uniform observability hypotheses for time varying systems). These were subsequently weakened (in the time invariant case) to stabilizability/detectability hypotheses [25].

In this paper, attention is limited to time-invariant systems. For by so doing, even when X is a separable Hilbert space, we can give conditions for filter stability under merely stabilizability/detectability hypotheses. *We stress that these conditions are no mere technical refinements of more tractable conditions.* Indeed the significance of the results reported here rests on the fact that, in certain important cases, the hypotheses of stabilizability/detectability may be directly tested (in the sense that verification involves examining the properties of a collection of

* Received by the editors April 2, 1975, and in revised form June 29, 1976.
† Imperial College of Science and Technology, London SW7 2BZ, England.

matrices). We argue in fact that conditions for filter stability involving controllability/observability notions are not natural to the present setting. It is not difficult to construct counter-examples illustrating that *approximate* controllability does not assure existence of a limiting solution to the mild differential Riccati equation; on the other hand, although conditions for filter stability can be given in terms of *exact* controllability/*exact* observability, these hypotheses are not met with for most infinite dimensional systems of interest and are, in any case, stronger than the conditions given here [26].

Inevitably a considerable part of the paper is given to properties of the mild differential Riccati equation. Readers may be surprised at the absence of technical conditions in the statement of results here; for time-varying systems, introduction of a morass of technical conditions to justify setting up analogues of different aspects of the finite dimensional results seems inevitable [4], but under our assumptions of time-invariance, almost all of these fall away. The only technical condition that remains is a finite dimensionality assumption on the range of a certain operator to assure asymptotic convergence of the weak differential Riccati equation with respect to the *uniform*, not merely the strong, operator topology. But this finite dimensionality assumption has to be made anyway for the estimated signal to have representation as the output of a Kalman filter and therefore, in relation to the filtering problem, *no loss of generality is involved* (this is judged to be a crucial observation of the paper).

Investigation of the detailed structure of infinite dimensional Kalman filters and of the numerical aspects of their construction is still a largely unexplored area. The paper shows however that, even with present knowledge, it is possible to give tight, and in some cases easily tested, conditions that one important qualitative property holds, that of filter stability.

2. Notations and conventions. Linear spaces are assumed real throughout.

We refer the reader to the appendices for usage and basic results relating to perturbed evolution operators and separable Hilbert space valued random variables.

For brevity, when the meaning is clear from context, P_t will denote either the function $t \mapsto P_t$ or the value of the function at t, etc.

3. The weak differential Riccati equation. Take X, Y, U (real) Hilbert spaces, $\{T_t \in \mathscr{L}(X) | t \geqq 0\}$ a C^0 semigroup with infinitesimal generator A,

$$G \in \mathscr{L}(X), \quad B \in \mathscr{L}(U, X), \quad C \in \mathscr{L}(X, Y), \quad R \in \mathscr{L}(U).$$

It is assumed that G, R are self-adjoint, nonnegative operators, and that there exists some $\varepsilon > 0$ such that

$$\|Rx\| \geqq \varepsilon \|x\|, \qquad \text{all } x \in U.$$

THEOREM 3.1. *There exists a unique* P_t: $(-\infty, 0] \to \mathscr{L}(X)$ *in the class of strongly continuous functions such that* (i) P_t *is self-adjoint* $t \leqq 0$, *and* (ii) *for each* $h \in \mathscr{D}\{A\}$,

$t \mapsto \langle P_t h, h\rangle$ is locally absolutely continuous with

$$\frac{d}{dt}\langle P_t h, h\rangle + 2\langle Ah, P_t h\rangle + \langle [C^*C - P_t B R^{-1} B^* P_t]h, h\rangle = 0, \quad \text{all } t \leqq 0$$

$$P_0 = G. \tag{3.1}$$

For a condensed proof the reader is referred to § 11.

We remark that the theorem may be given an alternative statement, where (i) is dropped and (ii) is replaced by

$$\frac{d}{dt}\langle P_t h, \bar{h}\rangle + \langle Ah, P_t \bar{h}\rangle + \langle P_t h, A\bar{h}\rangle + \langle [C^*C - P_t B R^{-1} B^* P_t]h, \bar{h}\rangle = 0, \tag{3.2}$$

all $t \leqq 0$, each $h, \bar{h} \in \mathscr{D}\{A\}$. The unique solution to (3.2) is everywhere self-adjoint and coincides with P_t of the theorem.

The theorem draws known results from a number of sources: existence and uniqueness of solutions to the 'integral Riccati equation',

$$\begin{aligned}\langle P_t h, h\rangle = &\int_t^0 \langle \tilde{T}_{\sigma,t} h, [C^*C + P_\sigma B R^{-1} B^* P_\sigma]\tilde{T}_{\sigma,t} h\rangle \, d\sigma \\ &+ \langle \tilde{T}_{0,t} h, G\tilde{T}_{0,t} h\rangle, \quad \text{all } t \leqq 0, \quad h \in X\end{aligned} \tag{3.3}$$

(with $\tilde{T}_{t,s}$, T_t perturbed by $-BR^{-1}B^*P_t$) in the class of weakly continuous functions everywhere self-adjoint, was established in [2], where the interpretation

$$\begin{aligned}&\langle P_t h, h\rangle = \inf\{J_u \mid u \in L^2(t, 0; u)\}, \\ &J_u = \int_t^0 \{\langle x_\tau^u, C^*C x_\tau^u\rangle + \langle u_\tau, Ru_\tau\rangle\}\, d\tau + \langle x_0^u, G x_0^u\rangle, \\ &x_\tau^u = T_{\tau - t} h + \int_t^\tau T_{\tau-\sigma} B u_\sigma d\sigma, \qquad \tau \geqq t\end{aligned} \tag{3.4}$$

was given. (For definition of the perturbed evolution operator $\tilde{T}_{t,s}$ we refer to § 9.)

It was shown (for example) in (21) that when $t \mapsto P_t$ is strongly continuous (as here) then $t \mapsto P_t$ satisfies the weak differential equation (3.1).

Finally, the result on uniqueness of the solution to (3.1) follows an idea in (4). It improves on a number of results giving uniqueness under a variety of conditions (involving finite dimensionality of range $\{B\}$, inclusion of range $\{B\}$ in $\mathscr{D}\{A\}$ etc.) [2], [21] all of which are dispensed with here.

The next theorem gives the best available conditions under which the "algebraic Riccati equation" has a unique solution and its identification through the limiting solution of the Riccati equation.

Recall the definitions:

DEFINITION 3.1. Take X, T, U real Hilbert spaces and $B \in \mathscr{L}(U, X)$, $C \in \mathscr{L}(X, Y)$. Let $\{T_t \in \mathscr{L}(X) \mid t \geqq 0\}$ be a C^0 semigroup with infinitesimal generator A. Then

(i) (A, B) is *stabilizable* if there exists $K \in \mathscr{L}(X, U)$ such that $A + BK$ generates an exponentially stable semigroup,

(ii) (C, A) is *detectable* if there exists $M \in \mathscr{L}(Y, X)$ such that $A + MC$ generates an exponentially stable semigroup.

Of course in the above definition, by exponential stability of the semigroup T_t we mean that the growth

$$\lim_{t \to \infty} \frac{\log \|T_t\|}{t}$$

is negative.

THEOREM 3.2. *Consider the equation in P,*

$$\begin{aligned} &2\langle Ah, Ph\rangle + \langle [C^*C - PBR^{-1}B^*P]h, h\rangle = 0, \quad \text{all } h \in \mathscr{D}\{A\}, \\ &P \in \mathscr{L}(X), \quad P = P^*, \quad P \geqq 0. \end{aligned} \tag{3.5}$$

(i) *If (A, B) is stabilizable, then* (3.5) *has a solution.*

(ii) *If (C, A) is detectable, then* (3.5) *has at most one solution P^∞, and (if the solution exists) $A - BR^{-1}B^*P^\infty$ generates an exponentially stable semigroup.*

(iii) *If (A, B) is stabilizable and (C, A) is detectable, then, for any $G \in \mathscr{L}(X)$, $G \geqq 0$, $G = G^*$,*

$$P_t \to P^\infty \text{ (strongly)} \quad \text{as } t \to -\infty,$$

where P_t is as given in Theorem 3.1 *and P^∞ is the unique solution to* (3.5).

Theorem 3.2 is proved in Appendix C. Notice that in requiring merely detectability of (C, A) and stabilizability of (A, B) for existence and uniqueness of a solution to (3.5), these results are as strong as those available in [25] for the case that X is finite dimensional.[1]

That stabilizability of (A, B) implies existence of a solution is well known [6], [14], [7]. Uniqueness under the detectability hypothesis is a much more recent result and has awaited a lemma of Zabczyk (Lemma 11.1 of Appendix C), which replaces the crucial [25, Lem. 12.2, p. 299]. Finally, the convergence property (iii) is well known for X finite dimensional. It has also been established in the present setting [14], [6], [2] for $G = 0$. The general result here which is needed in our study of filter stability is apparently new.

The results given above are in a form convenient for the study of quadratic control problems. Finally we provide a rephrasing of results and extensions which will be useful in studying the filtering problem.

PROPOSITION 3.1. *Take X, U (real) Hilbert spaces, $\{T_t \in \mathscr{L}(X) | t \geqq 0\}$ a C^0 semigroup with generator A, $N \in \mathscr{L}(Y)$, $P_0 \in \mathscr{L}(X)$, $B \in \mathscr{L}(U, X)$, $C \in \mathscr{L}(X, Y)$. We assume that P_0, N are self-adjoint, nonnegative, and there exists some $\varepsilon > 0$ with*

$$\|Ny\| \geqq \varepsilon \|y\|, \quad \text{all } y \in Y.$$

Then there exists a unique P_t: $[0, \infty) \to \mathscr{L}(X)$ such that

(i) *$P_t = P_t^*$ each $t \geqq 0$ and $t \mapsto P_t$ is strongly continuous,*

[1] We observe however that (35) has a solution under the weaker, implicit, hypothesis that the infimum of the corresponding 'infinite time' problem, for arbitrary initial conditions, is less than infinity.

(ii) *for each* $h \in \mathcal{D}\{A^*\}$, $t \mapsto \langle P_t h, h\rangle$ *is locally absolutely continuous with*

$$\frac{d}{dt}\langle P_t h, h\rangle = 2\langle P_t h, A^* h\rangle + \langle h, [BB^* - P_t C^* N^{-1} C P_t] h\rangle, \quad \text{all } t \geqq 0, \tag{3.6}$$

and P_0 *is as given.*

Now suppose that (A, B) *is stabilizable and* (C, A) *is detectable. In this case*

(i) $P_t \to P^\infty$ *strongly as* $t \to \infty$ *where* P^∞ *is the unique solution of*

$$\begin{aligned} &P^\infty \in \mathcal{L}(X), \quad (P^\infty)^* = P^\infty, \quad P^\infty \geqq 0, \\ &2\langle P^\infty h, A^* h\rangle + \langle [BB^* - PC^* N^{-1} CP] h, h\rangle = 0, \quad \text{all } h \in \mathcal{D}\{A^*\}. \end{aligned} \tag{3.7}$$

(ii) $A^* - C^* N^{-1} C P^\infty$ *generates an exponentially stable semigroup* T_t^∞

(iii) *we have the estimate: for all* $t \geqq 0$, $h \in X$,

$$\langle P_t h, h\rangle \leqq \langle T_t^\infty h, P_0 T_t^\infty h\rangle + \int_0^t \langle T_s^\infty h, [BB^* - P^\infty C^* N^{-1} C P^\infty] T_s^\infty h\rangle \, ds.$$

Finally, if in addition we assume that the range of C *is finite dimensional, then writing* $\tilde{T}_{t,s}$ *for* T_t *perturbed by* $-P_t C^* N^{-1} C$,

$$\|\tilde{T}_{t,0}\| \to 0, \qquad t \to \infty. \tag{3.8}$$

The convergence property (3.8) is crucial for the results of succeeding sections. To conclude (3.8) we need to assume that C has finite dimensional range; this will however result in no loss of generality as regards applications to the filtering problem.

The proposition is proved in Appendix C.

4. The filtering problem. For usage relating to separable Hilbert space valued random variables we refer to Appendix B.

Consider now the stochastic evolution equation

$$\begin{aligned} dx_t &= Ax_t \, dt + B \, dw_t, \\ x_0 &\text{ given,} \end{aligned} \tag{4.1}$$

with observation process

$$\begin{aligned} dz_t &= Cx_t \, dt + F \, dv_t, \\ z_0 &= 0. \end{aligned} \tag{4.2}$$

Here, $(\Omega, \mathcal{S}, \mathcal{P})$ is a complete probability space; U, X, (real) separable Hilbert spaces;

$$B \in \mathcal{L}(U, X), \quad C \in \mathcal{L}(X, \mathbb{R}^k), \quad F \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^k),$$

$FF^* > 0$; $\{w_t | t \geqq 0\}$, U-valued separable Wiener process (on $(\Omega, \mathcal{S}, \mathcal{P})$) with (constant) incremental covariance W; $\{v_t | t \geqq 0\}$, $\mathbb{R}^k$-valued separable Wiener process with unit incremental covariance, A, the infinitesimal generator of a C^0 semigroup $\{T_t \in \mathcal{L}(X) | t \geqq 0\}$; x_0, a Gaussian random variable taking values in X with zero mean and covariance P_0. We assume that w_t, v_t are independent and that x_0 is independent of future increments of w_t, v_t.

Equation (4.1) is interpreted as

$$x_t = T_t x_0 + \int_0^t T_{t-\sigma} B \, dw_\sigma \tag{4.3}$$

with the last term a Wiener integral. There exists a measurable version of x_t with summable paths; such a version is used in evaluating z_τ, $\tau \geqq 0$, as

$$z_\tau = \int_0^\tau C x_\sigma \, d\sigma + F v_t.$$

For fixed $t > 0$, we find that $\{z_\tau | 0 \leqq \tau \leqq t\}$ takes values almost surely in $L^2(0, t; \mathbb{R}^k)$ and defines a second order $L^2(0, t; \mathbb{R}^k)$-valued random variable.

The *filtering problem* is that of characterizing the process $\hat{x}_t$, $t \geqq 0$, where

$$\hat{x}_t = E\{x_t | z_\tau, 0 \leqq \tau \leqq t\}. \tag{4.4}$$

Concerning the filtering problem we have the following results:

THEOREM 4.1. *$\hat{x}_t$ has representation*

$$\hat{x}_t = \int_0^t \tilde{T}_{t,\tau} P_\tau C^*(FF^*)^{-1} \, dz_\tau \tag{4.5}$$

where $\tilde{T}_{t,\tau}$ is T_t perturbed by $-P_t C^*(FF^*)^{-1}C$, $P_t: [0, \infty) \to \mathcal{L}(X)$ is the unique strongly continuous function such that P_t is self-adjoint for all $t \geqq 0$, and for $h \in \mathcal{D}\{A^*\}$, $t \mapsto \langle h, P_t h\rangle$ is locally absolutely continuous with

$$\frac{d}{dt}\langle h, P_t h\rangle = 2\langle A^* h, P_t h\rangle + \langle h, [BWB^* - P_t C^*(FF^*)^{-1} C P_t] h\rangle, \tag{4.6}$$

$$P_0 \quad \textit{as given.}$$

Proof. Take $\tilde{T}_{t,s}$ to be T_t perturbed by $-P_t C^*(FF^*)^{-1}C$. It is known [15] that $\hat{x}_t$ has representation (4.5) where now $P_t: [0, \infty) \to \mathcal{L}(X)$ is the unique strongly continuous function such that P_t is self-adjoint, $t \geqq 0$, and for all $h \in X$, $t \geqq 0$,

$$\begin{aligned} \langle P_t h, h\rangle &= \langle \tilde{T}^*_{t,0} h, P_0 \tilde{T}_{t,0} h\rangle \\ &\quad + \int_0^t \langle \tilde{T}^*_{t,\sigma} h, [BWB^* + P_\sigma C^*(FF^*)^{-1} C P_\sigma] \tilde{T}^*_{t,\sigma} h\rangle \, d\sigma \end{aligned} \tag{4.7}$$

(see also [3]).

Using the strong continuity of $t \mapsto P_t$ we may justify term-by-term differentiation of (4.7) to obtain (4.6) as in the proof of Theorem 3.1.

That P_t is the unique solution in the specified class of the weak differential equation (4.6) now follows from Theorem 3.1, on considering a change of variable $t \mapsto -t$.

Let us recall that in order to have $\hat{x}_t$ represented as the output of a Kalman filter it is necessary that the incremental covariance of v_t be invertible. Since the incremental covariance is necessarily a trace class operator [1], this in turn constrains the process z_t to take values in a *finite dimensional linear space*. In

particular C has finite dimensional range (a technical condition introduced in Proposition 3.1).[2]

5. Filter stability. Let us write

$\mathscr{Z} = \{\mathbb{R}^k$-valued processes on $[0, \infty)$, $z_t | dz_t = g_t\,dt + d\xi_t$ with g_t an $\mathbb{R}^k$-valued measurable process with summable paths, ξ_t an $\mathbb{R}^k$-valued separable Wiener process$\}$.

We call the map $\mathscr{K}$ carrying processes in $\mathscr{Z}$ into X-valued processes on $[0, \infty)$ through the map (4.5) the *Kalman filter.*

Notice that in particular, the process $\{z'_t | t \geqq 0\}$ lies in the domain of the filter where

$$\begin{aligned} x'_t &= T_t x'_0 + \int_0^t T_{t-\sigma} B\, dw_\sigma, \\ z'_t &= \int_0^t C x'_\tau\, d\tau + F v_t. \end{aligned} \tag{5.1}$$

Equation (5.1) is identical with (4.1), (4.2) *except* that we now allow x'_0 to be an arbitrary X-valued random variable. It is important to note that x'_0 is not assumed Gaussian so that $\hat{x}'_t$ defined by

$$\hat{x}'_t = (\mathscr{K} z')(t), \qquad t \geqq 0 \tag{5.2}$$

will not in general be Gaussian or even second order.

We wish now to introduce the concept of filter stability in the present setting. First we recall some definitions concerning weak convergence of measures:

DEFINITION 5.1. Take X a topological space, and write $\mathscr{B}_x = \sigma\{$open sets in $X\}$. A sequence of finite measures $\{\mu_i\}$ on $\mathscr{B}_X$ *converges weakly* (with respect to the X topology) to μ a finite measure on $\mathscr{B}_X$ written

$$\mu_i \Rightarrow \mu \quad \text{(w.r.t. } X \text{ topology)}$$

when for every bounded function $f: X \to \mathbb{R}$, continuous with respect to the X topology,

$$\int_X f\, d\mu_i \to \int_X f\, d\mu.$$

We recall that weak convergence of a sequence of probability measures $\mu_i \to \mu$ on $\mathscr{B}_X$ (X as defined above) is equivalent to convergence of the distribution functions induced by the μ_i's on the range of f (in the sense of pointwise convergence at continuity points of the limiting distribution) for every bounded $f: X \to \mathbb{R}$ continuous with respect to the topology under consideration.

Now a highly desirable property of the filter is that the error process $x_t - \hat{x}_t$ should be both 'stable' and insensitive to errors in the modeling of x_0.

[2] Note however, that when C does not have finite dimensional range, then the filter may still be defined as the best linear estimator. Study of the asymptotic properties of the filter in this general setting awaits investigation.

This property is made precise in the definition of filter stability; here stability of the error process is taken to mean that the measures induced on $\mathcal{B}_X$ by the process converge to a limit, while insensitivity to modeling errors requires that this limit be independent of x_0.

DEFINITION 5.2. Take X as in § 4 and write $\mathcal{B}_X = \sigma\{\text{strongly open sets in } X\}$. Let $\mathcal{T}$ be a topology on X which we assume to be not finer than the strong X topology. Let x'_t, $\hat{x}'_t$ be processes as defined by (5.1), (5.2).

Then the Kalman filter is *stable (w.r.t. $\mathcal{T}$ topology on X)* if there exists some measure μ on $\mathcal{B}_X$ such that given any X-valued random variable x_0, then

$$\mu_t \Rightarrow \mu \quad (\text{w.r.t. } \mathcal{T} \text{ topology on } X), \tag{5.3}$$

where μ_t is the measure on $\mathcal{B}_X$ induced by $\hat{x}'_t - x'_t$.

The stipulation that the $\mathcal{T}$ topology on X be not finer than the strong topology on X assures that the X-valued random variable $\hat{x}'_t - x'_t$ defines a measurable map from the underlying pre-probability space $(\Omega, \mathcal{S})$ to $(X, \sigma\{\mathcal{T}\text{-open sets}\})$, whence (5.3) is meaningful.

We now state the main result of the paper (here $W^{1/2}$ is taken to be the nonnegative, self-adjoint square root of W as defined for example in [18]):

THEOREM 5.1. *Suppose that $(A, BW^{1/2})$ is stabilizable, and that (C, A) is detectable. Then the Kalman filter is stable with respect to the weak topology on X.*

If $x_0 = 0$ and we modify the definition of filter stability to stipulate that x'_0 be independent of future increments of w_t, v_t, then the filter is stable with respect to the strong topology.

In either case the limiting measure is a zero mean Gaussian measure on X with covariance P, P being the unique $P \in \mathcal{L}(X)$ such that $P = P^*$, $P \geqq 0$ and

$$2\langle A^*h, Ph\rangle + \langle h, [BWB^* - PC^*(FF^*)^{-1}CP]h\rangle = 0 \quad \text{all } h \in \mathcal{D}\{A^*\}.$$

6. Proof of Theorem 5.1. Take x'_0 an arbitrary X-valued random variable, and define x'_t, $\hat{x}'_t$ as in (5.1), (5.2). Proof of the theorem will hinge on the following representation of the error process $e'_t = \hat{x}'_t - x'_t$:

$$e'_t = \tilde{T}_{t,0}x'_0 + \int_0^t \tilde{T}_{t,\sigma}[B\,dw_\sigma + P_\sigma C^*(FF^*)^{-1}C\,dv_\sigma]$$

(see (15, Prop. 10.1)), where P_σ, $\sigma \geqq 0$, is as in Theorem 4.1, and $\tilde{T}_{t,s}$ is T_t perturbed by $-P_tC^*(FF^*)^{-1}C$.

We shall express e'_t as the sum of two processes

$$e'_t = a_t + b_t$$

where

$$a_t = \tilde{T}_{t,0}(x'_0 - x_0) \tag{6.1}$$

$$b_t = \tilde{T}_{t,0}x_0 + \int_0^t \tilde{T}_{t,\sigma}[B\,dw_\sigma + P_\sigma C^*(FF^*)^{-1}C\,dv_\sigma] \tag{6.2}$$

and x_0 is a zero mean Gaussian random variable with covariance P_0, independent of future increments of w_t, v_t.

DEFINITION 6.1. Take X a topological space, $\mathcal{B}_X = \sigma\{\text{open sets in } X\}$. We say that a family of finite measures on $\mathcal{B}_X$ is *weakly sequentially precompact* (w.r.t. X topology), when every sequence in the family contains a subsequence weakly convergent (w.r.t. X topology) to a finite measure on $\mathcal{B}_X$.

LEMMA 6.1. *Take X a topological space, and let $\{\mu_t | t \in \mathcal{I}\}$ be a family of measures on $\mathcal{B}_X$ with $\sup_{t\in\mathcal{I}}\{\mu_t(X)\} < \infty$. Suppose there exists an increasing sequence K_m in $\mathcal{B}_X$ such that*

(a) *each K_m is compact,*

(b) *each K_m (with the topology induced by X) is separable, metrizable,*

(c) *$\mu_t(X \setminus K_m) \to 0$ as $m \to \infty$, uniformly in $t \in \mathcal{I}$. Then $\{\mu_t\}$ is weakly sequentially precompact (w.r.t. X topology).*

Proof.[3] This result is well known for X a complete, separable metric space [10]. But scrutiny of the sufficiency part of the proof of [10, Thm. 1, p. 441] reveals that only separability and metrizability of the K_m's is required for the conclusions of the lemma to hold.

Proof of Theorem 5.1. Under the hypotheses that $(A, BW^{1/2})$ is stabilizable and (C, A) is detectable we prove that the filter is stable (w.r.t. the weak X topology).

The operator C has finite dimensional range. By Proposition 3.1 then, $\|\tilde{T}_{t,0}\| \to 0$ as $t \to \infty$. Note also that by standard theory $\|\tilde{T}_{t,0}\|$ is locally bounded, whence $\|\tilde{T}_{t,0}\|$ is *uniformly bounded on* $[0, \infty)$.

Writing $K_m = \{x \in X | \|x\| \leqq m\}$, we have that $\cup_{m=1,2,\ldots} K_m = X$ so that from the sigma-additivity of the measure induced on $\mathcal{B}_X$ by $x_0' - x$

$$\mathcal{P}\{\|x_0' - x_0\| \leqq m\} \to 1 \quad \text{as } m \to \infty.$$

But for each $t \geqq 0$, taking a_t as in (6.1),

$$\mathcal{P}\{\|a_t\| \leqq m\} \geqq \mathcal{P}\{\|\tilde{T}_{t,0}\| \cdot \|x_0' - x_0\| \leqq m\}.$$

The uniform bound on $\|\tilde{T}_{t,0}\|$ then gives

$$\mathcal{P}\{\|a_t\| \leqq m\} \to 1 \quad \text{as } m \to \infty \tag{6.3}$$

uniformly in $t \geqq 0$.

Now suppose that we can show that

$$\mathcal{P}\{\|b_t\| \leqq m\} \to 1 \quad \text{as } m \to \infty \quad \text{uniformly in } t \geqq 0. \tag{6.4}$$

It will follow that

$$\mathcal{P}\{\|e_t'\| \leqq m\} \to 1 \quad \text{as } m \to \infty \quad \text{uniformly in } t \geqq 0. \tag{6.5}$$

Indeed

$$\begin{aligned}\mathcal{P}\{\|e_t'\| > m\} &\leqq \mathcal{P}\{\|a_t\| + \|b_t\| > m\} \\ &\leqq \mathcal{P}\{\|a_t\| > m/2\} + \mathcal{P}\{\|b_t\| > m/2\} \\ &\to 0 \quad \text{as } m \to \infty \quad \text{uniformly in } t \geqq 0.\end{aligned}$$

[3] The reviewer has pointed out that the lemma also follows from Prokhorov's criterion on $U_m K_m$ and the fact that the space of measures on $\mathcal{B}_{U_m K_m}$ with the weak topology is a Lusin space so that its compact subspaces are metrizable (see [27, Thms. 3 and 7, Appendix paragraph 3]).

The family $\{K_m; m = 1, 2, \cdots\}$ satisfies the conditions (a), (b) of Lemma 6.1 w.r.t. weak topology on X. Indeed for each m, K_m is weakly compact by Alaoglu's theorem [8, p. 424]. Each K_m is also metrizable [8, p. 426] and obviously separable (being separable by assumption w.r.t. a stronger topology). Furthermore writing μ_t for the measure induced on $\mathcal{B}_X$ by e'_t, we have from (6.5)

$$\mu_t(K_m) \to 1 \quad \text{as } m \to \infty \quad \text{uniformly in } t \geqq 0.$$

It follows from Lemma 6.1 that the family $\{\mu_t | t \geqq 0\}$ is weakly sequentially pre-compact (w.r.t. weak topology on X). Take a sequence $\{t_j\}$ increasing to infinity. We know that there exists a subsequence, also written $\{t_j\}$, such that

$$\mu_{t_j} \Rightarrow \bar{\mu} \quad \text{(w.r.t. weak topology on } X)$$

for some measure $\bar{\mu}$ on $\mathcal{B}_X$. In particular, since $x \mapsto e^{i\langle x, x^*\rangle}$ defines a weakly continuous bounded functional on X for $x^* \in X$ we have, writing χ_ν for the characteristic function of the measure ν,

$$\chi_{\mu_{ij}}(x^*) \to \chi_{\bar{\mu}}(x^*) \quad \text{as } j \to \infty, \quad \text{each } x^* \in X.$$

b_t is a zero mean Gaussian random variable with covariance P_t [15]. From Proposition 3.1 $P_t \to P$ strongly, with P as given in Theorem 5.1. It follows that

$$E\{e^{i\langle b_t, x^*\rangle}\} = e^{-\langle x^*, P_t x^*\rangle} \to e^{-\langle x^*, Px^*\rangle} \quad \text{as } t \to \infty, \quad \text{each } x^* \in X.$$

Now $\|\tilde{T}_{t,0}\| \to 0$ as $t \to \infty$ implies that

$$a_t \to 0 \quad \text{almost surely,} \quad \text{as } t \to \infty. \tag{6.6}$$

For each $t \geqq 0$ we have that

$$\chi_{\mu_t}(x^*) = E\{e^{i\langle b_t, x^*\rangle}\} + E\{e^{i\langle a_t, x^*\rangle} - 1)\, e^{i\langle b_t, x^*\rangle}\}.$$

But

$$E\{e^{i\langle a_t, x^*\rangle} - 1)\, e^{i\langle b_t, x^*\rangle}\} \to 0 \quad \text{as } t \to \infty$$

for each $x^* \in X$ by (6.6) and dominated convergence. Taking the limit $j \to \infty$, we have

$$\chi_{\bar{\mu}}(x^*) = e^{-\langle x^*, Px^*\rangle}, \quad \text{each } x^* \in X.$$

However the values of the characteristic function uniquely define a measure on $\mathcal{B}_X$ [16, p. 152]. It follows that

$$\mu_{tj} \Rightarrow \bar{\mu} \quad \text{(w.r.t. the weak } X \text{ topology)},$$

where $\bar{\mu}$ is zero mean Gaussian measure with covariance P. We have then that every sequence $\{t_j\}$ increasing to infinity contains a subsequence, also written $\{t_j\}$ such that μ_{t_j} converges to a *unique* measure $\bar{\mu}$. An elementary argument now gives

$$\mu_t \Rightarrow \bar{\mu} \quad \text{as } t \to \infty.$$

To complete the proof of the first assertion of the theorem it remains to verify (6.4).

Write T_t^∞ for the semigroup generated by $A - PC^*(FF^*)^{-1}C$. Then by Proposition 3.1,

$$\langle P_t h, h\rangle \leqq \langle (T_t^\infty)^* h, P_0 (T_t^\infty)^* h\rangle + \int_0^t \langle (T_\sigma^\infty)^* h, [BWB^* + PC^*(FF^*)^{-1}CP](T_\sigma^\infty)^* h\rangle \, d\sigma.$$

But P_0, W, (FF^*) are trace class operators. Using well-known properties of such operators [1], we show that P_t is trace class with

$$\|P_t\|_{\mathrm{tr}} \leqq \|T_t^\infty\|^2 \|P_0\|_{\mathrm{tr}} + [\|B\|^2 \|W\|_{\mathrm{tr}} + \|CP\|^2 \cdot \|FF^*\|_{\mathrm{tr}}] \int_0^t \|T_\sigma^\infty\|^2 \, d\sigma.$$

($\|\cdot\|_{\mathrm{tr}}$ denotes the trace norm.) By Proposition 3.1, T_t^∞ is exponentially stable; it follows that $\{P_t | t \geqq 0\}$ is uniformly bounded in trace norm. Thus $\{b_t | t \geqq 0\}$ is a family of zero mean Gaussian random variables whose covariances are uniformly bounded in trace norm.

However,

$$E\{\|b_t\| > m\} \leqq (1/m^2) E\{\|b_t\|^2\}. \tag{6.7}$$

We note though that by the special properties of b_t,

$$\begin{aligned} E\{\|b_t\|^2\} &= E\{\lim_{n\to\infty} \sum_{i=1}^n \langle x, e_i\rangle^2\} \\ &= \lim_{n\to\infty} E\{\sum_{i=1}^n \langle x, e_i\rangle^2\} \quad \text{(by monotone convergence)} \\ &= \lim_{n\to\infty} \sum_{i=1}^n \lambda_t^i = \|P_t\|_{tr}. \end{aligned}$$

Here $\{e_i\}$ is an orthonormal basis for X, comprising eigenvectors of P_t corresponding to all nonzero eigenvalues together with a countable set which spans the null space of P_t and $\lambda_t^i = \|P_t e_i\|$. In view of the uniform bound on $\|P_t\|_{\mathrm{tr}}$, (6.4) now follows from (6.6). Thus the first assertion is proved.

Under the further assumptions that $P_0 = 0$ and that x_0' is independent of future increments of w_t, v_t we proceed to show that the filter is stable w.r.t. the strong X topology.

Take e_i', a_t, b_t as above (x_0 under the present assumptions is zero almost surely). Write μ_t, μ_t^a, μ_t^b for the measures induced on $\mathcal{B}_X$ by e_t', a_t, b_t respectively. Consider a_t. We have $a_t \to 0$ (strongly) as $t \to \infty$, almost surely, whence for arbitrary $f: X \to \mathbb{R}$, bounded and strongly continuous,

$$\int_X f \, d\mu_t^a = E\{f(a_t)\} \to \int_X f \, d(\delta(0)) \quad \text{as } t \to \infty$$

($\delta(0)$ probability measure concentrated on $\{0\}$).

Consider now $\mu_t^b \cdot \mu_t^b$ is a zero mean Gaussian measure with covariance P_t. As in the proof of Theorem 3.1 we show that

$$\langle P_t h, h\rangle \to \langle \tilde{P}h, h\rangle, \quad \text{all } h \in X, \tag{6.8}$$

where

$$\tilde{P} = \int_0^\infty \langle (T_t^\infty)^* h, [BWB^* + PC^*(FF^*)^{-1}CP](T_t^\infty)^* h\rangle \, dt$$

(recall T_t^∞ is exponentially stable, so that $\tilde{P}$ is well-defined), and *because of the assumption* $P_0 = 0$,

$$\langle P_t h, h\rangle \leqq \langle \tilde{P}h, h\rangle, \quad \text{all } h \in X. \tag{6.9}$$

The same reasoning as above establishes that $\tilde{P}$ is a trace class operator.

By a well-known result [11, p. 142], (6.8) and (6.9) imply that

$$\mu_t^b \Rightarrow \bar{\mu} \text{ weakly} \quad (\text{w.r.t. the strong } X \text{ topology}),$$

where $\bar{\mu}$ is zero mean Gaussian measure with covariance P.

By assumption x_0' is independent of future increments of w_t, v_t; it follows that a_t, b_t are independent random variables. In consequence

$$\mu_t = \mu_t^a * \mu_t^b \quad (\text{convolution}).$$

But [16, p. 57] the operation of convolution on measures defined on the Borel sets of separable metric spaces is continuous (w.r.t. the weak measure topology); it follows that

$$\mu_t = \mu_t^a * \mu_t^b \Rightarrow \delta(0) * \bar{\mu} = \bar{\mu}$$

(w.r.t. the strong X topology) as required.

7. Verification of the stabilizability/detectability assumptions. In this section we indicate how the hypotheses of stabilizability and detectability under which filter stability is assured are in certain circumstances *directly verifiable*. The development here will be in Banach spaces.

Take X, a (real) Banach space. The closed linear map $A: X \to X$ with dense domain is termed *discrete* in case for some λ in the resolvent set of A $(\lambda I - A)^{-1}: X \to X$ is compact (terminology of [9]).

Let $\{T_t \in \mathcal{L}(X) | t \geqq 0\}$ be a C^0 semigroup with generator A. In the case that A is a discrete operator and $T_{\bar{t}}$ is compact for some $\bar{t} > 0$, the semigroup has very special properties. In particular we find that the set

$$\mathcal{S} = \{\lambda \in \text{spectrum}\,\{A\} | \text{Re}\,\{\lambda\} \geqq 0\}$$

is a discrete operators and $T_{\bar{t}}$ is compact for some $\bar{t} > 0$, the semigroup has very finite-dimensional range.

The following theorem is proved in [24]. See also [17].

THEOREM 7.1. *Take X, U, Y (real) Banach spaces. Let be given $B \in \mathcal{L}(U, X)$, $C \in \mathcal{L}(X, Y)$ and $\{T_t \in \mathcal{L}(X) | t \geqq 0\}$ a C^0 semigroup with generator A. Assume*

(i) *A is a discrete operator,*

(ii) *$T_{\bar{t}}$ is compact for some $\bar{t} > 0$.*

Taking P as above, we have that

(A, B) *is stabilizable is and only if*

$$\dim\{P\mathcal{B} \oplus AP\mathcal{B} \oplus \cdots \oplus A^{(d-1)}P\mathcal{B}\} = d.$$

(C, A) *is detectable if and only if*

$$\bigcap_{j=0}^{d-1} \text{null space}\{C|_{\mathcal{M}^+}A^j|_{\mathcal{M}^+}\} = \{0\}.$$

Here $\mathcal{B}$ denotes range $\{B\}$, C^* denotes range $\{C^*\}$ and d is the dimension of the range of P, $\mathcal{M}^+$.

The theorem gives conditions for stabilizability in terms of properties of the restriction of A to a finite dimensional invariant subspace and of a projection of the range of the operator B onto this subspace (likewise dual results characterizing detectability). Verification of the conditions amounts to checking algebraic properties of certain matrices, on computing bases for the range of the projection operator P, and its adjoint P^* (see [24] for details).

We take up a class of semigroups to which the results of this section are applicable in the next section.

8. Filtering for linear stochastic differential delay equations. Consider the stochastic differential delay equation

$$\begin{aligned} dx_t &= L(\tilde{x}_t)\,dt + B\,dw_t, \\ x_0 &\text{ given,} \end{aligned} \tag{8.1}$$

$$L(\tilde{x}_t) = \sum_{i=0}^{k} A_i \begin{Bmatrix} x_{t+\theta_i} & t+\theta_i \geqq 0 \\ h_{t+\theta_i} & t+\theta_i < 0 \end{Bmatrix} + \int_{-b}^{0} A_\theta \begin{Bmatrix} x_{t+\theta} & t-\theta \geqq 0 \\ h_{t+\theta} & t+\theta < 0 \end{Bmatrix} d\theta,$$

$$-b \leqq \theta_k < \theta_{k-1} < \cdots < \theta_0 = 0 \quad \text{for } b > 0,$$

with solution x_t, an $\mathbb{R}^n$-valued process. Here w_t is an m-dimensional separable Wiener process with constant incremental covariance; $A_0, \cdots, A_k$, B are matrices of appropriate dimension; $\theta \mapsto A_\theta$ is an essentially bounded, measurable, matrix valued function; $(x_0; h)$ is an $\mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n)$-valued Gaussian random variable independent of future increments of w_t.

Associated with this equation is a stochastic evolution equation

$$d\tilde{x}_t = A\tilde{x}_t + \tilde{B}\,dw_t$$

$$\tilde{x}_0 = (x_0, h)$$

which defines an $\mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n)$-valued process $\tilde{x}_t$. Here A generates a C^0 semigroup T_t and $\tilde{B}$ is computed from B (see [23]). Writing $\tilde{x}_t$ as $(x_t; x_t(\alpha), -b \leqq \alpha \geqq 0)$ we know [23] that x_t is the solution of (8.1).

Taking an observation process

$$dz_t = C\tilde{x}_t\,dt + F\,dv_t$$

as in § 4 (X is now understood as $\mathbb{R}^n \times L^2(-b, 0; \mathbb{R}^n)$), we may define a filtering problem as before. Thus we seek to characterize

$$\hat{\tilde{x}_t} = E\{\tilde{x}_t | z_\tau, 0 \leqq \tau \leqq t\}.$$

In this application however, the random variable of primary interest is not so much $\hat{\tilde{x}}_t$ as $\hat{x}_t$, the random variable obtained by projecting the range of $\hat{\tilde{x}}_t$ on $\mathbb{R}^n \times \{0\}$; indeed, in view of the representation theorem [23] relating $\tilde{x}_t$, x_t, $\hat{x}_t = E\{x_t | z_\tau, 0 \leqq \tau \leqq t\}$.

We make two important observations; on these observations, in fact, the significance of the results of this paper is judged to rest.

Firstly, the semigroup satisfies the conditions of Theorem 7.1 [22]. Thus the filter stability criteria of § 4 reduce to statements about properties of certain matrices (though computation of sets of basis elements for the finite dimensional subspaces involved is by no means a trivial task).

Secondly, the strongest results of the main stability theorem, Theorem 5.1, in a sense apply here. We make this precise; it is natural in the context of this application to define stability in relation to the $\mathbb{R}^n$-valued process $\hat{x}_t - x_t$ rather than the $\mathbb{R}^n \times L^2$-valued process $\hat{\tilde{x}}_t - \tilde{x}_t$. That is to say, we take the Kalman filter to be stable when the distributions on $\mathbb{R}^n$ induced by $\hat{x}_t - x_t$ converge (pointwise at continuity points of the limiting distribution) to a unique measure, whatever the "initial condition".

But for $f: \mathbb{R}^n \to \mathbb{R}$ a continuous function, $f \circ S: \mathbb{R}^n \times L^2 \to \mathbb{R}$ is *weakly continuous* (for S the projection $\mathbb{R}^n \times L^2 \to \mathbb{R}^n \times \{0\}$). It is not difficult to see that, in consequence, the conditions in Theorem 5.1 which assure filter stability in relation to $\tilde{x}_t$ (w.r.t. the weak $\mathbb{R}^n \times L^2$ topology) also assure filter stability in relation to x_t *in the usual sense.*

Appendix A. Perturbed evolution operators. Take X, (real) Hilbert space.

DEFINITION A.1. $T_{t,s}: \mathscr{P} \to \mathscr{L}(X)$ is a *mild evolution operator* when

(a) $T_{t,s}: \mathscr{P} \to \mathscr{L}(X)$ is strongly continuous, and

(b) $T_{t,r}T_{r,s} = T_{t,s}$, $T_{t,t} = I$ (identity in $\mathscr{L}(X)$), $t \geqq r \geqq s$.

Here,

$$\mathscr{P} = \{(t, s) \in \mathbb{R}^2 | t \geqq s \geqq 0\}.$$

We shall of course view a C^0 semigroup T_t as a mild evolution operator defined through $(t, s) \mapsto T_{t-s}$.

DEFINITION A.2. Let $T_{t,s}: \mathscr{P} \to \mathscr{L}(X)$ be a mild evolution operator. Suppose that $B_t: [0, \infty) \to \mathscr{L}(X)$ is strongly measurable and locally essentially bounded. Then the mild evolution operator $\tilde{T}_{t,s}$ is referred to as $T_{t,s}$ *perturbed by* B_t when, for all $t \geqq s \geqq 0$, $x \in X$,

$$\tilde{T}_{t,s}x = T_{t,s}x + \int_s^t T_{t,\sigma}B_\sigma \tilde{T}_{\sigma,s}x \, ds. \tag{A.1}$$

It is known that for given $T_{t,s}$, B_t, (A.1) uniquely defines the perturbed evolution operator $\tilde{T}_{t,s}$ (see [21]).

Appendix B. Separable Hilbert space valued random variables. For definitions relating to separable Hilbert space valued random variables we refer to [1]. Note that the separability assumption assures that we do not have to distinguish between weak and strong measurability, independence, etc. Expectation is written $E\{\cdot\}$.

We adopt the definition of a Wiener process taking values in a separable Hilbert space given in [1, p. 168], limiting attention however to processes where the incremental covariance is constant.

Take X, U (real) separable Hilbert spaces. In order that (4.3) be meaningful, the definition of Wiener integral

$$\int_0^t B_\sigma \, dw\sigma$$

(w_t a U-valued Wiener process on $[0, t]$, $\sigma \mapsto B_\sigma$ a weakly measurable, essentially bounded $\mathscr{L}(U, X)$-valued function) here employed differs slightly from that in [1, p. 180 et seq.], where attention is limited to integrands measurable with respect to the *uniform* operator topology. For the rather obvious modification involved, we refer to [23].

Conditioning of a first order separable Hilbert space valued random variable on a sub-sigma field follows the definition of the (strong) conditional expectation given, for example, in [20, p. 356]; definition of conditional expectation of one first order separable Hilbert space valued random variable on another is clear. Note that if x, y are random variables taking values in the separable Hilbert spaces X, Y respectively, with x second order, then x conditioned on y, written

$$E\{x|y\},$$

coincides with $\omega \mapsto \hat{x}(y(\omega))$, where $\hat{x}$ is the unique element in $L^2(\mathscr{B}, \sigma; X)$ such that

$$E\{\langle \hat{x}(z(\omega)), g(y(\omega))\rangle\} = E\{\langle x(\omega), g(y(\omega))\rangle\}$$
$$\text{for all } g \in L^2(\mathscr{B}, \sigma; X).$$

Here, $\mathscr{B}$ are the Borel sets of Y, and σ, the probability measure induced on $\mathscr{B}$ by y.

Appendix C. Proofs of results in § 3.

Proof of Theorem 3.1. (i) *Existence of a solution.* We readily deduce from results in [2] that there exists a unique P_t: $(-\infty, 0] \to \mathscr{L}(X)$ such that

(a) $P_t = P_t^*$, all $t \leqq 0$,

(b) $t \mapsto P_t$ is weakly continuous, and

$$\langle P_t h, h\rangle = \langle \tilde{T}_{0,t} h, G\tilde{T}_{0,t} h\rangle + \int_t^0 \langle \tilde{T}_{\sigma,t} h, [C^*C + P_\sigma BR^{-1}B^*P_\sigma]\tilde{T}_{\sigma,t} h\rangle \, d\sigma \tag{C.1}$$

(where $\tilde{T}_{t,s}$ is T_t perturbed by $-BR^{-1}B^*P_\sigma$), all $t \leqq 0$, $h \in X$.

Let us suppose that P_t is *strongly* continuous on $(-\infty, 0]$. For $h \in \mathscr{D}\{A\}$ we may differentiate the right-hand side of (C.1), as is justified for example in [21] to give, for each $h \in \mathscr{D}\{A\}$,

$$\frac{d}{dt}\langle P_t h, h\rangle = -2\langle Ah, P_t h\rangle - \langle [C^*C - P_t BR^{-1}B^*P_t]h, h\rangle, \qquad \text{all } t \leqq 0. \tag{C.2}$$

Now the right-hand side is locally integrable in view of the strong continuity of P_t. Since $\langle P_t h, h\rangle$ is everywhere differentiable it follows from [19, p. 168] that $\langle P_t h, h\rangle$ is locally absolutely continuous. Thus it remains to prove that P_t is strongly continuous.

To this end we show first that $\tilde{T}^*_{t,s}\colon \mathscr{P} \to \mathscr{L}(X)$ is strongly continuous. Indeed T^*_t is a C^0 semigroup with generator A^* (this readily follows from [12, Thm. 10.6.3., p. 324]). Choose $(t, s) \in \mathscr{P}$ and $t_1 > t$. Using this property and Fubini's theorem one may show that $(\tau, \sigma) \mapsto \tilde{T}^*_{t_1-\sigma,t,-\tau}$ is $\tilde{T}^*_t$ perturbed by $t \mapsto -BR^{-1}B^*P_{t_1-t}$. In view of the definition of mild evolution operators therefore, $\tilde{T}^*_{t,s}$ is in particular strongly continuous at (t, s).

Next we note that for each $h \in X$,

$$\sigma \mapsto \tilde{T}^*_{\sigma,t}(C^*C + P_\sigma BR^{-1}B^*P_\sigma)\tilde{T}_{\sigma,t}h$$

is locally essentially bounded and (strongly) measurable. Since $P_t = P^*_t$, (C.1) may equivalently be written

$$P_t h = \tilde{T}^*_{0,t}G\tilde{T}_{0,t}h + \int_t^0 \tilde{T}^*_{\sigma,t}(C^*C + P_\sigma BR^{-1}B^*P_\sigma)T_{\sigma,t}h\, d\sigma,$$

$$\text{all } t \leqq 0, \quad h \in X.$$

Using dominated convergence and the strong continuity of $\tilde{T}^*_{t,s}$, $\tilde{T}_{t,s}$ one can now easily deduce that P_t is strongly continuous.

(ii) *Uniqueness of solutions.* Let P_t satisfy the hypotheses of the theorem. For $t \leqq 0$ define $Q_t \in \mathscr{L}(X)$ by $Q_t = Q^*_t$, and

$$\langle Q_t h, h\rangle = \langle \tilde{T}_{0,t}h, G\tilde{T}_{0,t}h\rangle + \int_t^0 \langle \tilde{T}_{\sigma,t}h, (W + P_\sigma BR^{-1}B^*P_\sigma)\tilde{T}_{\sigma,t}h\rangle\, d\sigma,$$

all $h \in X$, $t \leqq 0$, (where $\tilde{T}_{t,s}$ is T_t perturbed by $-BR^{-1}B^*P_t$). Using the strong continuity of P_t we may show as in the uniqueness part of the proof that Q_t is strongly continuous and that, for $h \in \mathscr{D}\{A\}$, $t \mapsto \langle Q_t h, h\rangle$ is everywhere differentiable with

$$\frac{d}{dt}\langle Q_t h, h\rangle = -2\langle Q_t h, (A - BR^{-1}B^*P_t)h\rangle - \langle (C^*C + P_t BR^{-1}B^*P_t)h, h\rangle.$$

Let us write $\Psi_t = P_t - Q_t$. Then Ψ_t is strongly continuous, $\Psi^*_t = \Psi_t$, all $t \leqq 0$ and, for $h \in \mathscr{D}\{A\}$, $t \mapsto \langle \Psi_t h, h\rangle$ is locally absolutely continuous on $(-\infty, 0]$ with

$$\text{(C.3)} \qquad \begin{aligned} &\frac{d}{dt}\langle \Psi_t h, h\rangle = -2\langle (A - BR^{-1}B^*P_t)h, \Psi_t h\rangle, \quad \text{all } t \leqq 0, \\ &\Psi_0 = 0. \end{aligned}$$

Now suppose that we can show that the only solution Ψ_t to (C.3) is the trivial solution (in the class of strongly continuous, self-adjoint functions such that $\langle \Psi_t h, h\rangle$ is locally absolutely continuous for each $h \in \mathscr{D}\{A\}$); then it will follow that $P_t = Q_t$ for all t. In other words P_t satisfies (C.1). But (C.1) admits a unique weakly

continuous solution, whence (3.1) admits at most one solution, which is what we set out to prove.

So turning to (C.3), fix $s<0$ and define $\sum_t = T^*_{t-s}\Psi_t T_{t-s}$ on $[s, 0]$. Using the strong continuity and differentiability properties of Ψ_t together with the properties of the C^0 semigroup T_t, we easily verify that $\langle \sum_t h, h\rangle$ is *everywhere* differentiable on $[s, 0]$ for $h \in \mathscr{D}\{A\}$ with

$$\frac{d}{dt}\langle \sum_t h, h\rangle = 2\langle \Psi_t h, BR^{-1}B^*P_t h\rangle.$$

By [19, p. 168] then (recall that $\sum_0 = 0$)

$$\langle \Sigma_s h, h\rangle = \langle \Psi_s h, h\rangle = -\int_s^0 \langle \Psi_t h, BR^{-1}B^*P_t h\rangle\, dt, \qquad \text{all } h \in \mathscr{D}\{A\}.$$

But $\|P_t\|$ is bounded on bounded intervals, so that given $t<0$ there exists a constant K_t depending on t such that

$$\|\Psi_s\| \leqq K_t \int_s^0 \|\Psi_\sigma\|\, ds \quad \text{on } [t, 0]$$

(we have used the density of $\mathscr{D}\{A\}$ in X). By Gronwall's lemma then, the strongly continuous function $\|\Psi_s\|$ is identically zero on the arbitrary interval $[t, 0]$ and the theorem is proved.

We now turn to proof of Theorem 3.2, which will require the following lemma due to Zabczyk:

LEMMA C.1. *Take A, B, C, R as in § 3. Suppose that (C, A) is detectable and that there exist $Q \in \mathscr{L}(X)$ with $Q = Q^*$, $Q \geqq 0$ and some $K \in \mathscr{L}(X, U)$ such that*

$$\langle (Q(A-BK)+C^*C+K^*RK)h, h\rangle \leqq 0, \quad \textit{all } x \in \mathscr{D}\{A\}. \tag{C.4}$$

Then (A − BK) *generates an exponentially stable semigroup.*

Proof (details in [26]). For clarity we write T^G_t for the C^0 semigroup on X with generator G, and in particular we write T^A_t for T_t. The detectability assumption gives existence of $S \in \mathscr{L}(Y,X)$ such that T^{A-SC}_t is exponentially stable. The proof depends on noting that $(A-BK)$ can be written $(A-SC)+(SC-BK)$ whence the perturbation formula gives

$$T^{A-BK}_t x = T^{A-SC}_t x + \int_0^t T^{A-SC}_{t-s}(SC-BK)T^{A-BK}_s x\, ds. \tag{C.5}$$

We may deduce from (C.4) and the invertibility of R that for arbitrary $x \in X$,

$$\begin{aligned} &\int_0^\infty \|CT^{A-Bk}_s x\|^2\, ds < \infty, \\ &\int_0^\infty \|KT^{A-BK}_s x\|^2\, ds < \infty. \end{aligned} \tag{C.6}$$

Equations (C.5), (C.6) with some standard estimates give that

$$\int_0^\infty \|T_t^{A-BK}x\|^2\,dt<\infty, \quad \text{all } x\in X,$$

from which it may be deduced [5] that T_t^{A-BK} is exponentially stable.

Proof of Theorem 3.2. (i) Let P_t^0 be the solution to (3.1) for $G=0$. Then optimality considerations (see (3.4)) and the stabilizability assumption give that P_t^0 is monotone nonincreasing and bounded on $(-\infty, 0]$ in the class of self-adjoint, bounded linear maps $X\to X$, with respect to the natural partial ordering. It follows [18, p. 263] that P_t^0 has a strong limit $\bar{P}$ as $t\to-\infty$. For each $h\in\mathscr{D}\{A\}$ then

$$\begin{aligned}\frac{d}{dt}\langle P_t^0h, h\rangle &= -2\langle Ah, P_t^0h\rangle\\ &\quad -\langle (CC^*-P_t^0BR^{-1}B^*P_t^0)h, h\rangle\\ &\to 2\langle Ah, \bar{P}h\rangle-\langle (CC^*-\bar{P}BR^{-1}B^*\bar{P})h, h\rangle\\ &= k\end{aligned}$$

for some constant k. We argue that k must take value zero. Indeed

$$\begin{aligned}k &= \lim_{t\to-\infty}\int_{t-1}^t \frac{d}{d\tau}\langle P_\tau h, h\rangle\,d\tau\\ &= \lim_{t\to-\infty}\langle (P(t)-P(t-1))h, h\rangle = 0.\end{aligned}$$

We may therefore take P^∞ as $\bar{P}$ and (i) is proved.

(ii) Suppose that P, Q both satisfy (3.5). Under the detectability assumption, as a special case of Lemma C.1, we have that $(A-BK)$ generates an exponentially stable semigroup T_t^{A-BK} where $K=R^{-1}B^*Q$. Write $K_0=R^{-1}B^*P$ and take note of the identity

$$\begin{aligned}&\langle (2P(A-BK_0)+K_0^*RK_0+C^*C)h, h\rangle\\ &\quad=\langle (2Q(A-BK)+K^*RK+C^*C)h, h\rangle+2\langle (P-Q)(A-BK)h, h\rangle\\ &\qquad+\langle (K-K_0)^*R(K-K_0)h, h\rangle, \qquad \text{all } h\in\mathscr{D}\{A\}.\end{aligned}$$

But P, Q both satisfy (3.5) by assumption whence

$$\langle 2(P-Q)(A-BK)h, h\rangle+\langle (K-K_0)^*R(K-K_0)h, h\rangle=0 \qquad \text{all } h\in\mathscr{D}\{A\}. \tag{C.7}$$

Recalling that T_t^{A-BK} is exponentially stable, for $h\in\mathscr{D}\{A\}$

$$\begin{aligned}&\langle (P-Q)h, h\rangle\\ &\quad=-\int_0^\infty \frac{d}{d\tau}\{\langle T_\tau^{A-BK}h, (P-Q)T_\tau^{A-BK}h\rangle\}\,d\tau\\ &\quad=-2\int_0^\infty \langle T_t^{A-BK}h, (P-Q)(A-BK)T_t^{A-BK}h\rangle\}\,d\tau\end{aligned}$$

which is nonnegative by (C.7) and the assumption on R. Since $\mathscr{D}\{A\}$ is dense, we conclude that $P \geqq Q$. Likewise we prove that $Q \geqq P$ whence, by the properties of partial orderings $P = Q$.

(iii) Under the stabilizability and detectability assumptions, we have shown that

$$P_t^0 \to P^\infty \quad \text{(strongly)}$$

with P^∞ the unique solution to (3.5) in the specified class, and $A - BK^\infty$ (where $K^\infty = BR^{-1}B^*P^\infty$) generates an exponentially stable semigroup, $T_t^{A-BK^\infty}$.

Define $\Sigma_t \in \mathscr{L}(X)$, $\Sigma_t = \Sigma_t^*$, for each $t \leqq 0$, by

$$\langle \Sigma_t h, h\rangle = \int_0^{-t} \langle T_\sigma^{A-BK^\infty} h, (W + P^\infty BR^{-1}B^*P^\infty) T_\sigma^{A-BK^\infty} h\rangle \, d\sigma.$$

It is not difficult to show (cf. [6]) that, for $h \in X$,

$$\langle (\Sigma_t - P_t^0)h, h\rangle \to 0 \quad \text{as } t \to -\infty. \tag{C.8}$$

Optimality considerations (see (3.4)) now give

$$\langle \Sigma_t h, h\rangle + \langle T_{-t}^{A-BK^\infty} h, GT_{-t}^{A-BK^\infty} h\rangle \geqq \langle P_t h, h\rangle \geqq \langle P_t^0 h, h\rangle, \qquad h \in X,$$

where P_t is the solution to (3.1). It follows, for $h \in X$,

$$\begin{aligned} 0 &\leqq \langle (P_t - P_t^0)h, h\rangle \\ &\leqq \langle (\Sigma_t - P_t^0)h, h\rangle + \langle T_{-t}^{A-BK^\infty} h, GT_t^{A-BK^\infty} h\rangle. \end{aligned}$$

Define $\Delta P_t = P_t - P_t^0$. We see that $\Delta P_t \in \mathscr{L}(X)$, $\Delta P_t = \Delta P_t^*, \Delta P_t \geqq 0$ all $t \leqq 0$; further that

$$\langle \Delta P_t h, h\rangle \to 0 \quad \text{as } t \to -\infty, \qquad h \in X, \tag{C.9}$$

by (C.8) and the exponential stability of $T_t^{A-BK^\infty}$. We now use a standard trick to deduce that in fact

$$\Delta P_t \to 0 \quad \text{(strongly) as } t \to -\infty. \tag{C.10}$$

By the generalized Schwarz inequality [18, p. 262], for $h \in X$,

$$\|\Delta P_t\|^2 \leqq \langle \Delta P_t h, h\rangle \cdot \langle \Delta P_t^2 h, \Delta P_t h\rangle.$$

But $\|\Delta P_t\|$ is uniformly bounded on $(-\infty, 0]$; (C.10) now follows from (C.9). We know that $P_t^0 \to P^\infty$ strongly. It follows that $P_t \to P^\infty$ strongly, and the theorem is proved.

Proof of Proposition 3.1. Recall that A generates a C^0 semigroup T_t if and only if A^* generates a C^0 semigroup T_t^*. As an immediate consequence we have that (A, B) is stabilizable if and only if (B^*, A^*) is detectable and that (C, A) is detectable if and only if (A^*, C^*) is stabilizable. Bearing these properties in mind, all except for the final assertion of the proposition follow from the previous results, on consideration of a change of independent variable $t \mapsto -t$. We omit the details.

We now establish (3.8). Finite dimensionality of range $\{C\}$ assures that $C^*N^{-1}C: X \to X$ has representation

$$(C^*N^{-1}C)x = \sum_{i=1}^{k} \langle c_i, x\rangle b_i, \quad \text{all } x \in X,$$

for some integer k, some set $\{b_i, c_i\}_{i=1}^k$ in X. It follows that

$$\text{(C.11)} \qquad \|(P_tC^*N^{-1}C - P^\infty C^*N^{-1}C)\| \leqq \sum_{i=1}^{k} \|(P_t - P^\infty)c_i\| \cdot \|b_i\|.$$

Under the stabilizability and detectability hypotheses, however, $P_t \to P^\infty$ (strongly) as $t \to \infty$. We conclude from (C.11) that

$$P_tC^*N^{-1}C \to P^\infty C^*N^{-1}C \quad \text{(in the uniform topology)}.$$

Write K_t for $P_tC^*N^{-1}C$, K^∞ for $P^\infty C^*N^{-1}C$. By a basic property of evolution operators, the semigroup $T_t^{A-K^\infty}$ generated by $A-K^\infty$ can be expressed as $\tilde{T}_{t,s}$ perturbed by $K_t - K^\infty$ ($\tilde{T}_{t,s}$ as defined in the proposition statement). Thus

$$\begin{aligned}(\tilde{T}_{t+\delta,t} - T_\delta^{A-K^\infty})x &= -\int_t^{t+\delta} \tilde{T}_{t+\delta,\sigma}(K_\sigma - K^\infty)T_{\sigma-t}^{A-K^\infty} x \, d\sigma \\ &= \int_0^{\delta} \tilde{T}_{t+\delta,t+\delta}(K_{\sigma+t} - K^\infty)T_\sigma^{A-K^\infty} x \, d\sigma.\end{aligned}$$

Standard estimates give existence of a nondecreasing function k_δ such that

$$\sup_{0\leqq\sigma\leqq\delta, t\geqq 0} \|\tilde{T}_{t+\delta,t+\sigma}\| < k_\delta, \qquad \delta \geqq 0.$$

We may also arrange that $k_\delta > \|T_\delta^{A-K^\infty}\|$. By the foregoing, $A^* - (K^\infty)^*$ generates an exponentially stable semigroup, whence $T_t^{A-K^\infty}$ is exponentially stable. There exist therefore $m, \alpha > 0$ with $\|T_\delta^{A-K^\infty}\| \leqq m \cdot e^{-\alpha\delta}$ for $\delta \geqq 0$. It follows that

$$\|\tilde{T}_{t+\delta,t}\| \leqq k_\delta^2, \; \delta \cdot \sup_{\tau\geqq t} \|K_\tau - K^\infty\| + m \cdot c^{-\alpha\delta}.$$

Clearly we may choose $\delta, t > 0$ so that

$$\|\tilde{T}_{\tau+\delta,\tau}\| < \beta < 1, \quad \text{all } \tau \geqq t.$$

But,

$$\tilde{T}_{\tau,0} = (\tilde{T}_{\tau,t+n_\tau\delta})\left(\prod_{j=0}^{n_\tau - 1} \tilde{T}_{t+(j+1)\delta,t+j\delta}\right)\tilde{T}_{t,0}$$

where n_τ is taken to be the largest nonnegative integer n such that $\tau - t > n\delta$. In consequence,

$$\|\tilde{T}_{\tau,0}\| \leqq k_\delta \beta^{n_\tau} \cdot \|T(t,0)\| \to 0 \quad \text{as } \tau \to \infty.$$

This is the desired result.

Acknowledgments. The author wishes to thank J. M. C. Clark and J. Zabczyk for helpful discussions.

REFERENCES

[1] A. BENSOUSSAN, *Filtrage Optimale des Systemes Lineaires*, Dunod, Paris, 1971.

[2] A. BENSOUSSAN, M. C. DELFOUR AND S. K. MITTER, *Notes on infinite dimensional systems*, monograph, to appear.

[3] R. F. CURTAIN, *Filtering for stochastic evolution equations*, Tech. Rep., Warwick Univ., Coventry, England, 1975.

[4] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Riccati equation for systems defined by evolution operators*, Tech. Rep., Warwick Univ., Coventry, England, 1975.

[5] R. DATKO, *Extending a Theorem of A. M. Liapunov to Hilbert Space*, J. Math. Anal. Appl., 32 (1970), pp. 610–616.

[6] ———, *A linear control problem in abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[7] M. C. DELFOUR, C. MCCALLA AND S. K. MITTER, *Stability and the infinite time quadratic cost problem for linear hereditary differential systems*, this Journal, 13 (1975), pp. 48–88.

[8] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Part 1, Interscience, New York, 1957.

[9] ———, *Linear Operators*, Part 3, Interscience, New York, 1973.

[10] I. I. GIKHMAN AND A. V. SKOROKHOD, *Introduction to the Theory of Random Processes*, W. B. Saunders, Philadelphia, 1969.

[11] U. GRENANDER, *Probabilities on Algebraic Structures*, John Wiley, New York, 1966.

[12] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, No. 31, American Mathematical Society, Providence, 1957.

[13] R. E. KALMAN AND R. S. BUCY, *New results in linear filtering and prediction theory*, Trans. ASME, J. Basis Engrg., 83 (1961), pp. 95–107.

[14] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

[15] S. K. MITTER AND R. B. VINTER, *Filtering for Linear Stochastic Hereditary Differential Systems*, International Symposium on Control Theory, Numerical Methods and Computer Systems Modelling, IRIA (Institute de Recherche d'Informatique et d'Antomatique), Rocquencourt, France, June 1974.

[16] K. R. PARTHASARATHY, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

[17] A. J. PRITCHARD AND R. TRIGGIANI, *Modal control and stabilizability of control systems in Banach space*, Tech. Rep., Warwick Univ., Coventry, England, 1975.

[18] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis*, Frederick Ungar, New York, 1955.

[19] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[20] F. S. SCALORA, *Abstract martingale converge theorems*, Pacific J. Math., 11 (1961), pp. 347–374.

[21] R. B. VINTER, *Some results concerning perturbed evolution equations with applications to delay systems*, Tech. Rep. 74/62, Imperial College, London, England, 1974.

[22] ———, *On the evolution of the state of linear differential delay equations in M^2: Properties of the generator*, J. Inst. Math. Appl., to appear.

[23] ———, *A representation of solutions to stochastic delay equations*, Appl. Math. and Optimization, to appear.

[24] ———, *Stabilizability and semigroups*, J. Inst. Math. Appl., to appear.

[25] W. M. WONHAM, *Linear Multivariable Control*, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1974.

[26] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optimization, to appear.

[27] L. SCHWARTZ, *Radon Measures on Arbitrary Topological Spaces and Cylinder Measures*, Oxford University Press, Oxford, England, 1973.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

THE LINEAR MULTIVARIABLE REGULATOR PROBLEM*

BRUCE A. FRANCIS†

Abstract. The problem is considered of regulating in the face of parameter uncertainty the output of a linear time-invariant system subjected to disturbance and reference signals. This problem has been solved by other researchers. In this paper a new and simpler algebraic solution is given.

1. Introduction. This paper deals with the regulation of the linear multivariable system modeled by the equations

$$\dot{x}_1 = A_1x_1 + A_3x_2 + B_1u, \tag{1}$$

$$\dot{x}_2 = A_2x_2, \tag{2}$$

$$y = C_1x_1 + C_2x_2, \tag{3}$$

$$z = D_1x_1 + D_2x_2. \tag{4}$$

Here x_1 is the plant state vector, u the control input, x_2 the vector of exogenous signals, y the vector of measurements available for control, and z the output to be regulated. The vectors u, x_1, x_2, y, and z belong to fixed finite-dimensional real linear spaces

$$\mathcal{U}, \mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}, \mathcal{Z} \tag{5}$$

respectively, and the time-invariant linear maps in (1) to (4) are defined on the appropriate spaces as follows:

$$A_i: \mathcal{X}_i \to \mathcal{X}_i, \quad C_i: \mathcal{X}_i \to \mathcal{Y}, \quad D_i: \mathcal{X}_i \to \mathcal{Z}, \qquad i = 1, 2,$$

$$A_3: \mathcal{X}_2 \to \mathcal{X}_1, \quad B_1: \mathcal{U} \to \mathcal{X}_1.$$

The vector A_3x_2 in (1) represents a plant disturbance, and the vector D_2x_2 in (4) represents a reference signal which the plant output $-D_1x_1$ is required to track. Equation (2) then models the class of disturbance and reference signals (e.g., steps, ramps, sinusoids).

Control action is to be provided by a compensator processing the measurements $y(\cdot)$, generating the control $u(\cdot)$, and modeled by

$$\dot{x}_c = A_cx_c + B_cy, \tag{6}$$

$$u = F_cx_c + G_cy. \tag{7}$$

Here the compensator state vector x_c belongs to a finite-dimensional real linear space $\mathcal{X}_c$, and the linear maps A_c, B_c, F_c, G_c are time-invariant. It is convenient to consider a compensator formally as a 5-tuple

$$(\mathcal{X}_c, A_c, B_c, F_c, G_c),$$

* Received by the editors November 26, 1975, and in revised form April 19, 1976.

† University Engineering Department, Control and Management Systems Division, Cambridge University, Cambridge CB2 1RX, England. This research was conducted while the author held a postdoctorate fellowship from the National Research Council of Canada.

where

$$A_c: \mathscr{X}_c \to \mathscr{X}_c, \qquad B_c: \mathscr{Y} \to \mathscr{X}_c,$$
$$F_c: \mathscr{X}_c \to \mathscr{U}, \qquad G_c: \mathscr{Y} \to \mathscr{U}.$$

There are two control objectives: closed loop stability and output regulation. Consider a fixed compensator $(\mathscr{X}_c, A_c, B_c, F_c, G_c)$, and define the closed loop state vector, state space, and linear maps

$$x_L = \begin{bmatrix} x_1 \\ x_c \end{bmatrix}, \qquad \mathscr{X}_L = \mathscr{X}_1 \oplus \mathscr{X}_c, \tag{8a}$$

$$A_L = \begin{bmatrix} A_1 + B_1 G_c C_1 & B_1 F_c \\ B_c C_1 & A_c \end{bmatrix}: \mathscr{X}_L \to \mathscr{X}_L, \tag{8b}$$

$$B_L = \begin{bmatrix} A_3 + B_1 G_c C_2 \\ B_c C_2 \end{bmatrix}: \mathscr{X}_2 \to \mathscr{X}_L, \tag{8c}$$

$$D_L = [D_1 0]: \mathscr{X}_L \to \mathscr{Z}. \tag{8d}$$

From (1), (3), (4), (6), and (7), the closed loop is described by

$$\dot{x}_L = A_L x_L + B_L x_2, \tag{9}$$

$$z = D_L x_L + D_2 x_2. \tag{10}$$

Closed loop stability means that A_L is stable, that is, $\sigma(A_L) \subset \mathbb{C}^-$, and *output regulation* means that $z(t) \to 0$ as $t \to \infty$ for all $x_L(0)$ and $x_2(0)$. The compensator is called a *synthesis* if it provides closed loop stability and output regulation.

The spaces (5) are assumed to have fixed bases, so we regard $A_1, A_3, \cdots$ in (1) to (4) as linear maps or as real matrices, depending on the context. Similarly, in specifying a compensator, we shall suppose that a basis for $\mathscr{X}_c$ is specified, so we regard A_c, B_c, F_c, G_c also as real matrices.

Now consider a fixed synthesis $(\mathscr{X}_c, A_c, B_c, F_c, G_c)$. An n-dimensional *data point* $p \in \mathbb{R}^n$ is a list of n numbers selected from among the elements of the plant matrices A_1, A_3, B_1 together with the compensator matrices A_c, B_c, F_c, G_c. A property of points in $\mathbb{R}^n$ is said to be *stable at* p if it holds throughout some open neighborhood of p. We say that the synthesis is *structurally stable at* p if closed loop stability and output regulation are both properties which are stable at p. Clearly closed loop stability is a stable property (if A_L is stable it remains so under small perturbation), so the synthesis is structurally stable at p iff output regulation is a stable property at p. The requirement of structural stability evidently reflects an uncertainty of some system parameters or the desire to achieve a degree of insensitivity to slow drift in certain parameters.

Our object in this paper is to solve two problems.

PROBLEM 1. Find computable necessary and sufficient conditions (in terms of the given data $A_1, A_3, B_1, A_2, C_1, C_2, D_1, D_2$) for the existence of a synthesis. Give an algorithm to compute a synthesis when these conditions hold.

By a computable condition we mean one for which a verifying algorithm exists.

PROBLEM 2. This is Problem 1 with "synthesis" replaced by "structurally stable synthesis."

These or similar problems have been treated by many researchers, among whom we mention Bhattacharyya et al. [1], [2], Davison [3], Davison and Goldenberg [4], Grasselli [5], Johnson [6], [7], Müller and Lückel [8], Pearson et al. [9], Sebakhy and Wonham [10], Smith and Davison [11], Wonham and Pearson [12], Wonham [13], and Young and Willems [14]. In our view the algebraic solutions presented in this paper are simpler than previous solutions. With the exception of some technical facts, the treatment given here is self-contained.

2. Technical preliminaries. Notation $\mathbb{R}$ (resp. $\mathbb{C}$) denotes the field of real (resp. complex) numbers. $\mathbb{C}^+$ (resp. $\mathbb{C}^-$) is the closed right-half (resp. open left-half) complex plane. We use the standard notation of linear algebra: if $A: \mathscr{X} \to \mathscr{X}$ is a linear transformation (map, for short), Im A is its image, Ker A its kernel, $\sigma(A)$ its complex spectrum, and $A \mid \mathscr{V}$ is the restriction of A to $\mathscr{V}$. The dimension of $\mathscr{X}$ is denoted by $d(\mathscr{X})$. For linear spaces $\mathscr{R}$ and $\mathscr{S}$, $\mathscr{R} \cong \mathscr{S}$ means $\mathscr{R}$ and $\mathscr{S}$ are isomorphic and Hom $(\mathscr{R}, \mathscr{S})$ is the linear space of all maps $\mathscr{R} \to \mathscr{S}$. For maps M and N, $M \cong N$ means M and N are similar ($M = T^{-1}NT$ for some isomorphism T). While any linear space $\mathscr{X}$ is initially real, we shall introduce without comment its complexification. For example if $A: \mathscr{X} \to \mathscr{X}$ and $\lambda \in \sigma(A) \subset \mathbb{C}$, then Ker $(A - \lambda)$ is a complex subspace of the complexification of $\mathscr{X}$. $\mathbb{R}[s]$ (resp. $\mathbb{C}[s]$) is the ring of polynomials in s with coefficients in $\mathbb{R}$ (resp. $\mathbb{C}$). For polynomials $\alpha(s)$ and $\beta(s)$, $\alpha \mid \beta$ means α divides β. We abbreviate degree to deg and greatest common divisor to gcd. Finally, for $n \geqq 1$, $\underline{n}$ is the set $\{1, \cdots, n\}$.

We now recall some characterizations of stabilizability and detectability. For this consider a triple (C, A, B):

$$C: \mathscr{X} \to \mathscr{Y}, \quad A: \mathscr{X} \to \mathscr{X}, \quad B: \mathscr{U} \to \mathscr{X}.$$

Let

$$\mathscr{N} = \bigcap_{i \geqq 0} \text{Ker } (CA^i)$$

be the unobservable subspace of (C, A),

$$\langle A | \text{Im } B \rangle = \sum_{i \geqq 0} A^i \text{ Im } B$$

the controllable subspace of (A, B), and $\mathscr{X}^+(A)$ the unstable subspace of A. (See [13].) *Then the pair* (C, A) is detectable iff

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

or equivalently,

$$\text{Ker } C \cap \text{Ker } (A - \lambda) = 0, \qquad \lambda \in \mathbb{C}^+;$$

the pair (A, B) is stabilizable iff

$$\mathscr{X}^+(A) \subset \langle A | \text{Im } B \rangle,$$

or equivalently,

$$\mathscr{X} = \operatorname{Im}(A-\lambda)+\operatorname{Im} B, \qquad \lambda \in \mathbb{C}^+.$$

Throughout this paper the following are *standing assumptions*:

(11) $$\sigma(A_2) \subset \mathbb{C}^+,$$

(12) $$\operatorname{Im} C_1 + \operatorname{Im} C_2 = \mathscr{Y},$$

(13) $$\operatorname{Im} D_1 = \mathscr{Z},$$

(14) (A_1, B_1) is stabilizable,

(15) (C_1, A_1) is detectable.

Assumption (11) involves no loss of generality, for any stable exogenous modes can be included in the plant description as they affect neither closed loop stability nor output regulation. Assumption (12) also involves no loss of generality, for if (12) does not hold initially we may redefine $\mathscr{Y}$ to be $\operatorname{Im} C_1 + \operatorname{Im} C_2$. Similarly we may assume that

(16) $$\operatorname{Im} D_1 + \operatorname{Im} D_2 = \mathscr{Z}.$$

But a necessary condition for output regulation is clearly

(17) $$\operatorname{Im} D_2 \subset \operatorname{Im} D_1;$$

so (13) follows from (16) and (17). Finally, we claim that (14) and (15) are necessary for closed loop stability. Indeed, if A_L is stable, then

$$\mathscr{X}_L = \operatorname{Im}(A_L - \lambda), \qquad \lambda \in \mathbb{C}^+;$$

hence in particular, from (8b),

$$\mathscr{X}_1 = \operatorname{Im}(A_1 + B_1 G_c C_1 - \lambda) + \operatorname{Im}(B_1 F_c), \qquad \lambda \in \mathbb{C}^+,$$

and

$$\operatorname{Ker}(B_c C_1) \cap \operatorname{Ker}(A_1 + B_1 G_c C_1 - \lambda) = 0, \qquad \lambda \in \mathbb{C}^+.$$

These conditions imply respectively

$$\mathscr{X}_1 = \operatorname{Im}(A_1 - \lambda) + \operatorname{Im} B_1 \qquad (\lambda \in \mathbb{C}^+),$$

$$\operatorname{Ker} C_1 \cap \operatorname{Ker}(A_1 - \lambda) = 0 \qquad (\lambda \in \mathbb{C}^+),$$

which are equivalent to (14) and (15). To summarize, then, (11) to (15) either involve no loss of generality or are necessary for the existence of a synthesis.

We next introduce the mathematical setting in which we shall solve Problems 1 and 2. For any linear space $\mathscr{R}$, bring in the linear space

$$\underline{\mathscr{R}} = \operatorname{Hom}(\mathscr{X}_2, \mathscr{R}).$$

For any map $A: \mathscr{X} \to \mathscr{X}$, define the linear map $\underline{A}: \underline{\mathscr{X}} \to \underline{\mathscr{X}}$ by

$$\underline{A}X = AX - XA_2, \qquad X \in \underline{\mathscr{X}}.$$

Finally, for any map $C: \mathscr{X} \to \mathscr{Y}$, where $\mathscr{X}$ and $\mathscr{Y}$ are distinct, define the linear map

$\underline{C}: \underline{\mathscr{X}} \to \underline{\mathscr{Y}}$ by

$$\underline{C}X = CX, \qquad X \in \underline{\mathscr{X}}.$$

As an application of this notation we have the following very useful characterization of output regulation due to W. M. Wonham.

LEMMA 1. *Suppose A_L is stable. Then the output z in the system*

$$\dot{x}_L = A_L x_L + B_L x_2,$$
$$\dot{x}_2 = A_2 x_2,$$
$$z = D_L x_L + D_2 x_2$$

is regulated iff

$$\underline{D}_L \underline{A}_L^{-1} B_L = D_2, \tag{18}$$

or equivalently,

$$\begin{bmatrix} B_L \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} \underline{A}_L \\ \underline{D}_L \end{bmatrix}. \tag{19}$$

By (19) we mean of course that

$$B_L = \underline{A}_L X_L, \qquad D_2 = \underline{D}_L X_L$$

for some $X_L \in \underline{\mathscr{X}}_L$.

The closed loop transfer matrix in the above system is

$$D_L(s - A_L)^{-1} B_L + D_2.$$

If A_L is stable, output regulation is therefore equivalent to the condition

$$\lim_{s \to 0} s[D_L(s - A_L)^{-1} B_L + D_2](s - A_2)^{-1} = 0. \tag{20}$$

Thus conditions (18) and (20) are equivalent. The conciseness of (18) shows the power of the present algebraic approach. Notice that for constant exogenous signals, that is $A_2 = 0$, (18) and (20) both become

$$D_L A_L^{-1} B_L = D_2.$$

Lemma 1 is a restatement of Lemma 1 of [15]. We reprove it here for completeness.

Proof of Lemma 1. Since $\sigma(A_L) \cap \sigma(A_2) = \varnothing$, $\underline{A}_L$ is invertible. Hence we may define

$$X_L = \underline{A}_L^{-1} B_L \tag{21}$$

and

$$\tilde{x}_L = x_L + X_L x_2.$$

Then an equivalent system description is

$$\begin{aligned} \dot{\tilde{x}}_L &= A_L x_L + (B_L + X_L A_2) x_2 \\ &= A_L \tilde{x}_L, \\ \dot{x}_2 &= A_2 x_2, \\ z &= D_L \tilde{x}_L + (D_2 - D_L X_L) x_2. \end{aligned} \tag{22}$$

In (22) we used (21). It is now apparent that output regulation is equivalent to the condition $D_2 - D_L X_L = 0$, which in turn is equivalent to (18). □

Using (8) in (19) we obtain immediately:

COROLLARY. *A compensator* $(\mathscr{X}_c, A_c, B_c, F_c, G_c)$ *which provides closed loop stability also provides output regulation iff*

$$\begin{bmatrix} A_3 + B_1 G_c C_2 \\ B_c C_2 \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} \underline{A}_1 + \underline{B}_1 \underline{G}_c \underline{C}_1 & \underline{B}_1 \underline{F}_c \\ \underline{B}_c \underline{C}_1 & \underline{A}_c \\ \underline{D}_1 & 0 \end{bmatrix}. \tag{23}$$

We remark that if $y = z$, which is to say

$$\mathscr{Y} = \mathscr{Z}, \quad C_1 = D_1, \quad C_2 = D_2,$$

then (23) reduces to

$$\begin{bmatrix} A_3 \\ 0 \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} \underline{A}_1 & \underline{B}_1 \underline{F}_c \\ 0 & \underline{A}_c \\ \underline{D}_1 & 0 \end{bmatrix}. \tag{24}$$

As our final technical preliminary we condense the system description (1) to (4) by defining

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \qquad \mathscr{X} = \mathscr{X}_1 \oplus \mathscr{X}_2,$$

$$A = \begin{bmatrix} A_1 & A_3 \\ 0 & A_2 \end{bmatrix} : \mathscr{X} \to \mathscr{X}, \qquad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix} : \mathscr{U} \to \mathscr{X},$$

$$C = [C_1 \quad C_2] : \mathscr{X} \to \mathscr{Y}, \qquad D = [D_1 \quad D_2] : \mathscr{X} \to \mathscr{Z}.$$

Then (1) to (4) become

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

3. Solution of Problem 1. Before solving Problem 1 we pose a simpler problem; namely, we consider pure gain controllers of the form

$$u = F_1 x_1 + F_2 x_2 \tag{25}$$

instead of dynamic compensators. Substituting (25) into (1) and rewriting (2) and (4) we obtain

$$\dot{x}_1 = (A_1 + B_1 F_1) x_1 + (A_3 + B_1 F_2) x_2, \tag{26a}$$

$$\dot{x}_2 = A_2 x_2, \tag{26b}$$

$$z = D_1 x_1 + D_2 x_2. \tag{26c}$$

PROBLEM 0. Find necessary and sufficient conditions for the existence of $F_1: \mathscr{X}_1 \to \mathscr{U}$ and $F_2: \mathscr{X}_2 \to \mathscr{U}$ so that $A_1 + B_1F_1$ is stable and the output z in (26) is regulated.

We shall call such a pair (F_1, F_2) a *pure gain synthesis.* Problem 0 is easily solved as follows.

PROPOSITION 1 (Solution of Problem 0). *A pure gain synthesis exists iff*

$$\begin{bmatrix} A_3 \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} A_1 & B_1 \\ D_1 & 0 \end{bmatrix}. \tag{27}$$

Proof. Necessity. Let (F_1, F_2) be a pure gain synthesis. Applying Lemma 1 with

$$A_L = A_1 + B_1F_1, \quad B_L = A_3 + B_1F_2, \quad D_L = D_1, \tag{28}$$

we find that

$$\begin{bmatrix} A_3 + B_1F_2 \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} A_1 + B_1F_1 \\ D_1 \end{bmatrix}, \tag{29}$$

and hence

$$\begin{bmatrix} A_3 + B_1F_2 \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} A_1 & B_1 \\ D_1 & 0 \end{bmatrix},$$

which clearly implies (27).

Sufficiency. We assume that (27) holds and shall construct suitable F_1 and F_2. First, select F_1 so that $A_1 + B_1F_1$ is stable. From (27) then

$$\begin{bmatrix} A_3 \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} A_1 + B_1F_1 & B_1 \\ D_1 & 0 \end{bmatrix},$$

and hence F_2 exists such that (29) holds. Using (28) and (29) we conclude from Lemma 1 that (F_1, F_2) provides output regulation. □

We now proceed to solve Problem 1; however, we shall make an additional assumption, namely,

$$(C, A) \quad \text{is detectable.} \tag{30}$$

This can be justified in the following manner. Let

$$\mathscr{N} = \bigcap_{i \geq 0} \operatorname{Ker}(CA^i)$$

be the unobservable subspace of (C, A) and $\mathscr{X}^+(A)$ the unstable subspace of A. Since (C_1, A_1) is detectable, the undetectable subspace $\mathscr{N} \cap \mathscr{X}^+(A)$ of the pair (C, A) is independent of $\mathscr{X}_1$:

$$\mathscr{N} \cap \mathscr{X}^+(A) \cap \mathscr{X}_1 = 0.$$

Hence we may decompose $\mathscr{X}$ as

$$\mathscr{X} = \mathscr{X}_1 \oplus \bar{\mathscr{X}}_2 \oplus \tilde{\mathscr{X}}_2,$$

where $\tilde{\mathscr{X}}_2 = \mathscr{N} \cap \mathscr{X}^+(A)$ and $\bar{\mathscr{X}}_2$ is any complement of $\tilde{\mathscr{X}}_2$ in $\mathscr{X}_2$. Corresponding to

this decomposition of $\mathscr{X}$, A, B, C, and D have representations of the form

$$\begin{bmatrix} A_1 & \bar{A}_3 & 0 \\ 0 & \bar{A}_2 & 0 \\ 0 & R & \tilde{A}_2 \end{bmatrix} \begin{bmatrix} B_1 \\ 0 \\ 0 \end{bmatrix},$$

$$[C_1 \quad \bar{C}_2 \quad 0] \quad [D_1 \quad \bar{D}_2 \quad \tilde{D}_2]$$

respectively. Here the pair

$$\left([C_1 \bar{C}_2], \begin{bmatrix} A_1 & \bar{A}_3 \\ 0 & \bar{A}_2 \end{bmatrix}\right)$$

is detectable and

$$A_2 \cong \begin{bmatrix} \bar{A}_2 & 0 \\ R & \tilde{A}_2 \end{bmatrix}.$$

These representations correspond to the system

(31a) $$\dot{x}_1 = A_1 x_1 + \bar{A}_3 \bar{x}_2 + B_1 u,$$

(31b) $$\dot{\bar{x}}_2 = \bar{A}_2 \bar{x}_2,$$

(31c) $$\dot{\tilde{x}}_2 = \tilde{A}_2 \tilde{x}_2 + R \bar{x}_2,$$

(31d) $$y = C_1 x_1 + \bar{C}_2 \bar{x}_2,$$

(31e) $$z = D_1 x_1 + \bar{D}_2 \bar{x}_2 + \tilde{D}_2 \tilde{x}_2.$$

It is readily apparent from (31) that a necessary condition for output regulation is $\tilde{D}_2 = 0$; that is,

(32) $$\mathscr{N} \cap \mathscr{X}^+(A) \subset \operatorname{Ker} D.$$

Conversely, if (32) is assumed, then in (31) $\tilde{x}_2$ is a superfluous exogenous signal: it is decoupled from the plant, the measurements y, and the output z.

To summarize, (32) is necessary for the existence of a synthesis, so we assume (32). The undetectability of (C, A) corresponds to a redundant description of the exogenous signal, so we assume (30). Since (30) trivially implies (32), we need only assume (30).

THEOREM 1 (Solution of Problem 1). *Assume* (30). *A synthesis exists iff*

(27bis) $$\begin{bmatrix} A_3 \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} A_1 & B_1 \\ D_1 & 0 \end{bmatrix}.$$

We observe that a synthesis exists iff a pure gain synthesis exists. For the system at hand, (27) apparently corresponds to the "steady-state invertibility condition" of [3] and to the "decomposability condition" of [12]. The proof of Theorem 1 is in three parts: first we prove necessity of (27), then present a synthesis algorithm, and finally show that the algorithm does indeed yield a synthesis.

Proof of Theorem 1. *Necessity.* If $(\mathscr{X}_c, A_c, B_c, F_c, G_c)$ is a synthesis, then by the Corollary to Lemma 1, (23) holds. Hence, in particular,

$$\begin{bmatrix} A_3+B_1G_cC_2 \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} \underline{A}_1+\underline{B}_1G_c\underline{C}_1 & \underline{B}_1F_c \\ \underline{D}_1 & 0 \end{bmatrix}$$

$$\subset \operatorname{Im} \begin{bmatrix} \underline{A}_1 & \underline{B}_1 \\ \underline{D}_1 & 0 \end{bmatrix}.$$

This implies (27). □

In view of assumption (30), an obvious synthesis procedure is the following: Use an observer to generate an estimate $x_c = \begin{bmatrix} x_{c1} \\ x_{c2} \end{bmatrix}$ of the state $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ of the system

$$\dot{x} = Ax + Bu, \qquad y = Cx.$$

Then apply the control $u = F_1x_{c1} + F_2x_{c2}$, where (F_1, F_2) is a pure gain synthesis. This is accomplished by the

SYNTHESIS ALGORITHM (SA).

Step 1. Let $\mathscr{X}_c = \mathscr{X}$ and select $B_c: \mathscr{Y} \to \mathscr{X}_c$ so that $A - B_cC$ is stable.

Step 2. Select $F_1: \mathscr{X}_1 \to \mathscr{U}$ so that $A_1 + B_1F_1$ is stable.

Step 3. Select $F_2: \mathscr{X}_2 \to \mathscr{U}$ so that

$$\begin{bmatrix} A_3+B_1F_2 \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} \underline{A}_1+\underline{B}_1\underline{F}_1 \\ \underline{D}_1 \end{bmatrix}. \tag{29bis}$$

Step 4. Set $F_c = [F_1, F_2]$, $A_c = A - B_cC + BF_c$, $G_c = 0$.

Sufficiency. Obviously Steps 1 and 2 of SA are possible, and if (27) holds we can choose F_2 to satisfy (29) just as we did in the proof of Proposition 1. So it remains to show that $(\mathscr{X}_c, A_c, B_c, F_c, F_c, 0)$ is a synthesis.

Writing

$$B_c = \begin{bmatrix} B_{c1} \\ B_{c2} \end{bmatrix}: \mathscr{Y} \to \mathscr{X}_1 \oplus \mathscr{X}_2,$$

we have

$$A_c = \begin{bmatrix} A_1 - B_{c1}C_1 + B_1F_1 & A_3 - B_{c1}C_2 + B_1F_2 \\ -B_{c2}C_1 & A_2 - B_{c2}C_2 \end{bmatrix}. \tag{33}$$

Hence

$$A_L = \begin{bmatrix} A_1 & B_1F_c \\ B_cC_1 & A_c \end{bmatrix}$$

$$\cong \begin{bmatrix} A_1+B_1F_1 & B_1F_1 & B_1F_2 \\ 0 & A_1-B_{c1}C_1 & A_3-B_{c1}C_2 \\ 0 & -B_{c2}C_1 & A_2-B_{c2}C_2 \end{bmatrix}$$

$$= \begin{bmatrix} A_1+B_1F_1 & B_1F_c \\ 0 & A-B_cC \end{bmatrix};$$

thus A_L is stable.

To show that output regulation holds, let

$$X_1 = (\underline{A}_1 + \underline{B}_1\underline{F}_1)^{-1}(A_3 + B_1F_2) \in \underline{\mathscr{X}}_1$$

and

$$X_c = \begin{bmatrix} X_1 \\ -I \end{bmatrix} \in \underline{\mathscr{X}}_c. \tag{34}$$

It is easily checked using (29) and (33) that

$$A_3 = \underline{A}_1X_1 + \underline{B}_1\underline{F}_cX_c, \tag{35a}$$

$$B_cC_2 = \underline{B}_c\underline{C}_1X_1 + \underline{A}_cX_c, \tag{35b}$$

$$D_2 = \underline{D}_1X_1. \tag{35c}$$

Thus (23) holds. Output regulation now follows from the Corollary to Lemma 1. □

A synthesis as computed by SA employs a full order dynamic observer of the state x. Such a synthesis may be inefficient in the sense of employing more integrators than is necessary. A reduced order synthesis may be obtained by using either a minimal order observer of the state x or a minimal order observer of Fx, where $F = [F_1F_2]$ is a pure gain synthesis.

A synthesis procedure of the latter type (see [16] and [17]) amounts to choosing $\mathscr{X}_c$ of minimal dimension such that there exist maps

$$H: \mathscr{X}_c \to \mathscr{X}_c, \quad K: \mathscr{Y} \to \mathscr{X}_c, \quad T: \mathscr{X} \to \mathscr{X}_c$$
$$F_c: \mathscr{X}_c \to \mathscr{U}, \quad G_c: \mathscr{Y} \to \mathscr{U}$$

with the properties

$$H \text{ is stable,}$$
$$TA - HT = KC,$$
$$F_cT + G_cC = F.$$

It is routine to verify that a synthesis is then $(\mathscr{X}_c, A_c, B_c, F_c, G_c)$, where

$$A_c = H + TBF_c, \qquad B_c = K + TBG_c.$$

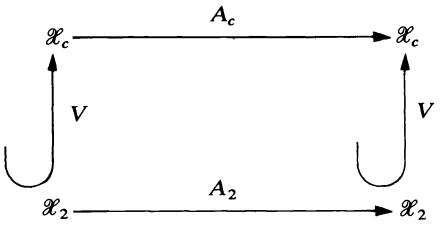
4. The structure of a feedback synthesis. We shall say that a synthesis is of *feedback type* if the compensator processes the regulated output z, that is, if $y = z$. Our object now is to point out a basic feature of a feedback synthesis as obtained by SA.

PROPOSITION 2. *Assume* (27), (30), *and* $y = z$, *and consider a synthesis* $(\mathscr{X}_c, A_c, B_c, F_c, G_c)$ *obtained by* SA. *There is a monomorphism*[1] $V: \mathscr{X}_2 \to \mathscr{X}_c$ *such that the*

[1] A monomorphism is an injective morphism, i.e. a one-to-one linear transformation.

following diagram commutes:

(36)



The interpretation of (36) is that A_c incorporates a copy of A_2; precisely,

$$A_c | \text{Im } V \cong A_2.$$

The use of a copy of A_2 in A_c is explicit in the controllers of Johnson [7] and Davison [3].

Proof of Proposition 2. Using the notation introduced in the proof of Theorem 1 (Sufficiency), if $C_1 = D_1$ and $C_2 = D_2$, we find from (35b) and (35c) that $A_c X_c = 0$. Since X_c is injective (see (34)) it suffices to take $V = X_c$. □

The above controller feature is not a result of using SA. Indeed, every feedback synthesis has this feature.

PROPOSITION 3. *Assume* (27), (30), *and* $y = z$. *For any synthesis* $(\mathscr{X}_c, A_c, B_c, F_c, G_c)$ *there is a monomorphism* $V: \mathscr{X}_2 \to \mathscr{X}_c$ *such that* (36) *commutes.*

Proof. Let $(\mathscr{X}_c, A_c, B_c, F_c, G_c)$ be any synthesis. From the Corollary to Lemma 1 we know that (24) holds; that is, there exist $X_1 \in \underset{\sim}{\mathscr{X}}_1$ and $V \in \underset{\sim}{\mathscr{X}}_c$ such that

$$A_3 = A_1 X_1 - X_1 A_2 + B_1 F_c V, \tag{37a}$$

$$0 = A_c V - V A_2, \tag{37b}$$

$$D_2 = D_1 X_1. \tag{37c}$$

It remains to show that V is injective.

For a proof by contradiction suppose that Ker $V \neq 0$. From (37b),

$$\text{Ker } (V A_2^i) \supset \text{Ker } V, \qquad i \geqq 0;$$

hence (V, A_2) is not observable. Therefore there exist $\lambda \in \sigma(A_2)$ and $x_2 \in \mathscr{X}_2$, $x_2 \neq 0$, such that

$$V x_2 = 0, \qquad A_2 x_2 = \lambda x_2. \tag{38}$$

Set $x_1 = -X_1 x_2 \in \mathscr{X}_1$. Then from (37a) and (38),

$$(A_1 - \lambda) x_1 + A_3 x_2 = (A_1 - \lambda) x_1 + (A_1 X_1 - X_1 A_2) x_2 = 0,$$

and from (37c),

$$D_1 x_1 + D_2 x_2 = 0.$$

Consequently,

$$0 \neq \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \text{Ker } D \cap \text{Ker } (A - \lambda),$$

which contradicts (30). □

Proposition 3 is not true if the assumption $y = z$ is dropped, as the following example shows.

Example. Consider a first order stable plant whose output is to follow a step reference signal:

$$\dot{x}_1 = -x_1 + u,$$
$$\dot{x}_2 = 0,$$
$$z = -x_1 + x_2.$$

Suppose the reference signal is available for measurement:

$$y = x_2.$$

It is easily checked that (27) and (30) hold. A synthesis is obtained by the feedforward control $u = y$: a dynamic compensator is not necessary. The closed loop transfer function is $s/(s+1)$. Thus the feedforward connection has, without dynamics, provided the necessary closed loop zero at $s = 0$ to cancel the reference signal pole at $s = 0$. Synthesis by feedforward is considered more generally by Davison [18].

5. Solution of Problem 2. Problem 2 is solved by Theorems 2a and 2b.

THEOREM 2a. *A synthesis which is structurally stable at A_3 exists only if*

$$\mathscr{X}_1 = \underline{A}_1 \operatorname{Ker} \underline{D}_1 + \operatorname{Im} \underline{B}_1. \tag{39}$$

It is not difficult to show that (39) is equivalent to the condition

$$\mathscr{X}_1 = (A_1 - \lambda) \operatorname{Ker} D_1 + \operatorname{Im} B_1, \qquad \lambda \in \sigma(A_2),$$

which in turn is equivalent to

$$\operatorname{Im} \begin{bmatrix} A_1 - \lambda & B_1 \\ D_1 & 0 \end{bmatrix} = \mathscr{X}_1 \oplus \mathscr{Z}, \qquad \lambda \in \sigma(A_2).$$

This latter condition is the one which arises in the work of Davison and Goldenberg [4] and Wonham [13].

Proof of Theorem 2a. If $(\mathscr{X}_c, A_c, B_c, F_c, G_c)$ is a synthesis which is structurally stable at A_3 then, by the Corollary to Lemma 1, (23) is a property which is stable at A_3. From (23) we have

$$\begin{bmatrix} A_3 + B_1 G_c C_2 \\ D_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} \underline{A}_1 + \underline{B}_1 G_c \underline{C}_1 & \underline{B}_1 F_c \\ \underline{D}_1 & 0 \end{bmatrix} \subset \operatorname{Im} \begin{bmatrix} \underline{A}_1 & \underline{B}_1 \\ \underline{D}_1 & 0 \end{bmatrix}. \tag{40}$$

Letting $D_1^\dagger : \mathscr{Z} \to \mathscr{X}_1$ be any right inverse of D_1, we find from (40) that

$$\begin{bmatrix} A_3 + B_1 G_c C_2 - \underline{A}_1 (D_1^\dagger D_2) \\ 0 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} \underline{A}_1 & \underline{B}_1 \\ \underline{D}_1 & 0 \end{bmatrix};$$

equivalently,

$$A_3 + B_1 G_c C_2 - \underline{A}_1 (D_1^\dagger D_2) \in \underline{A}_1 \operatorname{Ker} \underline{D}_1 + \operatorname{Im} \underline{B}_1. \tag{41}$$

Now clearly (41) is a property which is stable at $A_3 \in \mathscr{L}_1$ only if (39) holds. □

Our object now is to prove a converse of Theorem 2a; that is, to show that (39) is a sufficient condition. For this, however, we need an additional assumption. Recall from [15] the definition that *z is readable from y* if there is a map $Q: \mathscr{Y} \to \mathscr{Z}$ such that $z = Qy$, which is to say $D_1 = QC_1$ and $D_2 = QC_2$. It was shown in [15] (Theorem 1) that a necessary condition for structural stability (at a suitable data point) is that z be readable from y. Hence we here assume this.

If such Q exists we can imbed $\mathscr{Z}$ in $\mathscr{Y}$: write

$$\mathscr{Y} = \mathscr{W} \oplus \mathscr{Z}$$

for a suitable linear space $\mathscr{W}$. Then

$$C_1 = \begin{bmatrix} E_1 \\ D_1 \end{bmatrix}, \qquad C_2 = \begin{bmatrix} E_2 \\ D_2 \end{bmatrix}$$

for suitable maps $E_i: \mathscr{X}_i \to \mathscr{W}$ $(i = 1, 2)$, and

$$y = \begin{bmatrix} w \\ z \end{bmatrix},$$

where $w = E_1x_1 + E_2x_2 \in \mathscr{W}$. Here Q is the natural projection $\mathscr{W} \oplus \mathscr{Z} \to \mathscr{Z}$. Now for a compensator $(\mathscr{X}_c, A_c, B_c, F_c, G_c)$, define

$$B_{cw} = B_c \mid \mathscr{W}, \qquad B_{cz} = B_c \mid \mathscr{Z},$$
$$G_{cw} = G_c \mid \mathscr{W}, \qquad G_{cz} = G_c \mid \mathscr{Z}.$$

Then the overall system equations are

$$\begin{aligned} \dot{x}_1 &= A_1x_1 + A_3x_2 + B_1u, \\ \dot{x}_2 &= A_2x_2, \\ w &= E_1x_1 + E_2x_2, \\ z &= D_1x_1 + D_2x_2, \\ \dot{x}_c &= A_cx_c + B_{cw}w + B_{cz}z, \\ u &= F_cx_c + G_{cw}w + G_{cz}z. \end{aligned}$$

The compensator is now formally a 7-tuple

$$(\mathscr{X}_c, A_c, B_{cw}, B_{cz}, F_c, G_{cw}, G_{cz}).$$

THEOREM 2b. *Assume that z is readable from y and that* (39) *holds. Then there is a synthesis in which* $B_{cw} = 0$, $G_{cw} = 0$, *and* $G_{cz} = 0$ *and which is structurally stable at* $(A_1, A_3, B_1, B_{cz}, F_c)$.

Notice that the data point $(A_1, A_3, B_1, B_{cz}, F_c)$ includes the plant data (A_1, A_3, B_1) together with the nonzero compensator data excluding A_c: small arbitrary perturbations in A_c cannot be permitted if output regulation is to be maintained. Notice also that the compensator is of feedback type, processing only the output z $(B_{cw} = 0, G_{cw} = 0)$.

The format of the proof of Theorem 2b is the same as the proof of Theorem 1 (Sufficiency): first we give a synthesis procedure and then show that the resulting

compensator has the required properties. For the synthesis procedure we need some notation.

Let

$$\mathscr{X}_2 = \bigoplus_{i=1}^{k} \mathscr{X}_{2i}$$

be a rational canonical decomposition (rcd) of $\mathscr{X}_2$ relative to A_2. Thus $\mathscr{X}_{2i}$ is A_2-invariant ($i \in \underline{k}$), $A_{2i} \triangleq A_2|\mathscr{X}_{2i}$ is cyclic ($i \in \underline{k}$), the minimal polynomial (mp) of $A_{2,i+1}$ divides that of A_{2i} ($i \in \underline{k-1}$), and the mp of A_{21} is the same as that of A_2. Let $q = d(\mathscr{Z})$ and define

$$\mathscr{X}_{2e} = \mathscr{X}_{21} \oplus \cdots \oplus \mathscr{X}_{21} \quad (q\text{-fold direct sum})$$

and

$$A_{2e}: \mathscr{X}_{2e} \to \mathscr{X}_{2e}, \qquad A_{2e}|\mathscr{X}_{21} = A_{21}.$$

Thus A_{2e} is the q-fold direct sum of the largest cyclic component of A_2. Now if a is any one of the subscripts $1, \cdots, k, e$ and $\mathscr{R}$ is any linear space, define

$$\underline{\mathscr{R}}_a = \text{Hom}\,(\mathscr{X}_{2a}, \mathscr{R}).$$

Similarly if $A: \mathscr{X} \to \mathscr{X}$ define $\underline{A}_a: \underline{\mathscr{X}}_a \to \underline{\mathscr{X}}_a$ by

$$\underline{A}_a X_a = AX_a - X_a A_{2a}, \qquad X_a \in \underline{\mathscr{X}}_a,$$

and if $C: \mathscr{X} \to \mathscr{Y}$ define $\underline{C}_a: \underline{\mathscr{X}}_a \to \underline{\mathscr{Y}}_a$ by

$$\underline{C}_a X_a = CX_a, \qquad X_a \in \underline{\mathscr{X}}_a.$$

STRUCTURALLY STABLE SYNTHESIS ALGORITHM (SSSA).

Step 1. Define $\mathscr{X}_e = \mathscr{X}_1 \oplus \mathscr{X}_{2e}$ and select $A_{3e}: \mathscr{X}_{2e} \to \mathscr{X}_1$ so that (D_e, A_e) is detectable. Here

$$A_e = \begin{bmatrix} A_1 & A_{3e} \\ 0 & A_{2e} \end{bmatrix}: \mathscr{X}_e \to \mathscr{X}_e,$$

$$D_e = [D_1 \quad 0]: \mathscr{X}_e \to \mathscr{Z}.$$

The next four steps consist in obtaining a synthesis via SA for the system

$$\dot{x}_1 = A_1 x_1 + A_{3e} x_{2e} + B_1 u, \tag{42a}$$

$$\dot{x}_{2e} = A_{2e} x_{2e}, \tag{42b}$$

$$y = z = D_1 x_1. \tag{42c}$$

Step 2. Let $\mathscr{X}_c = \mathscr{X}_e$ and select $B_{cz}: \mathscr{Z} \to \mathscr{X}_c$ so that $A_e - B_{cz} D_e$ is stable.

Step 3. Select $F_1: \mathscr{X}_1 \to \mathscr{U}$ so that $A_1 + B_1 F_1$ is stable.

Step 4. Select $F_{2e}: \mathscr{X}_{2e} \to \mathscr{U}$ so that

$$\begin{bmatrix} A_{3e} + B_1 F_{2e} \\ 0 \end{bmatrix} \in \text{Im} \begin{bmatrix} \underline{A}_{1e} + \underline{B}_{1e} \underline{F}_{1e} \\ \underline{D}_{1e} \end{bmatrix}. \tag{43}$$

Step 5. Set $F_c = [F_1 F_{2e}]$, $A_c = A_e - B_c D_e + B_e F_c$, $B_{cw} = 0$, $G_{cw} = 0$, $G_{cz} = 0$.

Here

$$B_e = \begin{bmatrix} B_1 \\ 0 \end{bmatrix} : \mathcal{U} \to \mathcal{X}_e.$$

Before proceeding we require some technical facts. We first recall the notion of a generic property (see [13]). For any field K a property Π of points in K^n is *generic* if the set of points where Π fails lies in a proper algebraic variety in K^n. Suppose Π is generic on $\mathbb{C}^n$, that is, Π fails only on a proper variety in $\mathbb{C}^n$. Then Π is generic when restricted to $\mathbb{R}^n$. To see this let $\not{p} = (p_1, \cdots, p_n)$ be a representative point in $\mathbb{C}^n$. If Π is generic on $\mathbb{C}^n$, there is a nonzero polynomial

$$\phi(s_1, \cdots, s_n) \in \mathbb{C}[s_1, \cdots, s_n]$$

such that Π fails only at points $\not{p} \in \mathbb{C}^n$ where $\phi(p_1, \cdots, p_n) = 0$. Write ϕ as

$$\phi(s_1, \cdots, s_n) = \phi_1(s_1, \cdots, s_n) + i\phi_2(s_1, \cdots, s_n),$$

where $\phi_j \in \mathbb{R}[s_1, \cdots, s_n]$ $(j = 1, 2)$. Now ϕ_1 and ϕ_2 are not both identically zero; hence

$$\psi = \phi_1^2 + \phi_2^2 \in \mathbb{R}[s_1, \cdots, s_n]$$

is not identically zero. Now Π fails at $\not{p} \in \mathbb{R}^n$ only if $\psi(p_1, \cdots, p_n) = 0$. Hence Π is generic on $\mathbb{R}^n$.

Next we require

LEMMA 2. *Let $A: \mathcal{X} \to \mathcal{X}$, $\bar{A}: \bar{\mathcal{X}} \to \bar{\mathcal{X}}$ be maps with invariant factors $\alpha_i(s)$ $(i \in \underline{m})$, $\bar{\alpha}_i(s)$ $(i \in \underline{\bar{m}})$ respectively. Define L: Hom $(\bar{\mathcal{X}}, \mathcal{X}) \to$ Hom $(\bar{\mathcal{X}}, \mathcal{X})$ by*

$$LX = AX - X\bar{A}.$$

Then

(a) $d(\text{Ker } L) = \sum_{i,j} \deg \gcd(\alpha_i, \bar{\alpha}_j)$.

(b) *There exists a monomorphism $V \in \text{Ker } L$ iff $\bar{m} \leqq m$ and*

$$\bar{\alpha}_i \mid \alpha_i, \qquad i \in \underline{\bar{m}}.$$

Proof. (a) This is immediate from [19, Thm. 1, p. 219].

(b) Suppose there is a monomorphism $V: \bar{\mathcal{X}} \to \mathcal{X}$ such that $AV = V\bar{A}$. Then $\mathcal{V} = \text{Im } V$ is A-invariant and $A \mid \mathcal{V} \cong \bar{A}$. Thus the invariant factors of $A \mid \mathcal{V}$ are the same as those of $\bar{A}$. From [20, Lemma 1(i)] it now follows that $\bar{m} \leqq m$ and $\bar{\alpha}_i \mid \alpha_i$, $i \in \underline{\bar{m}}$.

Conversely, suppose $\bar{m} \leqq m$ and $\bar{\alpha}_i \mid \alpha_i$, $i \in \underline{\bar{m}}$. Let

$$\mathcal{X} = \bigoplus_{i=1}^{m} \mathcal{X}_i$$

be a rcd of $\mathcal{X}$ relative to A, and define

$$\mathcal{V}_i = \text{Ker } \bar{\alpha}_i(A \mid \mathcal{X}_i), \qquad i \in \underline{\bar{m}},$$

$$\mathcal{V} = \bigoplus_{i=1}^{\bar{m}} \mathcal{V}_i.$$

Then $\mathcal{V}$ is A-invariant and $A \mid \mathcal{V}$ has invariant factors $\bar{\alpha}_i$ $(i \in \underline{\bar{m}})$, so $A \mid \mathcal{V} \cong \bar{A}$. Thus there is an isomorphism $T: \bar{\mathcal{X}} \to \mathcal{V}$ such that

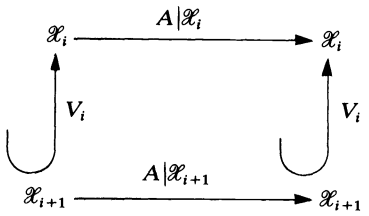
$$(A \mid \mathcal{V})T = T\bar{A}.$$

Hence it suffices to define $V: \bar{\mathcal{X}} \to \mathcal{X}$ by $V = T$ on $\bar{\mathcal{X}}$. □

As a simple application of Lemma 2b we find that if

$$\mathcal{X} = \bigoplus_{i=1}^{m} \mathcal{X}_i$$

is a rcd of $\mathcal{X}$ relative to A, then for each $i \in \underline{m-1}$ there is a monomorphism $V_i: \mathcal{X}_{i+1} \to \mathcal{X}_i$ such that the following diagram commutes:



Proof of Theorem 2b. We first show that when (39) holds, SSSA can be carried out, and second that the resulting compensator provides the required structural stability. As the proof is fairly long it is divided into four steps.

(i) We shall prove that Step 1 of SSSA is possible. Since (D_1, A_1) is detectable,

$$d[(A_1 - \lambda) \operatorname{Ker} D_1] = d(\operatorname{Ker} D_1) = d(\mathcal{X}_1) - q, \qquad \lambda \in \sigma(A_{2e}). \tag{44}$$

Furthermore (D_e, A_e) is detectable iff

$$(A_{3e}, A_{2e}) \quad \text{is observable} \tag{45}$$

and for each $\lambda \in \sigma(A_{2e})$,

$$A_{3e} \operatorname{Ker} (A_{2e} - \lambda) \cap (A_1 - \lambda) \operatorname{Ker} D_1 = 0. \tag{46}$$

Now by construction, A_{2e} has q cyclic components in a rcd, and since D_1 is surjective, $q \leqq d(\mathcal{X}_1)$. These two facts show that (45) is a generic property of A_{3e}. Similarly, (44) together with the fact

$$d[\operatorname{Ker} (A_{2e} - \lambda)] = q, \qquad \lambda \in \sigma(A_{2e}),$$

shows that (46) is a generic property of complex A_{3e} and hence of real A_{3e} for each $\lambda \in \sigma(A_{2e})$. Since the conjunction of a finite number of generic properties is generic, we find that

$$(D_e, A_e) \quad \text{detectable}$$

is a generic property of A_{3e}. Hence Step 1 of SSSA is accomplished by "almost any" $A_{3e}: \mathcal{X}_{2e} \to \mathcal{X}_1$.

(ii) Obviously Steps 2 and 3 are now possible, so we show that Step 4 is.

From (39) we have

$$\underline{\mathscr{X}}_{11} = \underline{A}_{11} \operatorname{Ker} \underline{D}_{11} + \operatorname{Im} \underline{B}_{11},$$

and hence

$$\underline{\mathscr{X}}_{1e} = \underline{A}_{1e} \operatorname{Ker} \underline{D}_{1e} + \operatorname{Im} \underline{B}_{1e};$$

equivalently,

$$\underline{\mathscr{X}}_{1e} = (\underline{A}_{1e} + \underline{B}_{1e}\underline{F}_{1e}) \operatorname{Ker} \underline{D}_{1e} + \operatorname{Im} \underline{B}_{1e}.$$

Thus there exists $F_{2e} \in \underline{\mathscr{U}}_e$ such that

$$A_{3e} + B_1 F_{2e} \in (\underline{A}_{1e} + \underline{B}_{1e}\underline{F}_{1e}) \operatorname{Ker} \underline{D}_{1e},$$

which is equivalent to (43).

We have now shown that SSSA can be carried out. Furthermore we know from the proof of Theorem 1 that

$$A_L = \begin{bmatrix} A_1 & B_1 F_c \\ B_{cz} D_1 & A_c \end{bmatrix} \tag{47}$$

is stable. So it remains to show that output regulation is a property which is stable at $(A_1, A_3, B_1, B_{cz}, F_c)$.

(iii) We claim that

$$\operatorname{Ker} \underline{B}_{cze} = 0 \tag{48}$$

and

$$\operatorname{Im} \underline{A}_{ce} \cap \operatorname{Im} \underline{B}_{cze} = 0. \tag{49}$$

To establish this, recall that $(\mathscr{X}_c, A_c, 0, B_{cz}, F_c, 0, 0)$ is a synthesis for system (42) and that (D_e, A_e) is detectable. Hence Proposition 2 implies the existence of a monomorphism $V_e: \mathscr{X}_{2e} \to \mathscr{X}_c$ such that

$$A_c V_e = V_e A_{2e}. \tag{50}$$

Since A_{2e} has exactly q invariant factors each of which is the mp of A_2, we conclude from (50) and Lemma 2b that the mp of A_2 divides at least q invariant factors of A_c.

Since A_L is stable,

$$\underline{\mathscr{X}}_{Le} = \operatorname{Im} \underline{A}_{Le}. \tag{51}$$

From (47) this implies that

$$\underline{\mathscr{X}}_{ce} = \operatorname{Im} \underline{A}_{ce} + \operatorname{Im} \underline{B}_{cze},$$

which in turn implies

$$\underline{\mathscr{X}}_{c1} = \operatorname{Im} \underline{A}_{c1} + \operatorname{Im} \underline{B}_{cz1}. \tag{52}$$

Now let $\{\alpha_{ci}(s)\}$ be the invariant factors of A_c and $\alpha_2(s)$ the mp of A_2. Applying Lemma 2a we have

$$d(\operatorname{Ker} \underline{A}_{c1}) = \sum_i \deg \gcd (\alpha_{ci}, \alpha_2).$$

Then, since α_2 divides at least q α_{ci}'s,

$$d(\operatorname{Ker} \underline{A}_{c1}) \geqq q \cdot \deg \alpha_2 = q \cdot d(\mathscr{X}_{c1}). \tag{53}$$

However

$$\begin{aligned} d(\operatorname{Im} \underline{B}_{cz1}) &\leqq d(\underline{\mathscr{Z}}_1) \\ &= d(\mathscr{Z}) \cdot d(\mathscr{X}_{21}) \\ &= q \cdot d(\mathscr{X}_{21}). \end{aligned} \tag{54}$$

So from (52), (53), and (54),

$$\operatorname{Im} \underline{B}_{cz1} \cong \underline{\mathscr{Z}}_1 \tag{55}$$

and

$$\operatorname{Im} \underline{A}_{c1} \cap \operatorname{Im} \underline{B}_{cz1} = 0. \tag{56}$$

Now (55) implies that $\operatorname{Ker} \underline{B}_{cz1} = 0$, which implies (48), and (49) follows from (56). This proves the claim.

(iv) Returning to (51), we know in view of (47) that for any $R_{1e} \in \underline{\mathscr{X}}_{1e}$, there exist $X_{1e} \in \underline{\mathscr{X}}_{1e}$ and $X_{ce} \in \underline{\mathscr{X}}_{ce}$ such that

$$\begin{aligned} R_{1e} &= \underline{A}_{1e} X_{1e} + \underline{B}_{1e} \underline{F}_{ce} X_{ce}, \\ 0 &= \underline{B}_{cze} \underline{D}_{1e} X_{1e} + \underline{A}_{ce} X_{ce}. \end{aligned}$$

But from (48) and (49) this implies

$$\mathscr{X}_{1i} = \underline{A}_{1e} \operatorname{Ker} \underline{D}_{1e} + \underline{B}_{1e} \underline{F}_{ce} \operatorname{Ker} \underline{A}_{ce}$$

from which there follows

$$\underline{\mathscr{X}}_{11} = \underline{A}_{11} \operatorname{Ker} \underline{D}_{11} + \underline{B}_{11} \underline{F}_{11} \operatorname{Ker} \underline{A}_{c1}. \tag{57}$$

Now for each $i \in \underline{k}$, A_{2i} is imbedded in A_{21} in the sense that $A_{21} V_i = V_i A_{2i}$ for some monomorphism V_i: $\mathscr{X}_{2i} \to \mathscr{X}_{21}$. A brief computation using this fact and (57) yields

$$\underline{\mathscr{X}}_{1i} = \underline{A}_{1i} \operatorname{Ker} \underline{D}_{1i} + \underline{B}_{1i} \underline{F}_{ci} \operatorname{Ker} \underline{A}_{ci}, \qquad i \in \underline{k},$$

and hence,

$$\underline{\mathscr{X}}_1 = \underline{A}_1 \operatorname{Ker} \underline{D}_1 + \underline{B}_1 \underline{F}_c \operatorname{Ker} \underline{A}_c. \tag{58}$$

We conclude the proof by showing that (24) is stable at $(A_1, A_3, B_1, B_{cz}, F_c)$. Clearly (58) is stable at this data point, so it suffices to show that (24) follows from (58).

If $D_1^\dagger$ is any right inverse of D_1, then (24) is equivalent to

$$\begin{bmatrix} A_3 - \underline{A}_1 (D_1^\dagger D_2) \\ 0 \\ 0 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} \underline{A}_1 & \underline{B}_1 \underline{F}_c \\ 0 & \underline{A}_c \\ \underline{D}_1 & 0 \end{bmatrix},$$

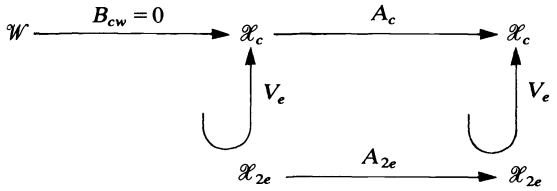
and this is equivalent to

$$A_3 - \underline{A}_1(D_1^{\dagger}D_2) \in \underline{A}_1 \operatorname{Ker} \underline{D}_1 + \underline{B}_1 \underline{F}_c \operatorname{Ker} \underline{A}_c.$$

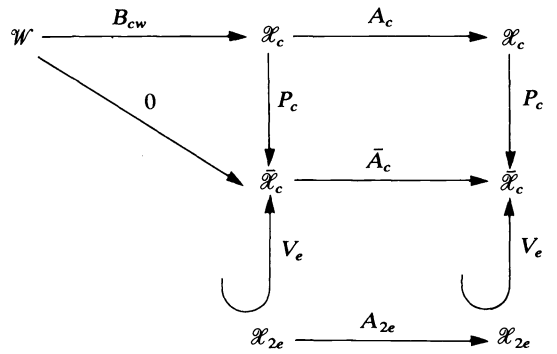
But this follows from (58). □

6. The structure of a structurally stable synthesis. We observed in § 4 that a feedback synthesis incorporates in A_c a copy of A_2. For the structurally stable feedback synthesis obtained by SSSA, a stronger statement is true: A_c incorporates a q-fold reduplication of the maximal cyclic component of A_2. More precisely, from (50) we have

PROPOSITION 4. *Assume that z is readable from y and that* (39) *holds, and consider a structurally stable synthesis computed by SSSA. There is a monomorphism $V_e: \mathscr{X}_{2e} \to \mathscr{X}_c$ such that the following diagram commutes:*



To complete the parallel of this section with § 4, we state without proof the following counterpart of Proposition 3. Assume z is readable from y and (39) holds. Let $(\mathscr{X}_c, A_c, B_{cw}, B_{cz}, F_c, G_{cw}, G_{cz})$ be a structurally stable synthesis. Then there is an A_c-invariant subspace $\mathscr{R}_c \subset \mathscr{X}_c$ and a monomorphism $V_e: \mathscr{X}_{2e} \to \bar{\mathscr{X}}_c = \mathscr{X}_c/\mathscr{R}_c$ such that the following diagram commutes:



Here P_c is the canonical projection and $\bar{A}_c$ the induced map in the factor space. We have stated this result informally, omitting the data point at which the synthesis is structurally stable. For a precise statement and proof the reader is referred to [15, Prop. 3 and Thm. 2].

7. Concluding remark. The synthesis theory presented in this paper deals with systems in state-space form. Uncertainty about the system is then taken to be uncertainty about parameters in the matrices in the state-space description. There is an implicit assumption here that the state-space description is derived from

physical laws rather than from a realization of an input–output impulse response. This is because the function (suitably defined) which maps an impulse response to its state-space realization is not continuous in the natural topologies, and hence "slight uncertainty" about the impulse response need not correspond to "slight uncertainty" about the state-space description. An important open problem therefore is a synthesis theory for systems modeled by input–output maps.

Acknowledgments. The present algebraic approach was inspired by discussions with W. M. Wonham. The criticism, expressed in the Concluding Remark, against posing the regulator problem in the state-space setting was raised by G. Zames.

REFERENCES

[1] S. P. BHATTACHARYYA AND J. B. PEARSON, *On error systems and the servomechanism problem*, Internat. J. Control, 15 (1972), no. 6, pp. 1041–1062.

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

[3] E. J. DAVISON, *The output control of linear time-invariant multi-variable systems with unmeasurable arbitrary disturbances*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 621–630.

[4] E. J. DAVISON AND A. GOLDENBERG, *The robust control of a general servomechanism problem: the servo compensator*, Automatica, 11 (1975), pp. 461–471.

[5] O. M. GRASSELLI, *Steady-state output insensitivity to step-wise disturbances and parameter variations*, Report R. 74-35, Ist. di Automatica, Universita' di Roma, Italy, 1974.

[6] C. D. JOHNSON, *Optimal control of the linear regulator with constant disturbances*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 416–421.

[7] ———, *Accommodation of external disturbances in linear regulator and servomechanism problems*, Ibid., AC-16 (1971), pp. 635–644.

[8] P. C. MÜLLER AND J. LÜCKEL, *Optimal multivariable feedback system design with disturbance rejection*, Tech. Rep., Lehrstuhl B für Mechanik, Technical University of Munich, Germany, 1975.

[9] J. B. PEARSON, R. W. SHIELDS AND P. W. STAATS, JR., *Robust solutions to linear multivariable control problems*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 508–517.

[10] O. A. SEBAKHY AND W. M. WONHAM, *A design procedure for multivariable regulators*, Proc. IFAC Third Multivariable Tech. Systems Symp., Manchester, U.K., 1974.

[11] H. W. SMITH AND E. J. DAVISON, *Design of industrial regulators: integral feedback and feedforward control*, Proc. Inst. Elec. Engrs., 119 (1972), no. 8, pp. 1210–1216.

[12] W. M. WONHAM AND J. B. PEARSON, *Regulation and internal stabilization in linear multivariable systems*, this Journal, 12 (1974), pp. 5–18.

[13] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Lecture Notes in Economics and Math. Systems, vol. 101, Springer-Verlag, New York, 1974.

[14] P. C. YOUNG AND J. C. WILLEMS, *An approach to the linear multivariable servomechanism problem*, Internat. J. Control, 15 (1972), no. 5, pp. 961–979.

[15] B. A. FRANCIS AND W. M. WONHAM, *The internal model principle for linear multivariable regulators*, J. Appl. Math. Optimization, 2 (1975), pp. 170–194.

[16] J. R. ROMAN AND T. E. BULLOCK, *Design of minimal order stable observers for linear functions of the state via realization theory*, IEEE Trans. Automatic Control, AC-20 (1975), pp. 613–622.

[17] J. B. MOORE AND G. F. LEDWICH, *Minimal order observers for estimating linear functions of a state vector*, Ibid., AC-20 (1975), pp. 623–632.

[18] E. J. DAVISON, *The feedforward control of linear multivariable time-invariant systems*, Automatica, 9 (1973), no. 5, pp. 561–573.

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

[20] J. P. CORFMAT AND A. S. MORSE, *Control of linear systems through specified input channels*, this Journal, 14 (1976), pp. 163–175.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 3, May 1977

FEEDBACK CONTROL OF PARTLY UNKNOWN SYSTEMS*

R. F. DRENICK†

Abstract. Some mathematical aspects are treated of the problem of controlling plants that are incompletely specified to the system designer. The view is taken that such plants are not defined by the usual input-output characterization, or some equivalent, but by one which assigns to every input a set of possible outputs. Their mathematical representation is then by set-valued rather than the conventional point-valued operators. Some results are accordingly derived regarding the feedback control of plants that are represented in that way. Indications are developed that a natural mode of operation for feedback controllers under the circumstances is of a kind that might be called "error-tolerant."

1. Introduction. This paper presents certain mathematical aspects of a theory for the control of systems that are partly unknown. The theory is based on the following point of view. The designer who is charged with the problem of specifying a controller for a partly unknown plant frequently will not be able to, or may not wish to, guarantee that the system will produce a certain desired response to every possible input. Instead, he may be willing to guarantee merely that the actual response lies in a certain set of possible responses, the size of the set being in some sense related to the uncertainty that surrounds the plant in the first place.

The paper (in § 2) develops the mathematical problem formulation to which one is led by this point of view, and in the sequel to certain results that follow from it. The formulation adopted here, however, is not the only one. Others based on probability theory (e.g., [6]), fuzzy sets (e.g., [2]) and minimax decision theory (e.g., [1]) have been proposed. They are surely well suited to certain situations but it is also easy to conceive of others to which they are not. In some of those, the point of view taken here may be more appropriate.

The approach to which it leads may be of interest because it draws on some mathematics which is not normally used in control theory, namely contraction mapping on spaces whose elements are subsets of other spaces. The approach may be of interest also because it seems to point to some new design principles, assuming of course, that it can be continued to the point of practicality. Indications are, roughly speaking, that controllers for partly unknown plants are most naturally designed as feedback controllers but in a way which might be called "error-tolerant." They should not, in other words, respond to every discrepancy between the actual and desired outputs but should remain inactive as long as both lie in the same "tolerance set." The choice of these sets appears to be in the nature of a design compromise in general: the smaller the set, the more involved and nonlinear the operations which the controller may have to perform. This is the upshot of §§ 3, 4 and 5 of the paper which describe, in sequence, some simple controllers with large tolerance sets, some complicated ones with tolerance sets that are points, and finally a compromise between the two.

2. Problem formulation. The usual assumption in control theory and practice is that the plant is completely known and specified, for instance, by its

* Received by the editors June 27, 1974, and in revised form August 14, 1975.

† Polytechnic Institute of New York, Brooklyn, New York 11201. This work was supported by the National Science Foundation under Grant GK-34179.

input-output characteristic. This specification is then in terms of an operator F which maps the control signal u into the output y according to

$$y = F(u). \tag{2.1}$$

The objective of the designer of the control system can then perhaps be stated as follows. He is presented with a set $\bar{X}$ of signals x and a corresponding set of desired responses $y^*(x)$, one for each x. He is to come up with a second operator G,

$$u = G(x, y), \tag{2.2}$$

which is to be chosen, if at all possible, that $y = y^*(x)$ is the unique response of the complete control system (2.1) and (2.2), or in other words, that $y^*(x)$ is the unique solution of (2.1) and (2.2), for every $x \in \bar{X}$. The device represented by G is the "controller." It is called a "feedback" or "open-loop" controller depending on whether or not u is a function of y as well as x.

In most formulations the problem is complicated by various restrictions which are placed on the choice of the controller G and which preclude the achievement of $y^*(x)$. Such complications are not considered in this paper, however. The complication that is to be considered is the incomplete specification of the plant.

A designer confronted with a partially unknown plant must in effect deal with a set of possible plants simultaneously, each representing a possible realization of the missing specifications. The members of the set can be visualized indexed by a parameter α, the "uncertainty parameter," and (2.1) accordingly replaced with

$$y = F(u; \alpha), \qquad \alpha \in \bar{A}, \tag{2.3}$$

where $\bar{A}$ is the range of α. An equivalent replacement is

$$Y = F(u; \bar{A}) \equiv \cup_{\alpha \in A} F(u; \alpha) \tag{2.4}$$

in which $F(\,:\,;\bar{A})$ is a set-valued operator representing at once all possible plant realizations, and Y the set of outputs y that could be generated by them from one control signal u. Equation (2.4) can be considered the input-output specification which is the counterpart to (2.1) for a partially unknown plant.

The objective of the designer is again the determination of an operator of the form of G in (2.2) and, if possible, in fact one that generates the desired output $y = y^*(x)$ for given x and regardless of $\alpha \in \bar{A}$. In those cases in which this objective is unachievable or impractical, the desired output might be expanded to a suitable set Y_x^*. That is, the criterion of desirability is changed to read that the control system consisting of F and G performs satisfactorily if its output y lies in Y_x^*, for given x and regardless of $\alpha \in \bar{A}$.

The question then is how the "target set" Y_x^* can be chosen and further, whether and how a controller G can be determined which assures satisfactory performance in this sense. This paper deals with this question and, more specifically, with certain somewhat unconventional mathematical aspects of it.

Such aspects are introduced by the fact that the operator in (2.4) representing the plant is not point-valued, as it is usually assumed to be, but set-valued. In other words, its domain is a function space $\bar{U}$ of signals u but its range is a space whose

elements are certain sets Y of points in another function space $\bar{Y}$. Along with operators such as F in (2.4) which map points into sets it will be necessary to consider others which carry sets into points, i.e., whose domains are spaces with elements that are subsets of others.

Such spaces are sometimes called "hyperspaces" [5]. Michael [8] has studied the hyperspace $\mathscr{C}(\bar{Y})$ of the nonempty closed compact subsets of a metric space $\bar{Y}$, and Radstroem [9] a hyperspace $\mathscr{R}(\bar{Y})$ which is constructed from the closed compact convex subsets of a normed linear space $\bar{Y}$. The results to be derived in this paper are valid in both. However, although $\mathscr{R}(\bar{Y})$ might be more suitable because it is the smaller of the two and thus perhaps closer to control practice, $\mathscr{C}(\bar{Y})$ will be used in what follows because it leads to slightly simpler proofs.

The underlying spaces $\bar{X}$, $\bar{Y}$, $\bar{U}$, and $\bar{A}$, will more specifically be assumed to be complete metric spaces, and $\bar{A}$ in particular will be assumed compact. Whenever product spaces are formed from these, the metrics on the latter will be assumed to be suitably weighted sums of the distances on the component spaces. Thus they will be convenient to use as distance on

$$
\begin{aligned}
&\bar{U}\times\bar{A}: d(u_1, u_2)+\mu_1 d(\alpha_1, \alpha_2),\\
&\bar{U}\times\bar{Y}: d(y_1, y_2)+\mu_2 d(u_1, u_2),\\
&\bar{Y}\times\bar{Y}\times\bar{A}: d(y_1, y_2)+d(z_1, z_2)+\lambda_1 d(\alpha_1, \alpha_2),\\
&\bar{Y}\times\bar{Y}\times\bar{U}: d(y_1, y_2)+d(z_1, z_2)+\lambda_2 d(u_1, u_2),
\end{aligned}
\tag{2.5}
$$

where the μ_i and λ_i are appropriate positive metric weighting constants. The choice of these constants is probably not very important in general. For the sake of consistency between (2.5b) and (2.5d), however, one presumably should set

$$\lambda_2 = 2\mu_2. \tag{2.6}$$

The hyperspace $\mathscr{C}(\cdot)$ will be assumed metrized by the Hausdorff distance. It is known [4, p. 94] that this distance is indeed a metric on the collection of all those nonempty closed subsets of a metric space Y whose Hausdorff distances are finite. But since the sets of $\mathscr{C}(\bar{Y})$ are also compact, the distance $d(Y_1, Y_2)$ between any pair is finite to begin with.

It is futher known [8] that the space $\mathscr{C}(\bar{Y})$ inherits many of the topological properties of $\bar{Y}$. In the sequel, completeness of $\mathscr{C}(\bar{Y})$ will be important. It can be shown that it inherits this property from $\bar{Y}$ as well [3, p. 61]. $\mathscr{C}(\bar{Y})$ can, incidentally, be interpreted as a linear space if $\bar{Y}$ is linear and if the operations (Y_1+Y_2) and λY_1 are, as usual, defined by

$$
\begin{aligned}
&Y_1+Y_2=\{y: y=y_1+y_2,\ y_1\in Y_1,\ y_2\in Y_2\},\\
&\mu Y_1=\{y: y=\mu y_1,\ y_1\in Y_1\}, \qquad (\mu \text{ real}).
\end{aligned}
\tag{2.7}
$$

Evidently, (Y_1+Y_2) and μY_1 lie in $\mathscr{C}(\bar{Y})$ whenever Y_1 and Y_2 do. Moreover, as Radstroem [9] has pointed out, the Hausdorff distance is a norm on $\mathscr{C}(\bar{Y})$.

The properties of operators, those mapping spaces into spaces, as well as those mapping spaces into hyperspaces, or vice versa, can be defined in terms of those metrics. This will be done below as needed.

3. Open-loop and linear feedback control. In the view taken in the preceding section, the design objective for the controller for a partially unknown plant is the production of a plant output $y(x)$ that lies in a target set Y_x^*, regardless of the value of the uncertainty parameter $\alpha \in A$. The first question that arises is how to choose this set. One such choice which is particularly undemanding will be described in this section. It will be shown that, with this choice, the controller can be designed as an open-loop controller or, if that is undesirable, as a linear feedback controller. In both cases, however, it should be designed as a device of the kind that has been called "error tolerant" in the Introduction.

The assumption to be made here regarding the target set is that it should always contain a subset Y which can be generated as in (2.4) from the partly unknown plant by some control signal $u = u^*(x)$; i.e.,

$$Y_x^* \supset Y = F(u; \bar{A}) \quad \text{for } u = u^*(x). \tag{3.1}$$

This choice might be based on the argument that, according to (2.4), any control signal u applied to such a plant will inevitably be mapped into a set Y and hence that no more can possibly be guaranteed by any controller than the placement of y into one of those sets. A target set should therefore always contain such a Y as a subset. It can of course also coincide with Y.

The argument is incorrect, as will be shown in the next sections. In fact, the choice (3.1) of the target set is quite undemanding: satisfactory control in the sense that $y(x) \in Y_x^*$ for all $\alpha \in \bar{A}$ can in principle be achieved even by an open-loop controller, namely

$$G(x) = u^*(x) = F^{-1}(Y_x^*; \bar{A}), \tag{3.2}$$

as is immediately evident from (3.1). In this equation, F^{-1} is the inverse of F in (3.1), i.e., an operator taking subsets of $\bar{Y}$ into points. If it is single-valued, it maps subsets Y_x^* of $\bar{Y}$ into points $u^*(x) \in \bar{U}$. Otherwise, it maps them into sets $U_x^* \in \bar{U}$. The intuitive interpretation of (3.1) is that, since any $y(x) \in Y_x^*$ is satisfactory, the controller need merely generate the control signal which achieves this, and this is $u^*(x)$. If there is more than one such signal, any one will do.

The controller (3.2) can be considered error-tolerant, in the sense that it produces the same signal $u^*(x)$ for all $y = Y_x^*$. This feature becomes more striking if the same performance is achieved by a feedback controller. That this is in fact possible in many cases, and even by a linear feedback controller, is shown in the following theorem.

THEOREM 3.1. *Suppose that the plant operators $F(\cdot\,; \alpha)$ obey the following boundedness conditions:*

(i) *They are bounded on $\bar{A}$, uniformly with respect to u, i.e., there exists a constant m such that*

$$d(F(u; \alpha_1), F(u; \alpha_2)) \leqq m d(\alpha_1, \alpha_2); \tag{3.3}$$

(ii) *they are furthermore bounded on* $\bar{U}\times\bar{A}$ *in the sense that*

$$\sup_{\alpha_2}\inf_{\alpha_1} d(F(u_1;\alpha_1), F(u_2;\alpha_2)) \leqq m'_F d(u_1, u_2),$$
$$\sup_{\alpha_1}\inf_{\alpha_2} d(F(u_1;\alpha_1), F(u_2;\alpha_2)) \leqq m''_F d(u_1, u_2) \tag{3.4}$$

for all $\alpha_1, \alpha_2 \in \bar{A}$.

Then there exist feedback controllers which insure that the complete control system has a unique output $y(x;\alpha)$ *for every* x *and* α *in the target set* Y_x^* *of* (3.1). *These controllers can in fact be linear in the feedback signal* y.

Proof. The method of proof is by the contraction mapping theorem. To be more specific, it is first pointed out that there exist controllers $G(x, Y)$ for which Y_x^* is a fixed element of the control system defined by

$$u = G(x, Y), \qquad Y = F(u;\bar{A}) \tag{3.5}$$

for any given x, and secondly that the system can be made contracting on $\mathscr{C}(\bar{Y})$ under the assumptions of the theorem.

For convenience, omit x and $\bar{A}$ notationally and assume $Y^* = F(u)$ for $u = u^*$, rather than (3.1). Suppose first G to have the domain $\mathscr{C}(\bar{Y})$ and to be so chosen that when $Y = Y^*$, $G(Y)$ reduces to

$$G(Y^*) = u^* = F^{-1}(Y^*). \tag{3.6}$$

Then

$$F(G(Y^*)) = Y^*, \tag{3.7}$$

showing that Y^* is a fixed element. It remains to be shown that, with suitable G, the complete control system (3.5) is contracting on $\mathscr{C}(\bar{Y})$.

To this end, one notes first that F is continuous. This follows readily from (3.3). The continuity of F, and the compactness of $\bar{A}$ assumed in § 2, together imply the compactness of the image under F of every $u \in \bar{U}$ [10, p. 63]. Hence, F maps $\bar{U}$ into $\mathscr{C}(\bar{Y})$.

It will now be shown that any G defined on $\mathscr{C}(\bar{Y})$ or, at any rate, on the range of F, and such that

$$d(u_1, u_2) = d(G(Y_1), G(Y_2)) \leqq m_G d(Y_1, Y_2) \tag{3.8a}$$

will lead to the desired contraction, provided only that the constant m_G obeys

$$m_G \leqq \rho \max(m'_F, m''_F) \equiv \rho m_F, \qquad \rho < 1. \tag{3.8b}$$

To see this, note that the Hausdorff distance between two elements Y'_1, Y'_2 in the range of F is, by (3.4),

$$d(Y'_1, Y'_2) = d(F(u_1), F(u_2)) \leqq d(u_1, u_2) \max(m'_F, m''_F).$$

But, since $\mathscr{C}(\bar{Y})$ is the domain of G, u_1 and u_2 are the images under G of two elements Y_1, $Y_2 \in \mathscr{C}(\bar{Y})$. Hence

$$d(Y'_1, Y'_2) \leqq m_F m_G d(Y_1, Y_2)$$

which shows that the complete system is contracting if G obeys (3.8). As pointed out in § 2, $\mathscr{C}(\bar{Y})$ is complete. It follows [7] that the system contracts on a unique element of $\mathscr{C}(\bar{Y})$ which is, of course, Y^*.

Controllers which satisfy this condition, and (3.5) as well, do exist. In fact,

$$G(Y)=\frac{\rho}{m_F}L(Y)-\left[F^{-1}(Y^*)+\frac{\rho}{m_F}L(Y^*)\right] \tag{3.9}$$

in which L is an arbitrary linear operator with norm $\|L\|=1$, satisfies both conditions. (L is understood to be an operator from $\mathscr{C}(\bar{Y})$ to $\bar{U}$, with $\mathscr{C}(\bar{Y})$ interpreted as a linear space.) The observation that (3.9) represents a linear controller completes the proof.

Comment 1. Conventional linear controllers differ from (3.9) in chiefly two respects. For one, they do not take sets Y into points u, as (3.9) does, but points y. In other words, they are not error-tolerant, in the terminology of this paper. For another, they do not usually include a term corresponding to $F^{-1}(Y^*)$. This omission is often defended by pointing out that the error due to it can be made small if the controller "gain" (ρ/m_F), or a similar factor, can be made large. The line of reasoning followed here, i.e., the use of contraction maps, is similar to that of Willems [12] and Zames [13] who employed such maps in conventional control problems.

4. Adaptive control. The control scheme described in the preceding section leads to a very broad class of controllers that perform satisfactorily, including even linear and open-loop controllers. One can surmise that the criterion of satisfactory operation adopted there is quite loose, and that consequently the target sets Y_x^* are unnecessarily large. This is indeed so. As will be shown in this section, these sets can in many cases be specified as single elements, namely the desired responses $y^*(x)$. The controllers which achieve this performance are of a kind that have been called "adaptive" (or "plant-adaptive," "dual," "self-adjusting," "learning," etc. [11]).

The characteristic feature of most adaptive controllers is that they achieve the desired control performance by identifying the plant at the same time. The block diagram of an adaptive control system is shown in Fig. 1. In fact, it is of fairly generic type which comprises as special cases many of those suggested at one time

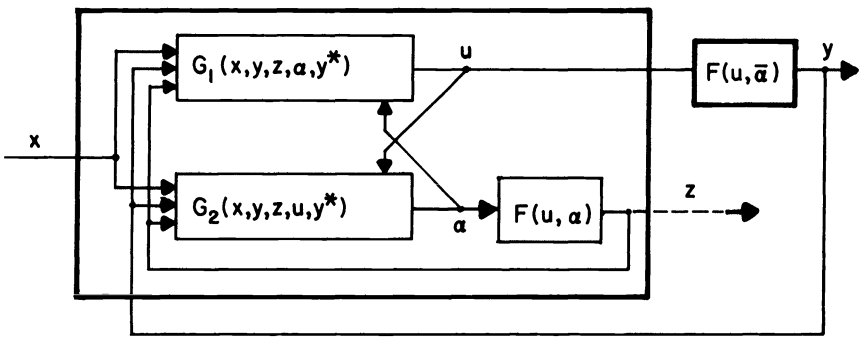


FIG 1. *An adaptive control system, with complete plant identification*

or other in the literature, to the writer's knowledge. The small block on the right, marked $F(u, \bar{\alpha})$, is the actual plant and $\bar{\alpha}$ is the unknown value of the uncertainty parameter for it. The controller G is the large block on the left. It consists of three subsystems, G_1, G_2, $F(\cdot\,; \alpha)$. The first of these produces the control signal u proper. The last, namely $F(\cdot\,; \alpha)$, is often called the "model." It is a simulated version of the plant but is characterized in place of $\bar{\alpha}$ by an adjustable parameter α. The adjustments of α are executed by the third subsystem of the controller, namely G_2, which can be called the "identifier." The output of the model is denoted with z in Fig. 1.

The intuitive idea for the operation of such a controller is to adjust α until $y = z$. One can expect that then also $\alpha = \bar{\alpha}$ at which point the actual plant has been identified. In fact, one can hope that, simultaneously with the adjustment of α, y can be so adjusted that it coincides with y^*. This intuitive idea is correct, at least under certain conditions, as will now be shown.

These remarks suggest that adaptive control might be feasible only when the plant is identifiable in the first place. Although this is not necessarily true, as will be explained later in this section, it will be good to avoid misunderstanding over terminology here and to adopt the following.

DEFINITION 4.1. A plant will be called *completely identifiable* on $(\bar{U} \times \bar{Y})$ if its operator $F(\cdot\,; \cdot)$ has a unique inverse

$$\alpha = K(u, y) \tag{4.1}$$

or, in other words, if there exists a single-valued operator K from $(\bar{U} \times \bar{Y})$ into $\bar{A}$ such that

$$y = F(u, K(u, y))$$

there. If K is not single-valued, the plant will be called *incompletely identifiable.*

The operators G_1, G_2, $F(\cdot\,; \alpha)$, and $F(\cdot\,; \bar{\alpha})$ which represent the blocks in the figure are now assumed to be of the conventional kind, namely mappings on function spaces which take points into points. The domains are product spaces, as the arguments shown in the figure imply. (However, x and $\bar{\alpha}$ can be considered fixed indices in what follows, and $y^* = y^*(x)$ could be omitted altogether because it is a function of x.)

THEOREM 4.1. *Assume that the plant is completely identifiable and that*

(i) *the operator F is bounded on $(\bar{U} \times \bar{A})$ in the sense of*

$$d(F(u_1, \alpha_1), F(u_2, \alpha_2)) \leqq m_F[d(u_1, u_2) + \mu_1 d(\alpha_1, \alpha_2)], \tag{4.2}$$

(ii) *the inverse operator K is bounded on $(\bar{U} \times \bar{Y})$ in the sense of*

$$d(K(u_1, y_1), K(u_2, y_2)) \leqq m_K[d(y_1, y_2) + \mu_2 d(u_1, u_2)]. \tag{4.3}$$

Then there exists a controller G which achieves satisfactory system performance, provided the bounds m_F and m_K also obey the condition

$$m_F m_K \mu_1 \mu_2 < \lambda_2. \tag{4.4}$$

(The μ_i and λ_2 are the metric weighting coefficients of (2.5).)

Proof. The proof proceeds along the same lines as the one of Theorem 3.1. It is first shown that the controller, or, in the present instance, the devices G_1 and G_2

can be so chosen that the point $y = y^*$, $z = z^*$ is a fixed element of the complete system, and then that they can in fact be so chosen that it is contracting. The contraction mapping theorem then implies that (y^*, y^*) is the only fixed element.

Regarding the first, assume G_1 and G_2 to have the following properties:

$$F(G_1(y, y, \alpha, y^*), \alpha) = y \quad \text{for all } \alpha \in \bar{A},\ y \in \bar{Y}, \tag{4.5}$$

$$G_2(y, y, u, y^*) = K(u, y^*) \quad \text{for all } y \in \bar{Y}. \tag{4.6}$$

Both conditions can be met. For the first, one can require analogously to Theorem 3.1 that for $y = z$,

$$G_1(y, y, \alpha, y^*) \subset F^{-1}(y; \alpha)$$

and that G_1 be linear in $(y - z)$ otherwise. The second can be met simply by choosing

$$G_2(y, z, u, y^*) = K(u, y^*). \tag{4.7}$$

Suppose now that the equation $y = z$ has been established in the system by a certain control signal $u = u^*$. Then

$$F(u^*, \alpha) = F(u^*, \bar{\alpha})$$

and because of the complete identifiability of the plant,

$$\bar{\alpha} = K(u^*, y) = \alpha.$$

In words, when actual plant and model produce the same output they are identical. But when $y = z$ and $u = u^*$, (4.6) takes the form

$$\alpha = G_2(y, y, u^*, y^*) = K(u^*, y^*).$$

It follows that

$$z = F(u^*; \alpha) = F(u^*; K(u^*, y^*)) = y^*,$$

and hence that $z = y = y^*$. The specification (4.5) for G_1, with $\alpha = \bar{\alpha}$, $y = z = y^*$, now reads

$$F(G_1(y^*, y^*, \bar{\alpha}, y^*); \bar{\alpha}) = y^*$$

and shows that $y = z = y^*$ is a fixed element of the system.

It remains to be demonstrated that the system can be made contracting by a suitable choice of G_1 and G_2, without however violating the requirements (4.5) and (4.6). It will in fact be shown that if G_1 and G_2 obey

$$\begin{aligned} d(u_1, u_2) &\leqq m_1[d(y_1, y_2) + d(z_1, z_2) + \lambda_1 d(\alpha_1, \alpha_2)], \\ d(\alpha_1, \alpha_2) &\leqq m_2[d(y_1, y_2) + d(z_1, z_2) + \lambda_2 d(u_1, u_2)], \end{aligned} \tag{4.8}$$

the bounds m_1 and m_2 can be selected in such a way that contraction is achieved. (The weighting coefficient λ_1 is the same as in (2.5).) To see this, note first that (4.8) is equivalent to

$$\begin{aligned} (1 - m_1 m_2 \lambda_1 \lambda_2) d(u_1, u_2) &\leqq m_1(1 + m_2 \lambda_1)[d(y_1, y_2) + d(z_1, z_2)], \\ (1 - m_1 m_2 \lambda_1 \lambda_2) d(\alpha_1, \alpha_2) &\leqq m_2(1 + m_1 \lambda_2)[d(y_1, y_2) + d(z_1, z_2)], \end{aligned} \tag{4.9}$$

which make sense only if

$$m_1 m_2 \lambda_1 \lambda_2 < 1. \tag{4.10}$$

This, therefore, is a first condition on m_1, m_2. Another comes from the requirement that the system must be contracting on $\bar{Y} \times \bar{Y}$. To arrive at it, note that

$$\begin{aligned} d(F(u_1, \bar{\alpha}), F(u_2, \bar{\alpha})) + d(F(u_1, \alpha_1), F(u_2, \alpha_2)) \\ \leqq m_F d(u_1, u_2) + m_F [d(u_1, u_2) + \mu_1 d(\alpha_1, \alpha_2)] \\ = 2 m_F d(u_1, u_2) + m_F \mu_1 d(\alpha_1, \alpha_2) \end{aligned}$$

for any pair (u_1, α_1) and (u_2, α_2) of elements of $\bar{U} \times \bar{A}$. Since each of these two pairs in turn is the image under (G_1, G_2) of two elements (y_1, z_1) and (y_2, z_2) of $\bar{Y} \times \bar{Y}$, one will further have, according to (4.8),

$$d(F(u_1, \bar{\alpha}), F(u_2, \bar{\alpha})) + d(F(u_1, \alpha_1), F(u_2, \alpha_2)) \leqq a[d(y_1, y_2) + d(z_1, z_2)],$$

where

$$a = (1 - m_1 m_2 \lambda_1 \lambda_2)^{-1} m_F [2 m_1 (1 + m_2 \lambda_1) + m_2 \mu_1 (1 + m_1 \lambda_2)].$$

Now, the space $\bar{Y} \times \bar{Y}$ is obviously complete since $\bar{Y}$ is complete by assumption. Contraction will therefore certainly prevail on it if

$$a < 1. \tag{4.11}$$

This is a second condition on m_1 and m_2.

A third condition is induced by requirement (4.6) according to which G_2 should coincide with K when $y = z$. Condition (4.8b) must therefore be consistent with (4.3) which will certainly be the case if

$$m_2 \lambda_2 = m_K \mu_2. \tag{4.12}$$

This constitutes a third condition. (If $\lambda_2 = 2\mu_2$, as suggested in (2.6), (4.12) is correspondingly simplified.)

It must next be shown that the conditions are consistent, i.e., that pairs of positive numbers m_1, m_2 exist which obey the three conditions (4.10), (4.11), (4.12). The first can certainly be satisfied because $m_1 m_2 \lambda_1 \lambda_2 = m_1 m_K \lambda_1 \mu_2$ can be made as small as desired by choosing m_1 sufficiently small, and (4.12) can of course be satisfied by choosing m_2 accordingly. However, no matter how small m_1, one will always have $a > m_K m_F \mu_1 \mu_2 / \lambda_2$. Therefore, in order to insure that (4.11) can be observed, one must have $m_K m_F \mu_1 \mu_2 < \lambda_2$. This however is the case, by (4.4).

The theorem is accordingly proved. Two corollaries may be useful.

The first deals with the question of whether the subsystem G_2 of the controller cannot always be assumed to be of the simpler form (4.7). The answer is that it can. In fact, the assumptions of the theorem can be relaxed at the same time.

COROLLARY 4.1. *If the identifier G_2 of the controller is*

$$G_2(x, y, z, u, y^*) = K(u, y^*),$$

the model $F(u;\alpha)$ in Fig. 1 and the signal z are superfluous. The boundedness requirements (4.7) *can be changed to*

$$d(u_1, u_2) \leqq m_1[d(y_1, y_2)+\lambda_1 d(\alpha_1, \alpha_2)],$$
$$d(\alpha_1, \alpha_2) \leqq m_K\mu_2 d(u_1, u_2)$$

in which m_1 must obey

$$m_1 m_K \lambda_1 \mu_2 < 1, \qquad m_1 m_F (1-m_1 m_K \lambda_1 \mu_2)^{-1} < 1.$$

Proof. The fact that the model and z become superfluous follows from the observation that now

$$z = F(u;\alpha) = F(u; K(u, y^*)) = y^*.$$

The fact that the boundedness conditions, together with the two inequalities for m_1 imply the contraction property for the system is then shown as in the theorem. The proof is completed by noting that those two inequalities can always be satisfied by making m_1 small enough.

The second corollary removes a restriction under which Theorem 4.1 was derived, namely the assumption that the plant be completely identifiable. Contrary perhaps to what one might expect, adaptive control can be executed also when the plant is incompletely identifiable and the desired system response y^* can be achieved. Some of the blocks in the controller of Fig. 1 must however be reinterpreted as devices of roughly the kind that have been called error-tolerant in the Introduction.

A plant that is incompletely identifiable in particular can be considered a device of exactly that kind: given some control signal u, it produces the same output y for all α in some set A (which will depend on u in general). In the notational convention that has been adopted earlier, one can therefore write

$$F(u; A) = y \tag{4.13}$$

and use the notation on the left also for the model in Fig. 1. The remaining two devices in the controller, namely G_1 and G_2 are similarly modified by writing A in place of α. The understanding regarding G_1 is more particularly the same as in (4.13). That is, G_1 is a device that produces the same signal u, for given x, y, z, and for all $\alpha \in A$. G_2, on the other hand, may produce any $\alpha \in A$ since the model responds to all in the same way. The output of G_2 can therefore also be labeled with A in place of α, as was just suggested. The same interpretation as on G_2 can be placed on the inverse of (4.13), namely the operation

$$K(u, y) = A.$$

The modification of the figure essentially carries over to Theorem 4.1 as well and leads to the following result.

COROLLARY 4.2. *Assume that the operator F is bounded in the sense of* (4.2). *Concerning K, assume that it takes on values that are nonempty closed subsets of $\bar{A}$ and that it is bounded on $(\bar{U}\times\bar{Y})$ as in* (4.3), *i.e.*,

$$d(K(u_1, y_1), K(u_2, y_2)) \leqq m_K[d(y_1, y_2)+\mu_2 d(u_1, u_2)], \tag{4.14}$$

but with the distance on the left interpreted as the Hausdorff distance between the two sets represented by $K(u_1, y_1)$ *and* $K(u_2, y_2)$. *Assume finally that the bounds* m_F *and* m_K *obey* (4.4). *Then there exist controllers* G *which achieve satisfactory system performance, provided* G_1 *and* G_2 *are reinterpreted as operators on sets, or to sets, in the manner described above.*

The proof is analogous to that of Theorem 4.1.

Comment 1. The last result shows that in principle it is always possible to construct a controller which identifies the plant as far as it needs to be identified in the first place, and which at the same time produces the described system output y^*. The controller which achieves this is of the adaptive kind. It is perhaps of interest that essentially only adaptive controllers can achieve this. This conclusion is readily reached by the following line of reasoning. Suppose that a controller G had been found which leads to the desired $y^*(x)$ for every $x \in \bar{X}$ and every $\alpha \in \bar{A}$. This would mean that it could produce a control signal $u^*(x)$ for every x such that

$$F(u^*; \alpha) = y^*, \quad \text{all } \alpha \in \bar{A}.$$

The indexing of the plant by α would thus be superfluous in this case, and no uncertainty of any kind would surround at least its desired operation. Suppose accordingly that, for some x, there exist at least two proper and distinct subsets A_1, A_2 of $\bar{A}$ such that the equation

$$F(u_1^*; A_1) = F(u_2^*; A_2) = y^*$$

holds only if $u_1^* \neq u_2^*$. But then

$$K(u_1^*, y^*) = A_1, \qquad K(u_2^*, y^*) = A_2,$$

showing that an incomplete plant identification is possible from the two (u, y)-pairs in these two equations.

Comment 2. For fixed u, F and K are each other's inverses. As one can readily establish, this implies a condition, namely

$$1 \leqq m_F m_K \mu_1 < 2$$

on the boundedness parameters m_F and m_K which may often fail in practice. One can however avoid it by choosing G_2 according to (4.7). The conditions (4.9) for contraction are then replaced by the simpler ones of Corollary 4.1.

5. "Partially" adaptive control. The control schemes described in the preceding two sections can be considered as two extremes. The specifications of the target set Y_x^* are so loose for the one of § 3 that they can in principle be met by an open-loop or a linear feedback system, while those in § 4 contract Y_x^* to single element $y^*(x)$ and hence require an adaptive (hence typically very complicated) controller for their execution. One can inquire whether or not compromises exist in which some of the complication of the latter is traded off against the looseness of the former. The object of this section is to show that this can be done. The control

schemes by which such trade-offs can be realized will be called "partially adaptive," for a reason which will become clear presently.

The objectionable operation in an adaptive controller in general is the one represented by the identifier G_2. G_1, as was pointed out after (4.6), can be chosen linear; G_2 on the other hand must reduce to $K(u, y^*)$ when $y = z$, and this is typically a rather involved nonlinear operation, even when the plant is completely identifiable. Suppose for the moment that it is in fact completely identifiable and that, as in (4.1), the parameter α is uniquely determined by $K(u, y)$. On the other hand, suppose that the realization of the operator K is impractical. One can then consider replacing it with another operation, $\hat{K}$ say, which does not determine α uniquely but only up to its membership in a set $\hat{A} \subset \bar{A}$

$$\hat{K}(u, y) = \hat{A}. \tag{5.1}$$

It follows then that

$$F(u; \hat{A}) = \bigcup_{\alpha \in \hat{A}} F(u; \alpha) \subset \bigcup_{\alpha \in \bar{A}} F(u; \alpha) = F(u; \bar{A}).$$

One can now proceed as in § 3 and designate, by anology to (3.1), a target set $\hat{Y}_x^*$ in such a way that

$$\hat{Y}_x^* \supset \hat{Y} = F(u; \hat{A}) \tag{5.2}$$

for some $u = u^*(x)$. One should however assume here that $\hat{Y}_x^* \subset Y_x^*$, with the inclusion proper, for otherwise $\hat{K}$ would achieve no effective reduction of the plant uncertainty and the control schemes of § 3 could be reverted to.

It will now be shown that controllers exist which place the system output into the target sets $\hat{Y}_x^*$, no matter how the partial identification operators $\hat{K}$ have been chosen.

The controller is in fact of the same general kind as the one shown in Fig. 1, i.e., with G_1, G_2 and a model as subsystems. The latter, however, is now represented by the operation

$$F(u; \hat{A}) = z \tag{5.3}$$

and is a device which responds, for given u, with the same z to every $\alpha \in \hat{A}$. G_1 and G_2 perform according to the equations

$$u = G_1(x, \hat{Y}, \hat{Z}, \hat{A}, \hat{Y}_x^*), \qquad \hat{A} = G_2(x, \hat{Y}, \hat{Z}, u, \hat{Y}_x^*). \tag{5.4}$$

These operations are to be interpreted in a way similar to those for adaptive control with incomplete plant identification. The subsystem G_1, for instance, receives four signals namely, x, y, z, and α. It responds to them with a control signal u and in fact with the same u to every $y \in \hat{Y}$, $z \in \hat{Z}$, and $\alpha \in \hat{A}$. G_2 receives four similar signals and responds to them with some $\alpha \in \hat{A}$. Since G_1 as well as the model respond in the same way to all $\alpha \in \hat{A}$; it does not matter which of these α G_2 generates. As the following theorem shows a controller of this kind can frequently be so chosen that the control system as a whole performs satisfactorily, i.e., that it places the response $y(x)$ into the smaller target set $\hat{Y}_x^*$.

THEOREM 5.1. *Suppose that the plant obeys the boundedness condition* (4.2). *Suppose further that a partial identification operator $\hat{K}$ has been chosen in such a*

way that the sets $\hat{A}$ *in* (5.1) *are elements of* $\mathscr{C}(\bar{A})$ *and furthermore that*

$$d(\hat{K}(u_1, y_1), \hat{K}(u_2, y_2)) \leqq \hat{m}_K[d(y_1, y_2) + \mu_2 d(u_1, u_2)] \tag{5.5}$$

with

$$m_F \hat{m}_K \mu_1 \mu_2 < \lambda_2$$

on $\bar{U} \times \bar{Y}$. *Suppose finally that a target set* $\hat{Y}_x^*$ *has been selected for every* x *in accordance with* (5.2). *Then there exists a controller of the form* (5.3), (5.4) *which achieves satisfactory system performance in the sense that* $y \in \hat{Y}_k^*$ *for every* x.

Proof. The proof of this theorem is a direct analogue in many respects to that of Theorem 4.1. It will therefore be only sketched here. Omitting, as usual, notational reference to x, G_1 and G_2 are first required to obey

$$\begin{aligned} F(G_1,(\hat{Y}, \hat{Y}, \hat{A}, \hat{Y}^*), \hat{A}) &= \hat{Y}, \\ G_2(\hat{Y}, \hat{Y}, \hat{u}, \hat{Y}^*) = \hat{K}(\hat{u}, \hat{Y}^*) &= \hat{A}, \end{aligned} \tag{5.6}$$

which are analogous to (4.5) and (4.6). It follows then, as in the proof of Theorem 4.1, that $\hat{Y}_x^*$ is a fixed element of the system. In order to achieve contraction, one now specifies by analogy to (4.7),

$$\begin{aligned} d(u_1, u_2) &\leqq m_1[d(\hat{Y}_1, \hat{Y}_2) + d(\hat{Z}_1, \hat{Z}_2) + \lambda_1 d(\hat{A}_1, \hat{A}_2)], \\ d(\hat{A}_1, \hat{A}_2) &\leqq m_2[d(\hat{Y}_1, \hat{Y}_2) + d(\hat{Z}_1, \hat{Z}_2) + \lambda_2 d(u_1, u_2)]. \end{aligned}$$

These relations should now be accompanied by the counterparts to (4.2) and (4.3), namely

$$\begin{aligned} d(F(u_1, \hat{A}_1), F(u_2, \hat{A}_2)) &\leqq m_F[d(u_1, u_2) + \mu_1 d(\hat{A}_1, \hat{A}_2)], \\ d(\hat{K}(u_1, \hat{Y}_1), \hat{K}(u_2, \hat{Y}_2)) &\leqq \hat{m}_K[d(\hat{Y}_1, \hat{Y}_2) + \mu_2 d(u_1, u_2)], \end{aligned} \tag{5.7}$$

which are in fact implied by the assumptions of the theorem. Consider for instance, (5.7a). The distance on the left is the Hausdorff distance between two sets $\in \mathscr{C}(\bar{Y})$, namely $F(u_1, \hat{A}_1)$ and $F(u_2, \hat{A}_2)$. This is the quantity

$$\begin{aligned} &d(F(u_1, \hat{A}_1), F(u_2, \hat{A}_2)) \\ &\quad = \max\{\sup_1 \inf_2 d(F(u_1, \alpha_1), F(u_2, \alpha_2)), \sup_2 \inf_1 d(F(u_1, \alpha_1) F(u_2, \alpha_2))\}, \end{aligned}$$

where, e.g., $\sup_1$ indicates the supremum over $\alpha_1 \in \hat{A}_1$. According to (5.5), therefore,

$$\begin{aligned} &d(F(u_1, \hat{A}_1), F(u_2, \hat{A}_2)) \\ &\quad \leqq m_F \max\{\sup_1 \inf_2 [d(u_1, u_2) + \mu_1 d(\alpha_1, \alpha_2)], \sup_2 \inf_1 [d(u_1, u_2) \\ &\qquad + \mu_1(\alpha_1, \alpha_2)]\} \\ &\quad = m_F \max\{[d(u_1, u_2) + \mu_1 \sup_1 \inf_2 d(\alpha_1, \alpha_2)], [d(u_1, u_2) \\ &\qquad + \mu_1 \sup_2 \inf_1 d(\alpha_1, \alpha_2)]\} \\ &\quad = m_F[d(u_1, u_2) + \mu_1 \max\{\sup_1 \inf_2 d(\alpha_1, \alpha_2), \sup_2 \inf_1 d(\alpha_1, \alpha_2)\}] \\ &\quad = m_F[d(u_1, u_2) + \mu_1 d(\hat{A}_1, \hat{A}_2)] \end{aligned}$$

which is (5.7a). Its counterpart, (5.7b), is proved analogously. The parallel with

the proof of Theorem 4.1 is now complete, and so is therefore the conclusion that contraction is assured on $\mathscr{C}(\bar{Y})$ if

$$m_F\hat{m}_K\mu_1\mu_2<\lambda_2, \quad m_1m_2\lambda_1\lambda_2<1, \quad m_2\lambda_2=\mu_2\hat{m}_K,$$

as was to be shown. Since $\mathscr{C}(\bar{Y})$ is complete the contraction converges on a unique fixed element, in the present case $\hat{Y}^*$. The theorem is accordingly proved.

REFERENCES

[1] D. P. BERTSEKAS AND I. B. RHODES, *On the minimax reachability of target sets and target tubes*, Automatica, 7 (1971), pp. 233–247.

[2] S. S. L. CHANG, *Fuzzy dynamic programming and the decision making process*, Proc. 3rd Princeton Conference on Information Sciences and Systems, 1969, pp. 200–203.

[3] J. DIEUDONNÉ, *Foundations of Modern Analysis*, vol. I, Academic Press, New York, 1969.

[4] F. HAUSDORFF, *Grundzuege der Mengenlehre*, Chelsea, New York, 1949.

[5] J. L. KELLEY, *Hyperspaces of a continuum*, Trans. Amer. Math. Soc., 52 (1942), pp. 23–36.

[6] H. KUSHNER, *Introduction to Stochastic Control*, Holt, Rinehart and Winston, New York, 1971.

[7] L. A. LUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Gordon and Breach, New York, 1969.

[8] E. MICHAEL, *Topologies on spaces of subsets*, Trans. Amer. Math. Soc., 71 (1971), pp. 152–182.

[9] H. RADSTROEM, *An embedding theorem for spaces of convex sets*, Proc. Amer. Math. Soc., 3 (1952), pp. 165–169.

[10] A. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.

[11] JA. S. TSYPKIN, *Adaption und Lernen in Automatischen Systemem*, Oldenbourg, Munich, 1966.

[12] J. C. WILLEMS, *The Analysis of Feedback Systems*, MIT Press, Cambridge, MA, 1971.

[13] G. ZAMES, *On the input-output stability of time-varying non-linear feedback systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238 and pp. 465–476.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

ON LOWER SEMICONTINUITY OF INTEGRAL FUNCTIONALS. I*

A. D. IOFFE†

Abstract. A necessary and sufficient condition for the integral functional $I(x(\cdot), y(\cdot)) = \int_G f(t, x(t), y(t))\, d\mu$ to be sequentially lower semicontinuous with respect to some kinds of strong convergence of $x(\cdot)$-components and weak convergence of $y(\cdot)$-components is proved. It is shown how many known and new sufficient conditions can be easily derived from this result. Such properties of the integrand as measurability, lower semicontinuity in (x, y) and convexity in y are also discussed. It appears that if $I(\cdot, \cdot)$ is lower semicontinuous then some other integrand $g(t, x, y)$ such that $g(t, x(t), y(t)) = f(t, x(t), y(t))$ a.e. for any measurable $x(\cdot)$, $y(\cdot)$ necessarily has these properties even if integrand f itself fails to satisfy some or any of them.

1. Introduction. Let G be a measure space with finite positive nonatomic complete measure μ. Let $f(t, x, y)$ be an extended-real-valued function on $G \times R^m \times R^n$ such that $f(t, x(t), y(t))$ is measurable for any measurable $x(\cdot)$, $y(\cdot)$ mapping G into R^m and R^n respectively. Such functions are usually called *integrands*. Consider two linear topological spaces, L and M, of summable mappings from G into R^m and R^n respectively and define the following integral functional on $L \times M$:

$$I(x(\cdot), y(\cdot)) = \int_G f(t, x(t), y(t))\, d\mu. \tag{1.1}$$

(Generally speaking, this integral can make no sense for certain $x(\cdot)$, $y(\cdot)$ but our further assumptions will exclude such a possibility.)

The purpose of this paper is to derive necessary and sufficient conditions for $I(\cdot, \cdot)$ to be sequentially lower semicontinuous (sequentially l.s.c.) on $L \times M$. This problem was motivated mainly by needs of the existence theory in calculus of variations and optimal control. A detailed description of interrelations between the problems can be found in a recent survey of Olech [10]. We only note that our problem is not the absolutely exact reflection of the situation typical for calculus of variations and optimal control. We consider $x(\cdot)$ and $y(\cdot)$ separately, with no connections between them, whereas in typical situations $x(\cdot)$ and $y(\cdot)$ are mutually dependent; as usual, $y(\cdot)$ is a collection of derivatives of $x(\cdot)$. The reduction to the problem with separated $x(\cdot)$ and $y(\cdot)$ in the framework of the lower semicontinuity and existence theory was undertaken long ago (it is hardly possible to say in what particular paper) mainly to cope with serious technical difficulties and it has proved to be very fruitful (see [2], [3], [9], [11], [13]). In each of the above works, strong convergence of $x(\cdot)s$ in some L_p and weak convergence of $y(\cdot)$ in some L_q was considered.

Here we shall study the problem in a more general setting but also with respect to certain kinds of "strong" topology in L and "weak" topology in M. Our assumptions cover in particular the case of Orlicz spaces with norm and weak topologies respectively.

* Received by the editors February 17, 1976.

† c/o R. T. Rockafellar, Department of Mathematics, University of Washington, Seattle, Washington 98195.

A lower semicontinuity theorem usually contains two groups of assumptions on the integrand. The first consists of qualitative assumptions concerning functional properties of the integrand, measurable linear and topological. The second includes so-called boundedness conditions containing some minorants (rarely majorants) for the integrand.

We prove here two theorems, the first containing what could be characterized as a necessary and sufficient boundedness condition and the second dealing with conditions of the first group. We show that many sufficient criteria, both known and new, can easily be derived from the first theorem. In this theorem, the integrand f is assumed to be measurable, l.s.c. in (x, y) and convex in y. These assumptions are very general and, no doubt, sufficient for practical purposes. The second theorem shows that they are also necessary in some sense. Thus both theorems together give a description of integrands generating integral functionals which are sequentially lower semicontinuous with respect to mixed strong-weak convergence.

2. Statements of the main theorems. Let T be a subset of G. By $\chi_T(t)$ we denote the characteristic function of T:

$$\chi_T(t) = \begin{cases} 1, & \text{if } t \in T, \\ 0, & \text{if } t \notin T. \end{cases}$$

Throughout the paper we assume that both L and M are *decomposable*, that is, if $z(\cdot)$ belongs to one of them, then $\chi_T(\cdot)z(\cdot)$ belongs to the same space whenever T is a measurable subset of G.

Recall that the integrand $f(t, x, y)$ is called $\mathcal{L} \otimes \mathcal{B}$-measurable if it is measurable with respect to the σ-algebra generated by products of measurable subsets of G and Borel subsets of $R^m \times R^n$.

We shall say that integrand $f(t, x, y)$ satisfies the *lower compactness property* on $L \times M$ if any sequence $f^-(t, x_k(t), y_k(t))$ is weakly precompact in L_1 whenever the $x_k(\cdot)$ converge in L, $y_k(\cdot)$ converge in M and $I(x_k(\cdot), y_k(\cdot)) \leqq a < \infty$ for all $k = 1, 2, \cdots$. Here $f^- = \min(f, 0)$.

Consider the following hypotheses on topologies in L and M:

(H_1) If $z_k(\cdot)$, $k = 1, 2, \cdots$, belong to one of the spaces and converge there to zero and if $\mu T_k \to 0$, then $\chi_{T_k}(\cdot)z_k(\cdot)$ also converge to zero.

(H_2) The topology in L is not weaker than the topology of convergence in measure; the topology in M is not weaker than the topology induced in M by the weak topology of L_1^n.

(H_3) L and M contain bounded measurable mappings and their topologies are not stronger than the topology of almost everywhere uniform convergence.

(H_4) If $y(\cdot) \in M$ and T_k, $k = 1, 2, \cdots$, is a sequence of measurable subsets of G such that $\alpha_k(\cdot) = \chi_{T_k}(\cdot)$ converge weakly* in L_∞ to some $\alpha(\cdot)$ then $\alpha_k(\cdot)y(\cdot)$ converge in M to $\alpha(\cdot)y(\cdot)$.

It is easy to see that spaces L_p with norm or weak topologies satisfy all of these hypotheses.

THEOREM 1 (Lower semicontinuity theorem). *Let L and M satisfy* (H_1) *and* (H_2). *Assume that $f(t, x, y)$ is $\mathcal{L} \otimes \mathcal{B}$-measurable, lower semicontinuous in (x, y) and convex in y. In order that $I(\cdot, \cdot)$ be lower semicontinuous on $L \times M$ and*

everywhere on $L \times M$ more than $-\infty$, it is necessary and (if $I(\cdot\cdot\cdot)$ is finite at least at one point in $L \times M$) sufficient that f satisfy the lower compactness property.

The proof of the sufficiency part is more or less traditional (cf. [2], [11]); it is, in fact, a modified version of the proof given in [2] for a less general situation. As to necessity, it is quite elementary.

To state the second theorem, we need some additional notions. We shall say that two integrands $f(t, x, y)$ and $g(t, x, y)$ are *measurably equivalent* if $f(t, x(t), y(t)) = g(t, x(t), y(t))$ a.e. for any measurable $x(\cdot)$ and $y(\cdot)$. If f satisfies some property up to measurable equivalence, we say that f *virtually* satisfies this property.

THEOREM 2 (Virtual measurability theorem). *Let L and M satisfy* (H_3). *Assume that $I(\cdot, \cdot)$ is lower semicontinuous on $L \times M$ and not everywhere on $L \times M$ equal to $\pm\infty$. Then $f(t, x, y)$ is measurably equivalent to some other integrand $g(t, x, y)$ which is $\mathcal{L} \otimes \mathcal{B}$-measurable and lower semicontinuous in (x, y). If in addition M satisfies* (H_4) *then g is convex in y.*

Remark. Here and below, assumptions and statements are referred to sets with measure-negligible projection on G. For instance, such words as "f is l.s.c. in (x, y) and convex in y" mean that there is $G' \subset G$ with $\mu G' = \mu G$ and such that the functions $(x, y) \to f(t, x, y)$ are l.s.c. for all $t \in G'$ and the functions $y \to f(t, x, y)$ are convex for all $t \in G'$, $x \in R^m$.

Note that two measurably equivalent *measurable* integrands coincide up to a set which has measure-negligible projection on G. This follows immediately from Aumann's selection theorem [1]. Hence the integrand g in Theorem 2 is uniquely defined.

In the case when f does not depend on x and M is L_1^n with weak topology this theorem was actually proved by Olech [12]; this proof, however, cannot be extended to the general case because it heavily depends on convexity.

It is interesting to note furthermore that in spite of Theorem 2, the integrand f itself may fail to have any of the properties inherent in g. Probably, the most striking example of this sort follows from a remarkable fact communicated to the author by M. A. Krasnosel'skii and A. V. Pokrovskii:

> under the continuum hypothesis, there exists a real-valued function $h(t, x)$ defined on $R \times R$ such that
>
> (i) for any $t \in R$, the set $\{x | h(t, x) = 0\}$ is denumerable;
>
> (ii) $h(t, x(t)) = 0$ a.e. for any measurable $x(\cdot)$.

In this case

$$\int_G h(t, x(t))\, dt = 0$$

for any measurable $x(\cdot)$ and $G \subset R$ whatever values h assumes outside the set on which it is equal to zero. Hence h can be neither measurable nor l.s.c. and convex in x but the integral will remain continuous in any topology.

3. Some corollaries. In this section we shall prove, using the lower semicontinuity theorem, some more usable sufficient criteria, both well known and new, for $I(\cdot, \cdot)$ to be l.s.c. To do this, we need a widely known characterization of weakly precompact sets in L_1 (see [4], [7]).

PROPOSITION 1. *Let A be a nonempty subset of L_1. Then the following three conditions are equivalent*:

(i) *A is weakly precompact*;

(ii) *A is equi-uniformly summable, that is,*

$$\sup_{x(\cdot)\in A}\int_T |x(t)|\,d\mu \to 0 \quad \text{if } \mu T \to 0;$$

(iii) *there exists a nonnegative nondecreasing function $h(\cdot)$ on $[0,\infty)$ such that*

$$\lim_{t\to\infty}\frac{h(t)}{t}=\infty, \qquad \int_G h(|x(t)|)\,d\mu \leqq 1 \quad \forall x(\cdot)\in A.$$

By $|\cdot|$ and $\langle\cdot,\cdot\rangle$ we shall denote the Euclidean norm and the inner product respectively; L_s^m $(1\leqq s\leqq\infty)$ will denote the space of all measurable mappings $x(\cdot)\colon G\to R^m$ such that $|x(\cdot)|$ belongs to L_s. We shall say that $I(\cdot,\cdot)$ is (s,q)-l.s.c. if it is l.s.c. with respect to the norm convergence of $x(\cdot)s$ in L_s^m and the weak (weak* if $q=\infty$) convergence of $y(\cdot)s$ in L_q^n.

In all of the following theorems, the integrand f is supposed $\mathscr{L}\otimes\mathscr{B}$-measurable, l.s.c. in (x, y) and convex in y.

THEOREM 3. *Assume in addition to the assumption of Theorem* 1 *that f is nonnegative. Then $I(\cdot,\cdot)$ is lower semicontinuous on $L\times M$.*

Proof. Here $f^-(t,x,y)\equiv 0$.

For (s,q)-lower semicontinuity, this fact follows from many recent results (see Theorems 4 and 5 below). Earlier versions may be found in [6], [9].

THEOREM 4 (Berkovitz [2]). *Let*

$$f(t,x,y)\geqq\langle a(t),y\rangle+b(t)$$

for some $a(\cdot)\in L_{q'}^n$, $b(\cdot)\in L_1$. Then $I(\cdot,\cdot)$ is (s,q)-l.s.c. for every s.

Here q' is defined by $1/q+1/q'=1$.

Proof. If the $y_k(\cdot)$ converge weakly (weakly* if $q=\infty$), then the $\langle a(\cdot), y_k(\cdot)\rangle$ converge weakly in L_1.

THEOREM 5 (Olech [11]). *Let*

$$f(t,x,y)\geqq -c(|x|+|y|)+b(t)$$

for some $c\in R$, $b(\cdot)\in L_1$. Then $I(\cdot,\cdot)$ is $(1,1)$-l.s.c.

Proof. If $x_k(\cdot)$ converge strongly in L_1^m and the $y_k(\cdot)$ converge weakly in L_1^n then the $|x_k(\cdot)|$ converge strongly in L_1 and the sequence of $|y_k(\cdot)|$ is weakly precompact in L_1 by Proposition 1.

Remark 1. The conditions of Theorem 5 are also necessary for $I(\cdot,\cdot)$ to be $(1,1)$-l.s.c. This fact was proved by Poljak [13] for integrands satisfying the Carathéodory condition. In our case, the proof needs almost no changes. On the other hand, the inequality in Theorem 5 is equivalent to the fact that $I(\cdot,\cdot)$ does not assume the value $-\infty$ on $L_1^m\times L_1^n$. Hence $I(\cdot,\cdot)$ is $(1,1)$-l.s.c. if and only if it is well-defined on $L_1^m\times L_1^n$. However for $q>1$, this nice result fails (see Remark 3 below). Note in this connection that Remark 1 in [2] is true only if all of q_i are equal to 1.

THEOREM 6 (Cesari [3]). *Assume that*

(i) $f(t, x, y) \geqq \beta > -\infty$ *for all* t, x *and* y *satisfying* $|y| \leqq 1$;

(ii) *there is a real* $\alpha \geqq \beta$ *such that for any fixed* t, x *either* $\alpha \geqq f(t, x, 0)$ *or* $f(t, x, y) = \infty$ *for all* $y \in R^n$.

Then $I(\cdot, \cdot)$ *is* $(1, 1)$*-l.s.c.*

Proof. Since f is convex in y, assumptions (i), (ii) imply that

$$f(t, x, y) \geqq \beta - (\alpha - \beta)|y|.$$

Apply Theorem 5.

THEOREM 7. *Let* $1 < q < \infty$ *and*

$$f(t, x, y) \geqq -c|x|^s - g(|y|) + b(t) \tag{3.1}$$

for some $c \in R$, $b(\cdot) \in L_1$ *and nonnegative nondecreasing function* $g(t)$ *on* $[0, \infty]$ *satisfying*

$$\lim_{t \to \infty} g(t)/t^q = 0. \tag{3.2}$$

Then $I(\cdot, \cdot)$ *is* (s, q)*-l.s.c.*

Proof. Let

$$h(\tau) = \inf \{t^q | g(t) = \tau\}$$

(assuming as usual that $\inf \varnothing = \infty$). We have

$$0 \leqq h(g(t)) \leqq t^q \quad \forall t \geqq 0, \tag{3.3}$$

$h(\cdot)$ does not decrease and

$$\lim_{\tau \to \infty} h(\tau)/\tau = \liminf_{t \to \infty} t^q/g(t) = \infty. \tag{3.4}$$

If now the $y_k(\cdot)$ converge weakly in L_q^n then the norms of $y_k(\cdot)$ are bounded by some N and hence by (3.3)

$$\int_G h(g(|y_k(t)|))\, d\mu \leqq \int_G |y_k(t)|^q\, d\mu \leqq N^q.$$

In this case (3.4) shows, by virtue of Proposition 1, that the sequence of $g(|y_k(t)|)$ is weakly precompact in L_1.

Remark 2. To get a corresponding result for $q = \infty$, one should only replace (3.2) by the assumption that $g(\cdot)$ is nowhere on $[0, \infty]$ equal to ∞. The proof is equally simple. An analogous assertion is true for $s = \infty$: it suffices to place an arbitrary continuous function of x in (3.1) instead of $c|x|^s$.

Remark 3. A well-known necessary and sufficient condition for $I(\cdot, \cdot)$ to be well defined on $L_s^m \times L_q^n$ is that

$$f(t, x, y) \geqq -c(|x|^s + |y|^q + b(t)). \tag{3.5}$$

We see that this condition differs from (3.1), (3.2) in one point: it permits the integrand to decrease in y as $-|y|^q$ while (3.1), (3.2) demand that this decrease be slower than $-|y|^q$. However neither (3.1), (3.2) are necessary for $I(\cdot, \cdot)$ to be

(s, q)-l.s.c. nor is (3.5) sufficient. Consider two examples.

If

$$f(t, x, y) = xy, \qquad m = n = 1, \quad s = q = 2,$$

then f does not satisfy (3.1), (3.2) no matter which g satisfying (3.2) is taken. However, $I(\cdot\,, \cdot\,)$ is $(2, 2)$-l.s.c.

On the other hand, let $G = (0, 1)$, $m = n = 1$, $q > 1$ and

$$f(t, x, y) = \frac{1}{q'}\left|\frac{x}{t}\right|^{q'} + \frac{x}{t}y. \tag{3.6}$$

We have

$$f(t, x, y) \geqq -|y|^q/q,$$

so that f satisfies (3.5) for every s. Nonetheless, the corresponding integral $I(\cdot\,,)$ is not (s, q)-l.s.c. To show this, consider two sequences

$$x_k(t) = \begin{cases} tk^{1/q'}, & \text{if } 0 < t \leqq 1/k, \\ 0, & \text{if } 1/k < t < 1; \end{cases}$$

$$y_k(t) = \begin{cases} -k^{1/q}, & \text{if } 0 < t \leqq 1/k, \\ 0, & \text{if } 1/k < t < 1. \end{cases}$$

Then $x_k(\cdot) \to 0$ uniformly and $y_k(\cdot) \to 0$ weakly in L_q. However, $I(0, 0) = 0$ whereas $I(x_k(\cdot), y_k(\cdot)) = -1/q$.

THEOREM 8. *Let $q > 1$, $s < \infty$. Assume that there exists a $\mathcal{L} \otimes \mathcal{B}$-measurable mapping $p(t, x)$: $G \times R^m \to R^n$ such that for some $c > 0$, $b(\cdot) \in L_1$ the following two inequalities hold*:

$$f(t, x, y) \geqq \langle p(t, x), y\rangle - c|x|^s + b(t); \tag{3.7}$$

$$|p(t, x)|^{q'} \leqq c|x|^s + b(t). \tag{3.8}$$

Then $I(\cdot\,, \cdot\,)$ is (s, q)-l.s.c.

Proof. If the $x_k(\cdot)$ converge strongly in L_s^m, then the functions $|x_k(\cdot)|^s$ are equi-uniformly summable; hence so are $|p(t, x_k(t))|^{q'}$. It follows that

$$\|\chi_T(\cdot)p(\cdot\,, x_k(\cdot))\|_{q'} \to 0 \quad \text{as} \quad \mu T \to 0.$$

If in addition $y_k(\cdot)$ converge weakly (weakly* if $q = \infty$) in L_q^n then their q-norms are bounded by some $N > 0$. Therefore

$$\left|\int_T \langle p(t, x_k(t)), y_k(t)\rangle \, d\mu\right| \leqq N\|\chi_T(\cdot)p(\cdot\,, x_k(\cdot))\|_{q'} \to 0$$

uniformly in k when $\mu T \to 0$. It remains to apply Proposition 1.

Two last theorems can be extended to Orlicz spaces. Recall some basic notions and facts about these spaces (see [5], [8]).

Let $\varphi(t, x)$ be a $\mathcal{L} \otimes \mathcal{B}$-measurable extended-real-valued function on $G \times R^m$ which is nonnegative, convex and l.s.c. in x and satisfying $\varphi(t, 0) = 0$,

$\varphi(t, -x) = -\varphi(t, x)$. Let

$$\varphi^*(t, u) = \sup_x (\langle u, x\rangle - \varphi(t, x))$$

be the Fenchel conjugate to $\varphi(t, \cdot)$. Then φ^* is also $\mathcal{L} \otimes \mathcal{B}$-measurable, nonnegative, convex in u and satisfies $\varphi^*(t, 0) = 0$, $\varphi^*(t, -u) = -\varphi^*(t, u)$. We shall say that φ is a *Young function* if there exists an $\varepsilon > 0$ such that the functions $t \to \varphi(t, x)$ and $t \to \varphi^*(t, u)$ are summable whenever $|x| < \varepsilon$, $|u| < \varepsilon$.

Let φ be a Young function and

$$J_\varphi(x(\cdot)) = \int_G \varphi(t, x(t))\, d\mu.$$

It is not difficult to show that $x(\cdot) \in L_1^m$ if $J_\varphi(x(\cdot)) < \infty$. The set

$$\hat{L}_\varphi = \{x(\cdot) \in L_1^m | J_\varphi(x(\cdot)) < \infty\}$$

is called the *Orlicz class* (generated by φ). The conical hull of $\hat{L}_\varphi$:

$$L_\varphi = \bigcup_{k=1}^{\infty} k\hat{L}_\varphi$$

is a linear space called an *Orlicz space*, which is converted into a Banach space having been supplied with the norm

$$\|x(\cdot)\|_\varphi = \inf\{\lambda > 0 | J_\varphi(\lambda^{-1}x(\cdot)) \leqq 1\}.$$

The space L_{φ^*} is defined in the same way. The connection between both spaces is the following: each of them is total on the other with respect to the pairing

$$\int_G \langle u(t), x(t)\rangle\, d\mu.$$

It is said that φ satisfies the Δ_2-*condition* if there is an $a > 0$ such that

$$\varphi(t, 2x) \leqq a\varphi(t, x).$$

If φ satisfies the Δ_2-*condition* (and only in this case)

$$\hat{L}_\varphi = L_\varphi, \qquad L_\varphi^* = L_{\varphi^*}.$$

PROPOSITION 2. *Let $\varphi(t, x)$ be a Young function on $G \times R^m$. Consider a sequence $\{x_k(\cdot)\} \subset L_\varphi$ norm converging to some $x(\cdot)$. Assume that $\|x(\cdot)\|_\varphi < 1/2$, $\|x_k(\cdot)\|_\varphi < 1/2$. Then the functions $\varphi(t, x_k(t))$ are equi-uniformly summable.*

Proof. Since φ is convex in x, we have

$$0 \leqq \varphi(t, x_k(t)) \leqq (1/2)(\varphi(t, 2x(t)) + \varphi(t, 2(x_k(t) - x(t)))). \tag{3.9}$$

By the assumptions $\|2x(\cdot)\|_\varphi < 1$ and hence $\varphi(t, x(t))$ is summable. On the other hand, $w_k(\cdot) = 2(x_k(\cdot) - x(\cdot))$ tend to zero; in particular $\|w_k(\cdot)\|_\varphi < 1$ if k is more than some k_0. For such k

$$\int_G \varphi(t, w_k(t))\, d\mu \leqq \|w_k(\cdot)\|_\varphi$$

which follows from the definition of the norm and from the fact that $J_\varphi(\cdot)$ is a convex function equal to zero at the origin. Thus $\varphi(\cdot, w_k(\cdot))$ $(k \geqq k_0)$ are nonnegative functions with integrals tending to zero. Therefore they are equi-uniformly summable. By virtue of (3.9), this proves that $\varphi(t, x_k(t))$ are equi-uniformly summable for $k \geqq k_0$ and hence for all k because $J_\varphi(x_k(\cdot)) \leqq 1$ by the assumptions.

PROPOSITION 3. *Let $\varphi(t, x)$ be a Young function on $G \times R^m$ satisfying the Δ_2-condition. Let A be a subset of L_1^m such that the set $\{\varphi(\cdot, x(\cdot)) | x(\cdot) \in A\}$ is equi-uniformly summable. Then*

$$\sup_{x(\cdot)\in A} \|\chi_T(\cdot)x(\cdot)\|_\varphi \to 0, \quad \text{if } \mu T \to 0.$$

Proof. Fix a positive integer r. It is sufficient to show that for some $\varepsilon > 0$, $\|\chi_T(\cdot)x(\cdot)\|_\varphi \leqq 2^{-r}$ if $\mu T < \varepsilon$ and $x(\cdot) \in A$. Choose $\varepsilon > 0$ such that

$$\int_T \varphi(t, x(t))\, d\mu \leqq a^{-r} \quad \forall x(\cdot) \in A,$$

if $\mu T < \varepsilon$ (a being the constant in the Δ_2-condition). If now $\mu T < \varepsilon$, $x(\cdot) \in A$, we get

$$\int_G \varphi(t, 2^r \chi_T(t) x(t))\, d\mu = \int_T \varphi(t, 2^r x(t))\, d\mu$$

$$\leqq a^r \int_T \varphi(t, x(t))\, d\mu \leqq 1,$$

that is, $\|\chi_T(\cdot)x(\cdot)\|_\varphi \leqq 2^{-r}$.

We shall apply Theorem 1 in the following situation. Let $\varphi(t, x)$ and $\psi(t, y)$ be two Young functions defined on $G \times R^m$ and $G \times R^n$ respectively. Consider the spaces L_φ and L_ψ, the first with the norm topology and the second with the $\sigma(L_\psi, L_{\psi^*})$-topology. It is easy to see that L_φ as L and L_ψ as M satisfy all of the hypotheses (H_1)–(H_4). For simplicity we shall say that $I(\cdot, \cdot)$ is (φ, ψ)-l.s.c. if it is sequentialy lower semicontinuous on $L_\varphi \times L_\psi$ with respect to the above mentioned topologies.

THEOREM 9. *Assume that for any $c > 0$ there are a summable function $b_c(t)$ and a $\mathcal{L} \otimes \mathcal{B}$-measurable integrand $g_c(t, \tau)$ on $G \times [0, \infty)$ which is nondecreasing in τ and satisfies*

$$\lim_{|y|\to\infty} \sup_{t\in G} \frac{g_c(t, |y|)}{\psi(t, \alpha y)} = 0, \quad \forall \alpha > 0;$$

$$f(t, x, y) \geqq -\varphi(t, cx) - g_c(t, |y|) + b_c(t).$$

Then $I(\cdot, \cdot)$ is (φ, ψ)-l.s.c.

Proof. Let $x_k(\cdot)$ converge strongly in L_φ and $y_k(\cdot)$ $\sigma(L_\psi, L_{\psi^*})$-converge in L_ψ. Then the norms of $x_k(\cdot)$ are bounded by some N_0 and the norms of $y_k(\cdot)$ are bounded by some N_1. Take $c < 1/(2N_0)$. Then $\|cx_k(\cdot)\|_\varphi < 1/2$ and by Proposition 2 the functions $\varphi(\cdot, cx_k(\cdot))$ are equi-uniformly summable.

To prove the theorem, it suffices to show that the functions $g_c(\cdot, |y_k(\cdot)|)$, $k = 1, 2, \cdots$, are also equi-uniformly summable. Take $\alpha < 1/N_1$ and let

$$h(t, \lambda) = \inf\{\psi(t, \alpha y) | g_c(t, |y|) = \lambda\},$$

$$h(\lambda) = \inf_{t \in G} h(t, \lambda).$$

Since ψ is convex in y and nonnegative and $\psi(t, 0) = 0$, $h(t, \lambda)$ and $h(\lambda)$ are nondecreasing and nonnegative. Furthermore

$$h(t, g_c(t, |y|)) \leqq \psi(t, \alpha y)$$

and

$$\lim_{\lambda \to \infty} \frac{h(\lambda)}{\lambda} = \lim_{\lambda \to \infty} \inf_{t \in G} \frac{h(t, \lambda)}{\lambda}$$

$$= \lim_{\lambda \to \infty} \inf_{t \in G} \inf \left\{ \frac{\psi(t, \alpha y)}{\lambda} \middle| g_c(t, |y|) = \lambda \right\}$$

$$= \liminf_{|y| \to \infty} \inf_{t \in G} \frac{\psi(t, \alpha y)}{g_c(t, |y|)} = \infty.$$

Finally

$$\int_G h(g_c(t, |y_k(t)|))\, d\mu \leqq \int_G h(t, g_c(t, |y_k(t)|))\, d\mu$$

$$\leqq \int_G \psi(t, \alpha y_k(t))\, d\mu \leqq 1,$$

according to the choice of α. Applying Proposition 1, we obtain what we need.

THEOREM 10. *Let ψ^* satisfy the Δ_2-condition. Assume that there is a $\mathcal{L} \otimes \mathcal{B}$-measurable mapping $p(t, x)$: $G \times R^m \to R^n$ such that for any $c > 0$ there are $k_c > 0$ and $b_c(\cdot) \in L_1$ satisfying*

$$f(t, x, y) \geqq \langle p(t, x), y \rangle - \varphi(t, cx) - b_c(t);$$
$$\psi^*(t, p(t, x)) \leqq k_c \varphi(t, cx) + b_c(t). \tag{3.10}$$

Then $I(\cdot, \cdot)$ is (φ, ψ)-l.s.c.

Proof. The proof is similar to the proof of Theorem 8. Let $x_k(\cdot)$ converge strongly in L_φ. As in the proof of the preceding theorem, we can choose $c > 0$ to make the functions $\varphi(t, cx_k(t))$ equi-uniformly summable. Then functions $\psi^*(t, p(t, x_k(t)))$ are also equi-uniformly summable and by Proposition 3, so are $\langle p(t, x_k(t)), y_k(t) \rangle$ if the $y_k(\cdot)$ converge in the $\sigma(L_\psi, L_{\psi^*})$-topology.

4. Proof of the lower semicontinuity theorem.

Necessity. Assume that $I(\cdot, \cdot)$ is everywhere greater than $-\infty$ and sequentially l.s.c. on $L \times M$. Consider a sequence $\{x_k(\cdot), y_k(\cdot)\}$ such that the $x_k(\cdot)$ converge in L to some $x(\cdot)$, the $y_k(\cdot)$ converge to some $y(\cdot)$ in M and $I(x_k(\cdot), y_k(\cdot)) \leqq a < \infty$ for all k. Then $I(x(\cdot), y(\cdot)) \leqq a$ and hence $|I(x(\cdot), y(\cdot))| < \infty$.

We must prove that the functions $f^-(t, x_k(t), y_k(t))$, $k = 1, 2, \cdots$, are equi-uniformly summable. Assume the contrary. Then for any integer $s = 1, 2, \cdots$,

there is a subscript $k_s \geqq s$ such that for some measurable set $T_s \subset G$ with $\mu T_s \leqq 1/s$,

$$\int_{T_s} f(t, x_{k_s}(t), y_{k_s}(t))\, d\mu \leqq -\delta < 0. \tag{4.1}$$

Let

$$u_s(t) = (1-\chi_{T_s}(t))x(t) + \chi_{T_s}(t)x_{k_s}(t),$$
$$v_s(t) = (1-\chi_{T_s}(t))y(t) + \chi_{T_s}(t)y_{k_s}(t).$$

Due to (H_1), $u_s(\cdot) \to x(\cdot)$, $v_s(\cdot) \to y(\cdot)$. But (4.1) implies that

$$I(u_s(\cdot), v_s(\cdot)) - I(x(\cdot), y(\cdot))$$
$$= \int_{T_s} (f(t, x_{k_s}(t), y_{k_s}(t)) - f(t, x(t), y(t)))\, d\mu \leqq -\delta/2,$$

for sufficiently large s which contradicts the fact that $I(\cdot, \cdot)$ is lower semicontinuous.

Sufficiency. Assume that f satisfies the lower compactness property on $L \times M$. Then obviously

$$\int_G f^-(t, x(t), y(t))\, d\mu > -\infty, \quad \forall x(\cdot) \in L, \quad \forall y(\cdot) \in M.$$

Consider again a sequence $\{x_k(\cdot), y_k(\cdot)\}$ converging to $(x(\cdot), y(\cdot))$ in $L \times M$ and such that

$$I(x_k(\cdot), y_k(\cdot)) \leqq a < \infty \quad \forall k = 1, 2, \cdots. \tag{4.2}$$

To prove the theorem, it suffices to show that

$$I(x(\cdot), y(\cdot)) \leqq a. \tag{4.3}$$

First we shall demonstrate that no loss of generality will follow if we assume that f is bounded from below. Indeed, let the theorem be true for integrands bounded from below. Let

$$f_N(t, x, y) = \max(f(t, x, y), -N),$$
$$I_N(x(\cdot), y(\cdot)) = \int_G f_N(t, x(t), y(t))\, d\mu.$$

Since the functions $f^-(t, x_k(t), y_k(t))$, $k = 1, 2, \cdots$, form a weakly precompact set in L_1, we have in particular

$$\int_G f^-(t, x_k(t), y_k(t))\, d\mu \geqq -\beta > -\infty \quad \forall k = 1, 2, \cdots \tag{4.4}$$

for some $\beta > 0$. Therefore $I_N(x_k(\cdot), y_k(\cdot)) \leqq a + \beta$ for all k and N. Since the theorem is true for f_N, as we have just assumed, it follows that

$$\liminf_{k \to \infty} I_N(x_k(\cdot), y_k(\cdot)) \geqq I_N(x(\cdot), y(\cdot)) \geqq I(x(\cdot), y(\cdot)). \tag{4.5}$$

On the other hand, (4.4) implies that for any k

$$\mu\{t \in G | f(t, x_k(t), y_k(t)) \leqq -N\} \to 0 \quad \text{if } N \to \infty.$$

But these are just the sets on which $f(t, x_k(t), y_k(t))$ differ from $f_N(t, x_k(t), y_k(t))$. Note that the functions $f^-(t, x_k(t), y_k(t))$ are equi-uniformly summable by Proposition 1. Therefore for any $\varepsilon > 0$, there is an N_ε such that

$$I_N(x_k(\cdot), y_k(\cdot)) \leqq I(x_k(\cdot), y_k(\cdot)) + \varepsilon$$

for all $N \geqq N_\varepsilon$ and all $k = 1, 2, \cdots$. This together with (4.5) implies (4.2).

Thus we may suppose that f is bounded from below and even nonnegative. By (H_2), the $y_k(\cdot)$ converge weakly in L_1^n to $y(\cdot)$. Proposition 1 shows that there is a nonnegative nondecreasing function $h_0(\tau)$ on $[0, \infty)$ such that

$$\lim_{\tau \to \infty} h_0(\tau)/\tau = \infty;$$

$$\int_G h_0(|y_k(t)|)\, d\mu \leqq 1 \quad \forall k = 1, 2, \cdots.$$

Choose another nonnegative nondecreasing function $h(\tau)$ on $[0, \infty)$ such that

$$\lim_{\tau \to \infty} h(\tau)/\tau = \lim_{\tau \to \infty} h_0(\tau)/h(\tau) = \infty \tag{4.6}$$

(for instance $h(\tau) = (\tau h_0(\tau))^{1/2}$). Let

$$\begin{aligned} h_1(\xi) &= \inf\{h_0(\tau) | h(\tau) = \xi\}, \\ \xi_k(t) &= h(|y_k(t)|). \end{aligned} \tag{4.7}$$

We see (cf. proofs of Theorems 7 and 9) that h_1 is nonnegative, nondecreasing and

$$\lim_{\xi \to \infty} h_1(\xi)/\xi = \infty. \tag{4.8}$$

Furthermore,

$$\int_G h_1(\xi_k(t))\, d\mu = \int_G h_1(h(|y_k(t)|))\, d\mu \leqq \int_G h_0(|y_k(t)|)\, d\mu \leqq 1$$

(because $h_1(h(\tau)) \leqq h_0(\tau)$ by definition). In view of (4.8) this means that the sequence $\{\xi_k(\cdot)\}$ is weakly precompact in L_1.

By Mazur's theorem, some sequence of convex combinations of $(y_k(\cdot), \xi_k(\cdot))$ converge strongly in $L_1^n \times L_1$. In other words, some functions of the form

$$v_j(t) = \sum_{i=1}^{s_j} \alpha_{ij} y_{k_j+i}(t), \tag{4.9}$$

$$\eta_j(t) = \sum_{i=1}^{s_j} \alpha_{ij} \xi_{k_j+i}(t) \tag{4.10}$$

where $k_j < k_j + s_j < k_{j+1}$, $\alpha_{ij} \geqq 0$, $\sum_{i=1}^{s_j} \alpha_{ij} = 1$, converge strongly when $j \to \infty$, the first ones to a certain $y(\cdot)$ in L_1^n and the second to a certain $\eta(\cdot)$ in L_1. Extracting, if necessary, a subsequence, we may consider $v_j(\cdot)$ and $\eta_j(\cdot)$ converging a.e. on G.

Let

$$\lambda_j(t) = \sum_{i=1}^{s_j} \alpha_{ij} f(t, x_{k_j+i}(t), y_{k_j+i}(t)). \tag{4.11}$$

Then obviously (since f is assumed nonnegative)

$$\lambda_j(t) \geqq 0 \quad \text{a.e.} \quad \text{and} \quad \int_G \lambda_j(t)\, d\mu \leqq a \quad \forall j = 1, \cdots. \tag{4.12}$$

To prove (4.3), it is sufficient to verify that

$$\liminf_{j \to \infty} \lambda_j(t) \geqq f(t, x(t), y(t)) \quad \text{a.e.} \tag{4.13}$$

in which case (4.3) follows from Fatou's lemma.

By (H_2), the $x_k(\cdot)$ converge to $x(\cdot)$ in measure. Therefore we may assume that the $x_k(\cdot)$ also converge to $x(\cdot)$ a.e. Fix some $t \in G$ such that $x_k(t) \to x(t)$, $v_j(t) \to y(t)$, $\eta_j(t) \to \eta(t)$. Let

$$\varepsilon_j = \max_{1 \leqq i \leqq s_j} |x(t) - x_{k_j+i}(t)|.$$

Then $\varepsilon_j \to 0$ because $k_j \to \infty$ as $j \to \infty$. Consider the set

$$A_j = \{(v, \eta, \lambda) \in R^{n+2} | \lambda \geqq f(t, x, y),\ \eta = h(|v|) \text{ for some } x \text{ satisfying } |x - x(t)| \leqq \varepsilon_j\}.$$

Then (4.7), (4.9)–(4.11) show that

$$(v_j(t), \eta_j(t), \lambda_j(t)) \in \operatorname{conv} A_j.$$

By Carathéodory's theorem, for every j, there exist three $(n+3)$-tuples

$$(\beta_{j1}, \cdots, \beta_{jn+3}), \quad (x_{j1}, \cdots, x_{jn+3}), \quad (v_{j1}, \cdots, v_{jn+3})$$

such that

$$\beta_{ij} \geqq 0, \quad \beta_{j1} + \cdots + \beta_{jn+3} = 1 \quad \forall j = 1, \cdots, \tag{4.14}$$

$$|x_{ji} - x(t)| \leqq \varepsilon_j \quad \forall j = 1, \cdots, \quad \forall i = 1, \cdots, n+3, \tag{4.15}$$

$$\sum_{i=1}^{n+3} \beta_{ji} v_{ji} = v_j(t), \qquad \sum_{i=1}^{n+3} \beta_{ji} f(t, x_{ji}, v_{ji}) \leqq \lambda_j(t), \tag{4.16}$$

$$\sum_{i=1}^{n+3} \beta_{ji} h(|v_{ji}|) = \eta_j(t). \tag{4.17}$$

Due to (4.6), v_{ji} cannot tend to ∞ as $j \to \infty$ for *every* $i = 1, \cdots, n+3$; otherwise

(4.17) would be violated. Therefore we may assume that there is $1 < s \leqq n+3$ such that

$$\begin{aligned} &v_{ji} \text{ tend to some } v_i && \text{if } i = 1, \cdots, s, \\ &|v_{ji}| \to \infty && \text{if } i = s+1, \cdots, n+3, \end{aligned} \tag{4.18}$$

as $j \to \infty$. It follows from (4.17) that

$$0 \leqq \sum_{i=s+1}^{n+3} \beta_{ji} h(|v_{ji}|) = \eta_j' \leqq \eta_j(t).$$

On the other hand,

$$\eta_j' = \sum_{i=s+1}^{n+3} \beta_{ji} |v_{ji}| \frac{h(|v_{ji}|)}{|v_{ji}|}.$$

Two latter relations and (4.6), (4.18) show that

$$\beta_{ji} |v_{ji}| \to 0 \quad \text{if } j \to \infty \quad \forall i = s+1, \cdots, n+3.$$

In particular $\beta_{ji} \to 0$ for such i. We may suppose furthermore, in view of (4.14), that for $i = 1, \cdots, s$, the β_{ji} converge to some β_i and

$$\beta_i \geqq 0, \quad i = 1, \cdots, s; \quad \sum_{i=1}^{s} \beta_i = 1; \quad \sum_{i=1}^{s} \beta_i v_i = y(t), \tag{4.19}$$

which follows from (4.16), (4.18) because $v_j(t) \to y(t)$.

Finally, since f is nonnegative, l.s.c. in (x, y) and convex in y, we get, using (4.15), (4.16) and (4.19), that

$$\begin{aligned} \liminf_{j \to \infty} \lambda_j(t) &\geqq \liminf_{j \to \infty} \sum_{i=1}^{s} \beta_{ji} f(t, x_{ji}, v_{ji}) \\ &\geqq \sum_{i=1}^{s} \beta_i f(t, x(t), v_i) \geqq f(t, x(t), y(t)). \end{aligned}$$

This proves (4.13) and thereby the theorem.

5. Proof of the virtual measurability theorem. The first part of Theorem 2 can be restated in the following equivalent form.

THEOREM 2'. *Let S be a linear topological space of measurable mappings from G into R^s. Assume that S is decomposable, contains all bounded measurable mappings and the topology in S is not stronger than the topology of almost everywhere uniform convergence. Let $\varphi(t, z)$ be an integrand on $G \times R^s$ such that $|J(\hat{z}(\cdot))| < \infty$ for some $\hat{z}(\cdot) \in S$, where*

$$J(z(\cdot)) = \int_G \varphi(t, z(t))\, d\mu.$$

Assume that for any $z(\cdot) \in S$, $J(z(\cdot))$ makes sense (regardless, finite or not) and that the function $J(\cdot)$ is lower semicontinuous on S. Then φ is measurably equivalent to some other integrand $g(t, z)$ which is $\mathcal{L} \otimes \mathcal{B}$-measurable and l.s.c. in z.

First we shall prove the following result.

LEMMA. *Let S satisfy the assumptions of Theorem 2'. Let $\varphi(t, z)$ be an integrand on $G \times R^s$ such that $\varphi(t, \hat{z}(t)) < \infty$ a.e. on G for some $\hat{z}(\cdot) \in S$. Then there is another integrand $g(t, z)$ which is $\mathcal{L} \otimes \mathcal{B}$-measurable, l.s.c. in z and satisfies the following condition: for any $z(\cdot) \in S$*

(i) $g(t, z(t)) \leqq \varphi(t, z(t))$ *a.e.;*

(ii) *there exists a sequence $\{z_k(\cdot)\}$ of elements of S converging uniformly to $z(\cdot)$ and such that*

$$\limsup_{k \to \infty} \varphi(t, z_k(t)) \leqq g(t, z(t)) \quad \text{a.e.}$$

Proof. Consider the set

$$A = \{(\alpha(\cdot), w(\cdot)) | w(\cdot) \in S, \alpha(\cdot): G \to R \text{ is measurable,}$$
$$\alpha(t) \geqq \varphi(t, w(t)) \text{ a.e.}\}.$$

According to the assumptions, $A \neq \varnothing$. Therefore we can choose a countable collection B of elements if A which is dense in A with respect to the convergence in measure. Let

$$B(t) = \{(\alpha, w) \in R \times R^s | \alpha = \alpha(t), w = w(t) \text{ for some } (\alpha(\cdot), w(\cdot)) \in B\},$$

and define $g(t, z)$ as follows:

$$g(t, z) = \inf \{\alpha | (\alpha, z) \in \operatorname{cl} B(t)\},$$

where $\operatorname{cl} B(t)$ denotes the closure of $B(t)$ and $\inf \varnothing$ is by convention equal to ∞. The multivalued mapping $t \to \operatorname{cl} B(t)$ is close-valued and contains a countable dense collection of measurable selections. Therefore this multivalued mapping is measurable (see Rockafellar [14]) and $g(t, z)$ is $\mathcal{L} \otimes \mathcal{B}$-measurable integrand. Obviously, g is l.s.c. in z.

Fix some $z(\cdot) \in S$ and let

$$T = \{t \in G | \varphi(t, z(t)) < \infty\}.$$

Then (i) trivially holds if $t \notin T$. If $\mu T > 0$, we consider

$$z_0(t) = (1 - \chi_T(t))\hat{z}(t) + \chi_T(t) z(t).$$

Then $z_0(\cdot) \in S$, $\varphi(t, z_0(t)) < \infty$ a.e. and there is a measurable function $\alpha(\cdot): G \to R$ such that $(\alpha(\cdot), z_0(\cdot))$ belongs to A. Fix such an $\alpha(\cdot)$ and choose a sequence $\{\alpha_k(\cdot), w_k(\cdot)\}$ of elements of B converging a.e. to $\alpha(\cdot), z_0(\cdot))$. By definition,

$$g(t, z_0(t)) \leqq \liminf_{k \to \infty} \alpha_k(t) = \alpha(t) \quad \text{a.e.}$$

This is true for any measurable $\alpha(\cdot)$ such that $\alpha(t) \geqq \varphi(t, z_0(t))$ a.e. Therefore $g(t, z_0(t)) \leqq \varphi(t, z_0(t))$ a.e.; in particular $g(t, z(t)) \leqq \varphi(t, z(t))$ a.e. on T which proves (i).

It follows from the definition of g that for any $\delta>0, t\in G$ such that $g(t, z(t))<\infty$ there is an $(\alpha, w)\in B(t)$ such that

$$|z(t)-w|<\delta, \tag{5.1}$$

$$\varphi(t, w)\leqq\alpha\leqq\begin{cases} g(t, z(t))+\delta & \text{if } g(t, z(t))>-\infty, \\ -1/\delta & \text{if } g(t, z(t))=-\infty, \end{cases}$$

or in other words

$$\varphi(t, w)\leqq\max\{g(t, z(t))+\delta, -1/\delta\}. \tag{5.2}$$

Denote by W the projection of B onto S. Let $w_1(\cdot), w_2(\cdot), \cdots$ be an arbitrary numbering of elements of W. Consider the following sets:

$$T_{ki}=\{t\in G\,|\,|z(t)-w_i(t)|\leqq 1/k,$$
$$\varphi(t, w_i(t))\leqq\max\{g(t, z(t))+1/k, -k\}\}.$$

Let

$$w_{kj}(t)=\begin{cases} w_1(t) & \text{if } t\in T_{kl}, \\ \cdots\cdot & \\ w_k(t) & \text{if } t\in T_{kj}\setminus\bigcup_{i=1}^{j-1} T_{ki}, \\ \\ z(t) & \text{if } t\in G\setminus\bigcup_{i=1}^{j} T_{ki}. \end{cases}$$

According to (5.1), (5.2),

$$\bigcup_{i=1}^{\infty} T_{ki}\supset\{t\in G\,|\,g(t, z(t))<\infty\}$$

up to a set of measure zero. Choose $j(k)$ to ensure

$$\mu\left\{t\in G\,|\,g(t, z(t))<\infty, t\notin\bigcup_{i=1}^{j(k)} T_{ki}\right\}\leqq 1/k,$$

and let

$$z_k(t)=w_{kj(k)}(t).$$

It is easy to see that the $z_k(\cdot)$ satisfy (ii).

Proof of Theorem 2'. Define $g(t, z)$ as in the Lemma. Let $z(\cdot)\in S$. Then by lemma, $g(t, z(t))\leqq\varphi(t, z(t))$ a.e. Assume that

$$\mu\{t\in G\,|\,g(t, z(t))<\varphi(t, z(t))\}>0. \tag{5.3}$$

For every t belonging to the above set, $g(t, z(t))<\infty$ and $\varphi(t, z(t))>-\infty$. Therefore we can choose a set $T\subset G$ such that $\mu T>0$,

$$\int_{T'} g(t, z(t))\,d\mu<\int_{T'}\varphi(t, z(t))\,d\mu \quad \forall T'\subset T, \quad \mu T'>0, \tag{5.4}$$

and both integrals make sense.

Take a sequence $\{z_k(\cdot)\} \subset S$ satisfying condition (ii) of the lemma. Then it is possible (extracting if necessary, a subsequence) to find a set $T' \subset T$ with $\mu T' > 0$ such that

$$\liminf_{k\to\infty} \int_{T'} \varphi(t, z_k(t))\, d\mu \geqq \int_{T'} g(t, z(t))\, d\mu. \tag{5.5}$$

Let

$$w_k(t) = (1-\chi_{T'}(t))\hat{z}(t) + \chi_{T'}(t) z_k(t),$$
$$w(t) = (1-\chi_{T'}(t))\hat{z}(t) + \chi_{T'}(t) z(t).$$

Then $w_k(\cdot)$ and $w(\cdot)$ belong to S and $w_k(\cdot) \to w(\cdot)$ uniformly and hence in S. Since $J(\cdot)$ is l.s.c.,

$$\liminf_{k\to\infty} J(w_k(\cdot)) \geqq J(w(\cdot)). \tag{5.6}$$

But

$$J(w_k(\cdot)) = \int_{G\setminus T'} \varphi(t, \hat{z}(t))\, d\mu + \int_{T'} \varphi(t, z_k(t))\, d\mu,$$
$$J(w(\cdot)) = \int_{G\setminus T'} \varphi(t, \hat{z}(t))\, d\mu + \int_{T'} \varphi(t, z(t))\, d\mu.$$

The first terms in both sums are finite and equal. Hence (5.6) implies

$$\liminf_{k\to\infty} \int_{T'} \varphi(t, z_k(t))\, d\mu \geqq \int_{T'} \varphi(t, z(t))\, d\mu$$

which contradicts (5.4), (5.5).

Thus (5.3) is contradictory, and $g(t, z(t)) = \varphi(t, z(t))$ a.e. for any $z(\cdot) \in S$. But since S contains all bounded measurable mappings, the latter is true for any measurable $z(\cdot)$. This proves Theorem 2′ and hence the first part of Theorem 2.

The proof of the second part of Theorem 2 follows a rather routine scheme based on measurable selection techniques. If the assertion in the second part is false, then (since g is $\mathscr{L}\otimes\mathscr{B}$-measurable and l.s.c. in y) Aumann's selection theorem [1] (see also Sainte-Beuve [15]) allows us to find a measurable set $T \subset G$, $\mu T > 0$, and measurable mappings $x(\cdot): T \to R^m$, $u(\cdot): T \to R^n$ and $v(\cdot): T \to R^n$ such that

$$g(t, x(t), y(t)) < -\infty, \qquad g(t, x(t), u(t)) < -\infty, \tag{5.7}$$

$$\frac{g(t, x(t), u(t)) + g(t, x(t), v(t))}{2} < g\left(t, x(t), \frac{u(t)+v(t)}{2}\right). \tag{5.8}$$

With no loss of generality we may assume that $x(\cdot)$, $u(\cdot)$ and $v(\cdot)$ are bounded. Let $\{\alpha_k(\cdot)\}$ be a sequence of measurable functions on G assuming only two values, 0 and 1, and converging weakly* in L_∞ to the function identically equal to 1/2. Take a set $T' \subset T$ of positive measure such that all functions in (5.7), (5.8) are integrable, not necessarily finitely but with the same inequalities valid for their integrals.

Take $\hat{x}(\cdot)\in L$, $\hat{y}(\cdot)\in M$ such that $I(\hat{x}(\cdot),\hat{y}(\cdot))$ is finite (such a pair exists according to the assumptions), and let

$$x'(t)=(1-\chi_{T'}(t))\hat{x}(t)+\chi_{T'}(t)x(t),$$
$$u'(t)=(1-\chi_{T'}(t))\hat{y}(t)+\chi_{T'}(t)u(t),$$
$$v'(t)=(1-\chi_{T'}(t))\hat{y}(t)+\chi_{T'}(t)v(t).$$

Then as we have shown

$$\frac{I(x'(\cdot),u'(\cdot))+I(x'(\cdot),v'(\cdot))}{2}<I\left(x'(\cdot),\frac{u'(\cdot)+v'(\cdot)}{2}\right).$$

Let

$$y_k(t)=(1-\alpha_k(t))u'(t)+\alpha_k(t)v'(t).$$

By (H_4), the $y_k(\cdot)$ converge to $y(\cdot)=(1/2)(u'(\cdot)+v'(\cdot))$. Since $I(\cdot,\cdot)$ is l.s.c. on $L\times M$, it follows that

$$\liminf_{k\to\infty} I(x'(\cdot),y_k(\cdot))\geqq I\left(x'(\cdot),\frac{u'(\cdot)+v'(\cdot)}{2}\right).$$

On the other hand, according to the definition of $\alpha_k(\cdot)$

$$g(t,x,y_k(t))=(1-\alpha_k(t))g(t,x,u(t))+\alpha_k(t)g(t,x,v'(t))$$

and since the $\alpha_k(\cdot)$ converge weakly* to $1/2$ in L_∞, we get

$$\lim_{k\to\infty} I(x'(\cdot),y_k(\cdot))=(1/2)(I(x'(\cdot),u'(\cdot))+I(x'(\cdot),v'(\cdot)).$$

It is easy to see that three latter relations are contradictory.

REFERENCES

[1] R. J. AUMANN, *Measurable utility and measurable choice theorems*, Proc. Colloque Internationale du CNRS (Aix-en-Provence, France 1967), Editions du CNRS, Paris, 1969, pp. 15–26.

[2] L. D. BERKOVITZ, *Lower semicontinuity of integral functionals*, Trans. Amer. Math. Soc., 192 (1974), pp. 51–57.

[3] L. CESARI, *Lower semicontinuity and lower closure theorems without seminormality conditions*, Anal. Math. Pura Appl., 98 (1974), pp. 381–397.

[4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators* 1. *General Theory*, John Wiley, New York, 1958.

[5] A. D. IOFFE, *B-spaces generated by convex integrands and multidimensional variational problems*, Dokl. Akad. Nauk SSSR, 195 (1970), pp. 1018–1021 = Soviet Math. Dokl., 11 (1970), pp. 1600–1604.

[6] V. I. KASIMIROV, *On lower semicontinuity of integrals of the calculus of variations*, Uspekhi Mat. Nauk, 69 (1956), pp. 125–130 (in Russian).

[7] M. A. KRASNOSEL'SKII, P. P. ZABREIKO, E. I. PUSTYL'NIK AND P. E. SOBOLEVSKII, *Linear Operators in Spaces of Summable Functions*, Nauka, Moscow, 1966.

[8] M. A. KRASNOSEL'SKII AND YA. B. RUTITZKII, *Convex Functions and Orlicz Spaces*, P. Noordhoff, Gröningen, the Netherlands, 1961.

[9] S. F. MOROSOV AND V. I. PLOTNIKOV, *On necessary and sufficient conditions for continuity and lower semicontinuity of functionals of the calculus of variations*, Mat. Sb., 57 (1962), pp. 265–280 (in Russian).

[10] C. OLECH, *Existence theory in optimal control problems—the underlying ideas*, International Conference on Differential Equations, Academic Press, New York, 1975, pp. 612–629.

[11] ———, *Weak lower semicontinuity of integral functionals*, J. Optimization Theory Appl., to appear.

[12] ———, *A necessary and sufficient condition for lower semicontinuity of certain integral functionals*, Mathematical Structures–Computational Mathematics–Mathematical Modeling, Sofia, 1975, pp. 373–379.

[13] B. T. POLJAK, *Semicontinuity of integral functionals and existence theorems on extremal problems*, Mat. Sb., 78 (1969), 65–84 = Math. USSR Sb., 7 (1969), pp. 59–77.

[14] R. T. ROCKAFELLAR, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

[15] M. F. SAINTE-BEUVE, *Sur la généralization d'un théorèm de section mesurable de von Neumann–Aumann*, Ibid., 17 (1974), pp. 112–129.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

PROBABILITY THEORY METHODS IN ZERO-SUM STOCHASTIC GAMES*

JEAN-MICHEL BISMUT†

Abstract. The purpose of this paper is to apply the methods of optimal stochastic control introduced by the author to a class of zero-sum stochastic games.

1. Introduction. The purpose of this paper is to apply the methods used by the author in his previous work on the optimal control of diffusions [1] to a wide class of stochastic games.

Let us consider two compact metrizable spaces U and U'.

f is a bounded function defined on $R^+ \times R^d \times U \times U'$ with values in R^d, Borel on $R^+ \times R^d$ for a given $(u, u') \in U \times U'$, and continuous on $U \times U'$ for $(t, x) \in R^+ \times R^d$. We consider the diffusion process

$$\begin{aligned} dx &= f(t, x_t, u(t, x_t), u'(t, x_t))\, dt + \sigma(t, x_t) \cdot d\beta \\ x_s &= x, \end{aligned} \tag{1.1}$$

where u and u' are Borel functions defined on $R^+ \times R^d$ with values in U and U'.

A is a Borel set in $R^+ \times R^d$.

T_A is the stopping time:

$$T_A = \inf \{t > s;\ (t, x_t) \in A\}. \tag{1.2}$$

p is a constant >0.

We consider the criterion

$$e^{ps} E \int_s^{T_A} e^{-pt} K(t, x_t, u(t, x_t), u'(t, x_t))\, dt, \tag{1.3}$$

where K is a bounded function satisfying the same assumptions as f.

The purpose of this paper is to prove the existence of a minimax couple of strategies where the minimum corresponds to u and the maximum to u'.

This problem has been considered by Friedman in [5], where partial differential equation techniques are used.

Friedman assumes that A^c may be written as $Q_T =]0, T[\times \Omega$ where Ω is an open domain of R^d, whose boundary is sufficiently regular. Differentiability assumptions are also done in [5] on the matrix $a = \sigma\sigma^*$.

Then Friedman proves in Theorem 3 of [5] that if H, X_1, X_2 are defined by

$$\begin{aligned} H(t, x, u, u', p) &= K(t, x, u, u') + \langle p, f(t, x, u, u') \rangle, \\ X_1(t, x, p) &= \inf_{u \in U} \sup_{u' \in U'} H(t, x, u, u', p), \\ X_2(t, x, p) &= \sup_{u' \in U'} \inf_{u \in U} H(t, x, u, u', p), \end{aligned} \tag{1.4}$$

* Received by the editors May 16, 1975, and in final revised form July 1, 1976.

† 191, rue d'Alésia 75014 Paris, France.

then under Isaac's condition $h = -X_1 = -X_2$, the partial differential equation

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sum_{i,j} a_{ij}(t,x) V_{x_i x_j} = h\left(t, x, V, \frac{\partial V}{\partial x}\right) \quad \text{on } Q_T, \tag{1.5}$$
$$V = 0 \quad \text{on }]0, T[\times \partial\Omega) \cup (\{T\} \times \bar{\Omega})$$

has one unique solution, which is the cost function of the game. V is then proved to be continuous on $\bar{Q}_T$, and $\nabla_x V$ is proved to be Hölder continuous on Q_T. Moreover, Theorem 4 of [5] shows that a solution of the game exists.

In this paper we consider the case where f and K may be written

$$f(t,x,u,u') = b(t,x,u) + b'(t,x,u'), \tag{1.6}$$
$$K(t,x,u,u') = L(t,x,u) + L'(t,x,u').$$

We do not have any differentiability assumptions on a, and we accept A to be any Borel set in $R^+ \times R^d$. The partial differential equation (1.5) is then not well defined.

A solution of the game is proved to exist by using the results of [1]. The main interest of the method is the use of the deep convex structure of some problems of stochastic control. Moreover, the method is extendable to games on processes other than diffusions [2].

Finally, we come back to the more general system (1.1)–(1.3), and derive conditions for existence of solutions of this game. The reader is referred to [5] for a precise comparison of theses results with the results of Friedman.

In a different framework, Duncan and Varaiya have examined in [3] the problem of existence in the general nonanticipating case, where f and K depend on the entire trajectory of x and where (u, u') are taken as nonanticipating functions of x. They also assume that K does not depend on (u, u'), and that the first equality holds in (1.6) with $b_i b'_i = 0 (i = 1, \cdots, d)$. They also have a convexity condition on $\{b(t,x,u) | u \in U\}$ and $\{b'(t,x,u') | u' \in U'\}$. In [4] Elliott has extended the results of Duncan and Varaiya and obtains an existence result under Isaac's condition. However, the techniques used in [3] and [4] are rather different from the method used here which applies more directly to Markov processes.

2. Definition of the problem. a is a continuous function defined on $R^+ \times R^d$ with values in $R^d \otimes R^d$ such that:

a is bounded,

a is positive definite.

b is a Borel bounded function defined on $R^+ \times R^d$ with values in R^d.

$Q^b_{(s,x)}$ is the unique measure on the space of continuous functions defined on R^+ with values in R^d, which is a solution of the martingale problem defined in [7], with (s, x) as the starting point.

$E^b_{(s,x)}$ is the expectation operator for $Q^b_{(s,x)}$.

σ is the positive square root of a.

p is a strictly positive constant.

A is a Borel set in $R^+ \times R^d$.

L is a real-valued bounded Borel function defined on $R^+ \times R^d$.

For $c=(b, L)$, we define the function V_c as

$$V_c(s,x)=e^{ps}E^b_{(s,x)}\int_s^{T_A} e^{-pt}L(t,x_t)\,dt. \tag{2.1}$$

K and K' are two bounded Borel set-valued mappings defined on $R^+\times R^d$ with nonempty compact values in $R^d\times R$.

$\mathscr{L}$ (resp. $\mathscr{L}'$) is the set of Lebesgue equivalence classes of the Borel selections of K (resp. K').

$\mathscr{L}$ (resp. $\mathscr{L}'$) has the topology $\sigma(L_\infty(R^+\times R^d), L_1(R^+\times R^d))$, which is metrizable.

DEFINITION 2.1. Problem R is defined as the search of

$$c_0=(b_0, L_0)\in\mathscr{L}$$

and

$$c'_0=(b'_0, L'_0)\in\mathscr{L}'$$

such that for $(c, c')\in\mathscr{L}\times\mathscr{L}'$,

$$V_{c_0+c'}\leqq V_{c_0+c'_0}\leqq V_{c+c'_0} \tag{2.2}$$

The relation between this formulation of problem R and the formulation given in § 1 is derived in the same way as in [1, Chap. IV, Part 1].

THEOREM 2.1. *Problem* R *has a solution.*

Proof. The next parts are devoted to the proof of the theorem.

3. The convex case. We define first two measures μ and ν on $R^+\times R^d$.

μ is a probability measure on $R^+\times R^d$ mutually absolutely continuous with the Lebesgue measure.

ν is the measure on $R^+\times R^d$ defined by

$$\nu(\varphi)=E^0_\mu e^{ps}\int_s^{T_A} e^{-pt}\varphi(t,x_t)\,dt. \tag{3.1}$$

ν is then absolutely continuous relative to the Lebesgue measure of $R^+\times R^d$, by Theorem 8.1 of [7].

We assume in this part that K and K' have convex values.

For $c\in\mathscr{L}$ (resp. $c'\in\mathscr{L}'$), we define Γ'_c (resp. $\Gamma_{c'}$) by

$$\Gamma'_c=\{c'\in\mathscr{L}';\ \forall(s,x)\in R^+\times R^d,\ V_{c+c'}(s,x)=\sup_{\tilde{c}'\in\mathscr{L}'} V_{c+\tilde{c}'}(s,x)\} \tag{3.2}$$

(resp.

$$\Gamma_{c'}=\{c\in\mathscr{L};\ \forall(s,x)\in R^+\times R^d,\ V_{c+c'}(s,x)=\inf_{\tilde{c}\in\mathscr{L}} V_{\tilde{c}+c'}(s,x)\}). \tag{3.2'}$$

PROPOSITION 3.1. Γ'_c *(resp.* $\Gamma_{c'}$*) has nonempty compact convex values.*

Proof. The nonemptiness and the compactness of Γ'_c (resp. $\Gamma_{c'}$) follow from [1, Thm. V-1]. Moreover, Theorems IV-5 and IV-8 of [1] applied to Part V of [1], prove that one can find a Borel function H_c (resp. $H_{c'}$) such that a necessary and sufficient condition for $c'=(b', L')$ to be in Γ'_c (resp. $c=(b, L)$ to be in $\Gamma_{c'}$) is that,

ν-a.e., the following relation holds:

$$(3.3)\quad L'(t,x)+\langle H_c(t,x),\sigma^{-1}(t,x)b'(t,x)\rangle=\max_{(\tilde{b}',\tilde{L}')\in K'(t,x)}\tilde{L}'+\langle H_c(t,x),\sigma^{-1}(t,x)\tilde{b}'\rangle$$

(resp.

$$(3.4)\quad L(t,x)+\langle H_{c'}(t,x),\sigma^{-1}(t,x)b(t,x)\rangle=\min_{(\tilde{b},\tilde{L})\in K(t,x)}\tilde{L}+\langle H_{c'}(t,x),\sigma^{-1}(t,x)\tilde{b}\rangle)$$

K and K' having convex values, the result follows. □

PROPOSITION 3.2. *The set-valued mapping defined on $\mathscr{L}\times\mathscr{L}'$ with values in $\mathscr{L}\times\mathscr{L}'$*

$$(3.5)\qquad (c,c')\to\Gamma_{c'}\times\Gamma'_c$$

is upper semicontinuous.

Proof. We have to prove that if

$$(c_n,c'_n)\to(c,c')$$

and if

$$(\tilde{c}_n,\tilde{c}'_n)\in\Gamma_{c'_n}\times\Gamma'_{c_n}\to(\tilde{c},\tilde{c}'),$$

then

$$(\tilde{c},\tilde{c}')\in\Gamma_{c'}\times\Gamma'_c.$$

We know that, for any $\gamma'\in\mathscr{L}'$,

$$(3.6)\qquad V_{c_n+\tilde{c}'_n}\geqq V_{c_n+\gamma'}.$$

By Theorem V-1 of [1], which proves the continuity of $c\to V_c$, we find that, for $\gamma'\in\mathscr{L}'$,

$$(3.7)\qquad V_{c+\tilde{c}'}\geqq V_{c+\gamma'}$$

and $\tilde{c}'\in\Gamma'_c$. Similarly, we find that $\tilde{c}\in\Gamma_{c'}$, and then

$$(\tilde{c},\tilde{c}')\in\Gamma_{c'}\times\Gamma'_c.\quad\square$$

We then have:

THEOREM 3.1. *Problem* R *has a solution.*

Proof. The set-valued mapping

$$(c,c')\to\Gamma'_{c'}\times\Gamma'_c$$

has nonempty compact convex values and is upper semicontinuous. By Kakutani's theorem, it has a fixed point (c_0,c'_0). This point has the property that

$$\text{if } \gamma'\in\mathscr{L}',\quad V_{c_0+c'_0}\geqq V_{c_0+\gamma'},$$

$$\text{if } \gamma\in\mathscr{L},\quad V_{c_0+c'_0}\leqq V_{\gamma+c'_0}.$$

It is then a solution of problem R. □

If (c_0,c'_0) is a solution of problem R, it is easy to check that, for $(s,x)\in R^+\times R^d$,

$$(3.8)\qquad V_{c_0+c'_0}(s,x)=\inf_{c\in\mathscr{L}}\sup_{c'\in\mathscr{L}'}V_{c+c'}(s,x)=\sup_{c'\in\mathscr{L}'}\inf_{c\in\mathscr{L}}V_{c+c'}(s,x).$$

The function $V_{c_0+c'_0}$ does not depend then on the particular solution of problem R which is considered.

DEFINITION 3.1. q is the function defined by

$$q(s,x)=\inf_{c\in\mathscr{L}}\sup_{c'\in\mathscr{L}'}V_{c+c'}(s,x)=\sup_{c'\in\mathscr{L}'}\inf_{c\in\mathscr{L}}V_{c+c'}(s,x). \tag{3.9}$$

THEOREM 3.2. *It is possible to find a Borel function H such that for*

$$(c_0,c'_0)=((b_0,L_0),(b'_0,L'_0))$$

to be a solution of problem R, *it is necessary and sufficient that, ν—a.e., the following relations hold*:

$$L_0(t,x)+\langle H(t,x),\sigma^{-1}(t,x)b_0(t,x)\rangle=\min_{(\tilde{b},\tilde{L})\in K(t,x)}\tilde{L}+\langle H(t,x),\sigma^{-1}(t,x)\tilde{b}\rangle, \tag{3.10}$$

$$L'_0(t,x)+\langle H(t,x),\sigma^{-1}(t,x)b'_0(t,x)\rangle=\max_{(b',L')\in K'(t,x)}\tilde{L}'+\langle H(t,x),\sigma^{-1}(t,x)b'\rangle. \tag{3.11}$$

Moreover, a choice of (c_0,c'_0) verifying (3.10)–(3.11) *ν-a.e., is possible.*

Proof. The method is the same as in [1, Thm. IV-5, Cor. of Thm. IV-7 and Thm. IV-8].

With the notations of [1], we know from [1, (5.28)], that for $c=(b,L)\in L_\infty(R^+\times R^d)$, we can find a Borel function H_c and an additive functional A^c such that:

$$V_c(t,x_t)=V_c(s,x_s)+\int_s^t(pV-L)(u,x_u)\,du+\int_s^t H_c(u,x_u)\cdot d\beta_u^b+\int_s^t dA_u^c \tag{3.12}$$

(by using the results of Annex 1 in [1] we cancel the term M_t-M_s in [1, (5.28)]).

If (c_0,c'_0) is a solution of Problem R,

$$V_{c_0+c'_0}=q.$$

We can then write

$$\begin{aligned}q(t,x_t)=q(s,x_s)&+\int_s^t(pq-(L_0+L'_0))(u,x_u)\,du\\&+\int_s^t H_{c_0+c'_0}(u,x_u)\cdot d\beta_u^{b_0+b'_0}+\int_u^t dA_u^{c_0+c'_0}\end{aligned} \tag{3.13}$$

By (3.13), $H_{c_0+c'_0}$ does not depend on a particular solution of the game (c_0,c'_0).

By reasoning as in [1, Cor. of Thm. IV-7], we take for H the fixed function $H_{c_0+c'_0}$, where (c_0,c'_0) is a solution of Problem R.

The result follows from [1, Thm. IV-5 and Thm. IV-8]. □

4. The general case. We now prove Theorem 1.1.

Proof. Let $\hat{K}(t,x)$ and $\hat{K}'(t,x)$ be the closed convex hulls of $K(t,x)$ and $K'(t,x)$.

$\hat{K}$ and $\hat{K}'$ are bounded Borel set-valued functions by Corollary 3.3 of [6]. They satisfy the assumptions of § 2.

Problem $\hat{R}$ associated to $\hat{K}$ and $\hat{K}'$ has a solution. Let H be the function defined in Theorem 3.2 associated with $\hat{R}$.

Then, as in [1, Chap. IV-5], it is possible to find Borel selections c of K and c' of K' such that (3.10) and (3.11) hold ν-a.e., because $K(t, x)$ and $\hat{K}(t, x)$ (resp. $K'(t, x)$ and $\hat{K}'(t, x)$) have the same extremal points.

By Theorem 3.2, (c, c') is a solution to problem $\hat{R}$. It is then seen immediately that it is also a solution to problem R. □

5. Extensions. By using the methods of [1], the previous results can be extended to criteria of the type

$$E^{b+b'}_{(s,x)} \int_s^{T_A} \exp -\left\{\int_s^t (m+m')(\sigma, x_\sigma)\, d\sigma\right\} (L+L')(t, x_t)\, dt, \tag{5.1}$$

where:

we ask (b, L, m) and (b', L', m') to be Borel selections of K and K' which are bounded Borel compact-valued functions from $R^+ \times R^d$ in $R^d \times R \times R^+$.

we can find $p > 0$ such that if $(b, L, m) \in K(t, x)$, $m > p$.

The results can be also extended to diffusions with boundary conditions [8] with the same methods.

Finally, in the time-homogeneous case, the solutions can also be taken to be time-homogeneous.

Another point of interest is to know if the previous methods apply to the more general systems (1.1)–(1.3).

DEFINITION 5.1. If V is a bounded finely continuous Borel function on $R^+ \times R^d$, and if h and H are Borel functions on $R^+ \times R^d$, we say that

$$\begin{aligned} \mathscr{L}^0 V &= h(t, x), \\ \partial V &= H \end{aligned} \tag{5.2}$$

if:

(a) $V(s, x) = 0$ if (s, x) is regular for A (i.e., $Q^0_{(s,x)}(T_A = 0) = 1$).

(b) For any $(s, x) \in R^+ \times R^d$,

$$l_{t<T_A} V(t, x_t) - l_{s<T_A} V(s, x_s) - \int_{s \wedge T_A}^{t \wedge T_A} (h + pV)(u, x_u)\, du \tag{5.3}$$

is a local martingale for $Q^0_{(s,x)}$.

(c) There is a predictable additive functional A such that

$$V(t, x_t) - V(s, x_s) = \int_s^t (h + pV)(u, x_u)\, du + \int_s^t dA_u + \int_s^t H(u, x_u) \cdot d\beta^0_u. \tag{5.4}$$

Let φ, ψ, h be defined by

$$\begin{aligned} \varphi(t, x, v) &= -\inf_{u \in U} \{L(t, x, u) + \langle v, b(t, x, u)\rangle\}, \\ \psi(t, x, v) &= \sup_{u' \in U'} \{L'(t, x, u') + \langle v, b'(t, x, u')\rangle\}, \\ h(t, x, v) &= \varphi(t, x, v) - \psi(t, x, v). \end{aligned} \tag{5.5}$$

Then Theorem 3.2 implies that for a bounded Borel finely continuous function V to be the cost function associated to the problem (1.6), it is necessary

and sufficient that

$$\mathscr{L}^0 V = h(t, x, \sigma^{-1}\partial V(t, x)). \tag{5.6}$$

Equation (5.6) is then obviously the weak extension of (1.5). For each $(t, x) \in R^+ \times R^d$, $h(t, x, v)$ is the difference of two bounded uniformly Lipschitz convex functions. As a function of v, $h(t, x, \cdot)$ has a very general form, because the difference of Lipschitz convex functions is dense in the space of continuous functions for the uniform convergence on compact sets of R^d.

Let us assume that in (1.4), $h = -X_1 = -X_2$. Then if h can be written as in (5.5), equation (5.6) has one unique bounded Borel finely continuous solution, by the results of §§ 1–4.

By proceeding as Friedman in [5], we know then that for each (t, x) in $R^+ \times R^d$

$$(u, u') \to L(t, x, u, u') + \langle \partial V(t, x), \sigma^{-1}(t, x) b(t, x, u, u') \rangle$$

has a saddle point.

By using a measurable selection theorem, it is then possible to derive an existence result for the nonseparable case.

REFERENCES

[1] J. M. BISMUT, *Théorie probabiliste du contrôle des diffusions*, Mem. Amer. Math. Soc., 4 (1976), no. 167.

[2] ———, *Control of jump processes and applications*, Bull. Soc. Math. France, to appear.

[3] T. DUNCAN AND P. VARAIYA, *On the existence of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.

[4] R. ELLIOTT, *The existence of value in stochastic differential games*, this Journal, 14 (1976), pp. 85–94.

[5] A. FRIEDMAN, *Stochastic differential games*, J. Differential Equations, 11, (1972), pp. 79–108.

[6] R. T. ROCKAFELLAR, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

[7] D. W. STROOCK AND S. R. S. VARADHAN, *Diffusion processes with continuous coefficients*, Comm. Pure and Appl. Math., 22 (1969), pp. 345–400, pp. 479–530.

[8] ———, *Diffusion processes with boundary conditions*, Ibid., 24 (1971), pp. 147–225.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

DUAL VARIABLE METRIC ALGORITHMS FOR CONSTRAINED OPTIMIZATION*

SHIH-PING HAN†

Abstract. We present a class of algorithms for solving constrained optimization problems. In the algorithm nonnegatively constrained quadratic programming subproblems are iteratively solved to obtain estimates of Lagrange multipliers and with these estimates a sequence of points which converges to the solution is generated. To achieve a superlinear rate of convergence the matrix appearing in the subproblem is required to be an approximate inverse of the Hessian of the Lagrangian. Some well-known variable metric updates such as the BFGS update are employed to generate the matrix and the resulting algorithm converges locally with a superlinear rate. When the penalty Lagrangian developed by Hestenes, Powell and Rockafellar is incorporated in the algorithm it turns out to be closely related to the recently developed method of multipliers. Unlike the method of multipliers, our algorithm takes only one step in the unconstrained minimization of the penalty Lagrangian. Besides, it possesses a superlinear rate of convergence even without requiring a penalty parameter going to infinity and therefore avoids the numerical instability so caused.

1. Introduction. The techniques for solving quadratic programming problems have been developed so extensively that it becomes feasible to deal with the general nonlinear programming problem by reducing it to a sequence of quadratic programming subproblems. Adopted by many authors [17], [18], [22], [23] and shown very effective, this approach allows us to approximate the nonlinear programming problem quadratically and affords an extension of Newton's and Newton-like methods to constrained optimization. Following this approach, we present in this work a class of algorithms in which we iteratively solve nonnegatively constrained quadratic programming subproblems to obtain estimates of Lagrange multipliers and with these estimates generate a sequence of points which converges to the solution. To achieve a superlinear rate of convergence the matrix appearing in the subproblem is required to be an approximate inverse of the Hessian of the Lagragian. We suggest variable metric updates to generate these matrices and justify our suggestion by showing that, when some well-known updates such as the BFGS update are employed in this context, the algorithm converges locally with a superlinear rate. The penalty Lagrangian developed by Hestenes [25], Powell [35] and Rockafellar [39] may also be incorporated into the algorithm to replace the ordinary Lagragian; the resulting algorithm turns out to be closely related to the recently developed multiplier method [8], [25], [35], [39], [40]—a very promising method which has lately attracted a great deal of attention. Unlike the multiplier method, our algorithm takes only one step in the unconstrained minimization of the penalty Lagrangian. Besides, it has a superlinear rate of convergence even without requiring a penalty parameter going to infinity and therefore avoids the numerical instability so caused.

* Received by the editors July 17, 1975, and in revised form August 30, 1976.

† Department of Computer Science, Cornell University, Ithaca, New York 14853. This research was supported in part by the National Science Foundation under Grants ENG 75-10486 and GJ 35292.

In § 2 we state the algorithm and compare it with the related results in the literature of nonlinear programming. Sufficient conditions for convergence of the algorithm and for superlinear rates of convergence are presented in §§ 3 and 4 respectively. In § 5 we embed the BFGS and some other updates into the algorithm and with the results obtained in §§ 3 and 4 we show that the algorithm converges locally with a superlinear rate. In § 6 the algorithm is modified by replacing the Lagrangian by a penalty Lagrangian in order to relax some assumptions in the convergence theorems. Some comments and computational results are contained in § 7.

We note here that all vectors are column vectors and a row vector will be indicated by superscript T. For convenience a column vector in R^{n+m+q} is sometimes written as (x, u, v). We use x^i to denote different vectors; i.e., x^1 and x^2. To avoid some cumbersome constants we restrict ourselves to the l_2 vector norm and operator norm and denote it by $\|\cdot\|$. An ε-neighborhood $N(x, \varepsilon)$ of a point x in R^n is the set $N(x, \varepsilon) = \{y \in R^n : \|y - x\| < \varepsilon\}$. We use $L(R^n)$ to indicate the set of $n \times n$ real matrices and write $f \in LC^2[x]$ if function f has Lipschitz continuous second-order derivatives in a neighborhood of x.

2. Algorithm. In this paper we consider the following nonlinear programming problem

$$\text{(P)} \qquad \begin{aligned} &\min_x && f(x) \\ &\text{subject to} && g(x) \leqq 0, \\ & && h(x) = 0 \end{aligned}$$

where f, g, and h are functions from R^n into R, R^m, and R^q respectively. The Lagrangian of problem (P) is the real-valued function $L(x, u, v) = f(x) + u^T g(x) + v^T h(x)$ defined on R^{n+m+q}, and a Kuhn–Tucker triple is a vector $z^* = (x^*, u^*, v^*)$ in R^{n+m+q} which satisfies the first-order Kuhn–Tucker conditions [30]. We define a quadratic programming problem $DQ(x, A)$

$$\begin{aligned} \min_{(u,v)} \quad & \tfrac{1}{2}(\nabla f(x) + \nabla g(x)u + \nabla h(x)v)^T A(\nabla f(x) + \nabla g(x)u \\ & + \nabla h(x)v) - u^T g(x) - v^T h(x) \\ & \text{subject to} \quad u \geqq 0 \end{aligned} \tag{2.1}$$

associated with any x in R^n and any A in $L(R^n)$.

DEFINITION 2.1. A vector $z = (\hat{x}, \hat{u}, \hat{v})$ in R^{n+m+q} is a z-solution of $DQ(x, A)$ if $(\hat{u}, \hat{v})$ is a Kuhn–Tucker point of $DQ(x, A)$ and

$$\hat{x} = x - A(\nabla f(x) + \nabla g(x)\hat{v} + \nabla h(x)\hat{v}). \tag{2.2}$$

It is noted that $DQ(x, A)$ has no constraint at all if (P) has no inequality constraint. Now we can state the algorithm as follows.

ALGORITHM.

Step 1. Start with an estimate of a Kuhn–Tucker triple $z^0 = (x^0, u^0, v^0)$ of problem (P) and an estimate of A_0 of the inverse of the Hessian of the Lagrangian.

Step 2. Set $k = 0$.

Step 3. Find a z-solution of $DQ(x^k, A_k)$ and call this z-solution $z^{k+1} = (x^{k+1}, u^{k+1}, v^{k+1})$. If there is more than one such z-solution, choose one which is closest to z^k.

Step 4. If $z^{k+1}=(x^{k+1}, u^{k+1}, v^{k+1})$ satisfies a prescribed convergence criterion, stop; otherwise, update A_{k+1} by some scheme, set $k=k+1$ and go to Step 3.

In the algorithm, with an estimate (u^{k+1}, v^{k+1}) of the Lagrange multipliers obtained from solving $DQ(x^k, A_k)$, we find a new point x^{k+1} by taking one step of a gradient method to minimize the Lagrangian $L(x, u^{k+1}, v^{k+1})$. When $A_k=\nabla_{xx}L(x^k, u^{k+1}, v^{k+1})^{-1}$, a Newton step is carried out. In this paper we are more interested in the variable metric way to generate the matrix A_k; for example, the very successful BFGS update in unconstrained optimization can be so employed here. It is perhaps worth mentioning that the updated matrix A_k is used to find not only x^{k+1} but also the multipliers (u^{k+1}, v^{k+1}).

By Dorn's duality theorem [30] and under the assumption that A is symmetric and positive definite, the quadratic program $DQ(x, A)$ is dual to the quadratic program

$$\begin{aligned} &\min_{s} && \nabla f(x)^T s+\tfrac{1}{2}s^T A^{-1}s \\ &\text{subject to} && g(x)+\nabla g(x)^T s\leqq 0, \\ &&& h(x)+\nabla h(x)^T s=0, \end{aligned} \tag{2.3}$$

which can be viewed as a quadratic approximation to problem (P) if A^{-1} is the Hessian of the Lagrangian. Some efficient algorithms [17], [18], [21], [22] based on (2.3) have been developed. However, our algorithm seems more promising since the subproblem (2.1) has only nonnegative constraints and no constraints at all if problem (P) has only equality constraints. Moreover, some unconstrained optimization updating schemes are more naturally incorporated in (2.1) than in (2.3).

The algorithm is related to the dual, feasible direction algorithm developed by Mangasarian [31]. But unlike it we do not require the generated points to be feasible for the original problem (P) and therefore never need an anti-zigzag procedure to avoid the jamming situation.

For solving $DQ(x^k, A_k)$ there are a number of effective methods in the extensive literature of quadratic programming. These include Beale's method [2], Wolfe's method [42], Ritter's method [38], Lemke's method [27] and the principal pivoting method [10], [11].

3. Convergence theorems. In this section we shall show that under suitable conditions the algorithm will generate a sequence of vectors in R^{n+m+q} which converge to a Kuhn–Tucker triple of problem (P). First, we define the following function $G(\tilde{x}, A, \cdot): R^{n+m+q}\to R^{n+m+q}$,

$$G(\tilde{x}, A, z)=\begin{bmatrix} A(\nabla f(\tilde{x})+\nabla g(\tilde{x})u+\nabla h(\tilde{x})v)+(x-\tilde{x}) \\ u_1(g_1(\tilde{x})+\nabla g_1(\tilde{x})^T(x-\tilde{x})) \\ \vdots \\ u_m(g_m(\tilde{x})+\nabla g_m(\tilde{x})^T(x-\tilde{x})) \\ h_1(\tilde{x})+\nabla h_1(\tilde{x})^T(x-\tilde{x}) \\ \vdots \\ h_q(\tilde{x})+\nabla h_q(\tilde{x})^T(x-\tilde{x}) \end{bmatrix} \tag{3.1}$$

which is associated with any $\tilde{x}$ in R^n and any A in $L(R^n)$. The function $G(\tilde{x}, A, \cdot)$ is related to the equalities of the Kuhn–Tucker conditions of problem $DQ(\tilde{x}, A)$ by the following lemma.

LEMMA 3.1. *If A in $L(R^n)$ is symmetric and $\hat{z} = (\hat{x}, \hat{u}, \hat{v})$ in R^{n+m+q} is a z-solution of $DQ(\bar{x}, A)$ then $G(\bar{x}, A, \hat{z}) = 0$.*

Proof. If $\hat{z} = (\hat{x}, \hat{u}, \hat{v})$ is a z-solution of $DQ(\bar{x}, A)$, then by Definition 2.1 $(\hat{u}, \hat{v})$ is a Kuhn–Tucker point of $DQ(\bar{x}, A)$. Hence there exists a vector w in R^m such that $w \geqq 0$ and

$$\nabla g(\bar{x})^T A(\nabla f(\bar{x}) + \nabla g(\bar{x})\hat{u} + \nabla h(\hat{x})\hat{v}) - g(\bar{x}) - w = 0, \tag{3.2}$$

$$\nabla h(\bar{x})^T A(\nabla f(\bar{x}) + \nabla g(\bar{x})\hat{u} + \nabla h(\bar{x})\hat{v}) - h(\bar{x}) = 0, \tag{3.3}$$

and for $i = 1, \cdots, m$, we have

$$w_i \hat{u}_i = 0. \tag{3.4}$$

It also follows from Definition 2.1 that

$$A(\nabla f(\bar{x}) + \nabla g(\bar{x})\hat{u} + \nabla h(\bar{x})\hat{v}) + (\hat{x} - \bar{x}) = 0. \tag{3.5}$$

Thus

$$w = -(g(\bar{x}) + \nabla g(\bar{x})^T(\hat{x} - \bar{x})),$$

which in conjunction with (3.4) implies that for $i = 1, \cdots, m$

$$\hat{u}_i(g_i(\bar{x}) + \nabla g_i(\bar{x})^T(\hat{x} - \bar{x})) = 0. \tag{3.6}$$

From (3.3) and (3.5) we have

$$h(\bar{x}) + \nabla h(\bar{x})^T(\hat{x} - \bar{x}) = 0 \tag{3.7}$$

which combines with (3.5) and (3.6) to lead to the desired result. □

COROLLARY 3.2. *Let A in $L(R^n)$ be symmetric and nonsingular and $\hat{z} = (\hat{x}, \hat{u}, \hat{v})$ be a z-solution of $DQ(\bar{x}, A)$. If $\hat{x} = \bar{x}$ then $\hat{z} = (\hat{x}, \hat{u}, \hat{v})$ is a Kuhn–Tucker triple of problem* (P).

Let $z^* = (x^*, u^*, v^*)$ be a Kuhn–Tucker triple of problem (P); the nonsingularity of $\nabla_z G(x^*, A, z^*)$ with $A = \nabla_{xx} L(x^*, u^*, v^*)^{-1}$ is essential for establishing our convergence theorems. To ensure such nonsingularity we need the following condition, which was first studied by Fiacco and McCormick [14] and has been called "the Jacobian uniqueness condition" [29]. However, we note that it is the nonsingularity of $\nabla_z G(x^*, A, z^*)$ that is really needed.

DEFINITION 3.3. A Kuhn–Tucker triple $z^* = (x^*, u^*, v^*)$ of problem (P) satisfies the Jacobian uniqueness condition if the following three conditions are simultaneously satisfied:

(a) $u_i^* > 0$ if $i \in I(x^*) = \{j : g_j(x^*) = 0\}$;

(b) the gradients $\{\nabla g_i(x^*)\}$ (all $i \in I(x^*)$), $\{\nabla h_j(x^*)\}$, $j = 1, \cdots, q$ are linearly independent;

(c) for every nonzero vector y satisfying $y^T \nabla g_i(x^*) = 0$ for all $i \in I(x^*)$ and $y^T \nabla h_j(x^*) = 0$, $j = 1, \cdots, q$, it follows that $y^T \nabla_{xx} L(z^*) y > 0$.

We note here that conditions (a) and (c) have also been called the strict complementarity condition and the second order sufficiency condition respectively.

Our convergence theorems also need the following two lemmas. The proof of the first one follows from the mean value theorem and appears in [21].

LEMMA 3.4. *If* $z^* = (x^*, u^*, v^*) \in R^{n+m+q}$ *and* f, g *and* $h \in LC^2[x^*]$, *then there exists a neighborhood* $N(x^*, \varepsilon)$ *and two positive numbers* $\bar{K}$ *and* $\hat{K}$ *such that for any* $\tilde{x}$ *and* $\hat{x}$ *in* $N(x^*, \varepsilon)$ *and any* $(\hat{u}, \hat{v})$ *in* R^{m+q} *we have*

$$\begin{aligned} &\|\nabla_x L(\tilde{x}, \hat{u}, \hat{v}) - \nabla_x L(\bar{x}, \hat{u}, \hat{v}) - \nabla_{xx} L(x^*, u^*, v^*)(\tilde{x} - \bar{x})\| \\ &\leqq (\bar{K} \max\{\|\tilde{x} - x^*\|, \|\bar{x} - x^*\|\} + \hat{K}\|(\hat{u}, \hat{v}) - (u^*, v^*)\|)\|\tilde{x} - \bar{x}\|. \end{aligned} \tag{3.8}$$

COROLLARY 3.5. *If all the assumptions of Lemma* 3.4 *hold and* $\nabla_{xx} L(z^*)$ *is nonsingular then there exist positive numbers* ε, η *and* ξ *such that whenever* $\bar{x}$, $\tilde{x} \in N(x^*, \varepsilon)$ *and* $(\hat{u}, \hat{v}) \in N((u^*, v^*), \varepsilon)$ *then*

$$\eta\|\bar{x} - \tilde{x}\| \leqq \|\nabla_x L(\tilde{x}, \hat{u}, \hat{v}) - \nabla_x L(\bar{x}, \hat{u}, \hat{v})\| \leqq \xi\|\bar{x} - \tilde{x}\|.$$

LEMMA 3.6. *If* f, g *and* $h \in LC^2[x^*]$ *and* $\nabla_{xx} L(z^*)$ *is nonsingular, then for any* $\alpha > 0$ *there exist two positive numbers* ε *and* δ *such that for any* x *in* R^n *and any* A *in* $L(R^n)$ *satisfying* $\|x - x^*\| \leqq \varepsilon$ *and* $\|A - \nabla_{xx} L(z^*)^{-1}\| \leqq \delta$ *it follows that*

$$\|A\nabla_x L(x, u^*, v^*) + (x^* - x)\| \leqq \alpha\|x^* - x\|.$$

Proof. Let $\alpha > 0$ be given and let

$$\lambda = \max\{\|\nabla_{xx} L(z^*)^{-1}\|, \|\nabla_{xx} L(z^*)\|\}. \tag{3.9}$$

Choose ε and δ such that

$$\delta < 1/2, \tag{3.10}$$

$$(\delta + \lambda)\left(\bar{K}\varepsilon + \frac{\lambda^2\delta}{1 - \lambda\delta}\right) \leqq \alpha \tag{3.11}$$

where $\bar{K}$ is the constant defined in Lemma 3.4. Since $\|A - \nabla_{xx} L(z^*)^{-1}\| \leqq \delta$, it follows from (3.9), (3.10) and the perturbation Lemma [26] that A is nonsingular and

$$\|A^{-1} - \nabla_{xx} L(z^*)\| \leqq \frac{\lambda^2\delta}{1 - \lambda\delta}. \tag{3.12}$$

Then

$$\begin{aligned} &\|A\nabla_{xx} L(x, u^*, v^*) + (x^* - x)\| \\ &\quad \leqq \|A\|\,\|\nabla_x L(x, u^*, v^*) + A^{-1}(x^* - x)\| \\ &\quad \leqq \|A\|\,\|\nabla_x L(x, u^*, v^*) - \nabla_{xx} L(z^*)(x^* - x)\| \\ &\qquad + \|A\|\,\|\nabla_{xx} L(z^*) - A^{-1}\|\,\|x^* - x\| \\ &\quad \leqq (\delta + \lambda)\left(\bar{K}\varepsilon + \frac{\lambda^2\delta}{1 - \lambda\delta}\right)\|x^* - x\| \quad \text{(by Lemma 3.4 and 3.12)} \\ &\quad \leqq \alpha\|x^* - x\|. \quad \text{(by 3.11).} \quad \square \end{aligned}$$

The following theorem guarantees that under suitable conditions the algorithm generates a better estimate of a Kuhn–Tucker point of problem (P) in each iteration.

THEOREM 3.7 *Let f, g and $h \in LC^2[x^*]$. If a Kuhn–Tucker triple $z^* = (x^*, u^*, v^*)$ of problem* (P) *satisfies the Jacobian uniqueness condition and $\nabla_{xx}L(z^*)$ is nonsingular, then for any $r \in (0, 1)$ there exist two positive numbers $\varepsilon(r)$ and $\delta(r)$ such that if $\|\tilde{z} - z^*\| \leqq \varepsilon(r)$ and A is a symmetric $n \times n$ matrix with $\|A - \nabla_{xx}L(z^*)^{-1}\| \leqq \delta(r)$ then a closest z-solution $\bar{z}$ of $DQ(\tilde{x}, A)$ to $\tilde{z}$ exists and $\|\bar{z} - z^*\| \leqq r\|\tilde{z} - z^*\|$.*

Proof. For any A in $L(R^n)$ let C_A in $L(R^{n+m+q})$ be defined as

$$C_A = \nabla_z G(x^*, A, z^*)$$

where G is defined in (3.1). Let C^* denote C_A when $A = \nabla_{xx}L(x^*)^{-1}$. Under the Jacobian uniqueness condition it can be shown that C^* is nonsingular.

Let $r \in (0, 1)$ be given; define

$$\lambda = \max\{\|C^{*-1}\|, \|\nabla g(x^*)\| + \|\nabla h(x^*)\|, \|\nabla_{xx}L(x^*)\|, \|\nabla_{xx}L(x^*)^{-1}\|\} \tag{3.13}$$

and

$$\tau = \frac{\lambda}{1-r}. \tag{3.14}$$

We first choose $\bar{\varepsilon} > 0$ such that the following conditions are satisfied:

(3.15)
(a) for any z and $\tilde{z}$ in $N(z^*, \bar{\varepsilon})$ and any A in $L(R^n)$ with $\|A\| \leqq r/\lambda^2 + \lambda$ we have $\|\nabla_z G(\tilde{x}, A, z) - C_A\| \leqq 1/(2\tau)$, and for all $i = 1, \cdots, m$
(b) $g_i(x^*) < 0$ implies $g_i(\tilde{x}) + g_i(\tilde{x})^T(x - \tilde{x}) < 0$,
(c) $u_i^* > 0$ implies $u_i > 0$.

Then choose $\varepsilon(r)$ and $\delta(r)$ to satisfy the following conditions, where for simplicity we write henceforth ε and δ for $\varepsilon(r)$ and $\delta(r)$ respectively.

(3.16)
(a) $\max\{\lambda\delta, \lambda^2\delta\} \leqq r$,
(b) $\varepsilon \leqq \bar{\varepsilon}/3$,
(c) for any $\tilde{z}$ in R^{n+m+q} and any A in $L(R^n)$ with $\|\tilde{z} - z^*\| \leqq \varepsilon$ and $\|A - \nabla_{xx}L(z^*)^{-1}\| \leqq \delta$ we have that $\|G(\tilde{x}, A, z^*)\| \leqq (r/(2\tau))\|\tilde{x} - x^*\|$.

The existence of such ε and δ follows from Lemma 3.6 and by observing that $u_i^* g_i(x^*) = 0$, $(i = 1, \cdots, m)$ and $h(x^*) = 0$.

Assume that a vector $\tilde{z} = (\tilde{x}, \tilde{u}, \tilde{v})$ in R^{n+m+q} and a symmetric A in $L(R^n)$ satisfy $\|\tilde{z} - z^*\| \leqq \varepsilon$ and $\|A - \nabla_{xx}L(x^*)^{-1}\| \leqq \delta$; then we have

$$\|C_A - C^*\| \leqq \|A - \nabla_{xx}L(z^*)^{-1}\|(\|\nabla g(x^*)\| + \|\nabla h(x^*)\|) \leqq \delta\lambda \leqq r < 1$$

(by 3.13 and 3.15(a)).

Hence by the nonsingularity of C^* and the perturbation Lemma [26], C_A is also nonsingular and

$$\|C_A^{-1}\| \leqq \frac{\lambda}{1-\lambda^2\delta} \leqq \frac{\lambda}{1-r} \leqq \tau \quad \text{(by 3.16 and 3.14)}. \tag{3.17}$$

Define the functions $S_{\tilde{x},A} : R^{n+m+q} \to R^{n+m+q}$, associated with $\tilde{x}$ and A as follows:

$$S_{\tilde{x},A}(z) = z - C_A^{-1}G(\tilde{x}, A, z).$$

For any z in $N(z^*, \bar{\varepsilon})$ we have that

$$\begin{aligned}\|\nabla_z S_{\tilde{x},A}(z)\| &= \|I - C_A^{-1}\nabla_z G(\tilde{x}, A, z)\| \\ &\leqq \|C_A^{-1}\|\,\|C_A - \nabla_z G(\tilde{x}, A, z)\| \\ &\leqq \tau\frac{1}{2\tau} \leqq \frac{1}{2} \quad \text{(by 3.15(a) and 3.17),}\end{aligned}$$

which implies that $S_{\tilde{x},A}$ is a contraction mapping in $N(z^*, \bar{\varepsilon})$. Since from (3.16(b)), (3.16(c)) and (3.17) we also have

$$\|z^* - S_{\tilde{x},A}(z^*)\| \leqq \tau\|G(\tilde{x}, A, z^*)\| \leqq \frac{\bar{\varepsilon}}{2},$$

thus, the contraction mapping theorem [28] implies that $S_{\tilde{x},A}$ has a unique fixed point, say $\bar{z}$, in $N(z^*, \bar{\varepsilon})$ which satisfies

$$\|\bar{z} - z^*\| \leqq 2\tau\|G(\tilde{x}, A, z^*)\| \leqq r\|\tilde{x} - x^*\| \leqq r\|\tilde{z} - z^*\|. \tag{3.18}$$

We now show that $\bar{z}$ is a z-solution of $DQ(\tilde{x}, A)$. Since $\bar{z}$ is the unique fixed point of $S_{\tilde{x},A}$ in $N(x^*, \bar{\varepsilon})$, $\bar{z}$ is the unique zero of $G(\tilde{x}, A\,\cdot)$ in $N(z^*, \bar{\varepsilon})$. Thus

$$A(\nabla f(\tilde{x}) + \nabla g(\tilde{x})\bar{u} + \nabla h(\tilde{x})\bar{v}) + (\bar{x} - \tilde{x}) = 0, \tag{3.19}$$

and for $i = 1, \cdots, m$,

$$\bar{u}_i(g_i(\tilde{x}) + \nabla g_i(\tilde{x})^T(\bar{x} - \tilde{x})) = 0, \tag{3.20}$$

and for $j = 1, \cdots, q$,

$$h_j(\tilde{x}) + \nabla h_j(\tilde{x})^T(\bar{x} - \tilde{x}) = 0. \tag{3.21}$$

By (3.15(b)) and (3.15(c)) in the choice of $\bar{\varepsilon}$ and also by (3.20) we have

$$\bar{u} \geqq 0 \tag{3.22}$$

and

$$g(\tilde{x}) + \nabla g(\tilde{x})^T(\bar{x} - \tilde{x}) \leqq 0. \tag{3.23}$$

If w in R^m is defined by

$$w = -(g(\tilde{x}) + \nabla g(\tilde{x})^T(\bar{x} - \tilde{x})) \tag{3.24}$$

then clearly,

$$w \geqq 0. \tag{3.25}$$

Premultiplying (3.19) by $\nabla g(\tilde{x})^T$ and taking (3.24) into account, we then have

$$\nabla g(\tilde{x})^T A(\nabla f(\tilde{x}) + \nabla g(\tilde{x})\bar{u} + \nabla h(\tilde{x})\bar{v}) - g(\tilde{x}) - w = 0. \tag{3.26}$$

Similarly, from (3.19) and (3.21) we can get

$$\nabla h(\tilde{x})^T A(\nabla f(\tilde{x}) + \nabla g(\tilde{x})\bar{u} + \nabla h(\tilde{x})\bar{v}) - h(\tilde{x}) = 0. \tag{3.27}$$

From (3.20) and (3.24) it is also clear that for $i = 1, \cdots, m$

$$\bar{u}_i w_i = 0,$$

which in conjunction with (3.26), (3.27), (3.22) and (3.25) imply that $(\bar{u}, \bar{v})$ is a Kuhn–Tucker point of $DQ(\tilde{x}, A)$ with Lagrange multiplier vector w. Therefore, it follows from (3.19) that $\bar{z}$ is a z-solution of $DQ(\tilde{x}, A)$.

We next show that $\bar{z}$ is the closest z-solution of $DQ(\tilde{x}, A)$ to $\tilde{z}$. Suppose that $\hat{z}$ is another z-solution of $DQ(\tilde{x}, A)$. Then by Lemma 3.1 we have that $\hat{z}$ is a zero of $G(\tilde{x}, A, \cdot)$. The uniqueness of the zero of $G(\tilde{x}, A, \cdot)$ in $N(z^*, \bar{\varepsilon})$ implies that $\|\hat{z} - z^*\| \geqq \bar{\varepsilon}$; hence

$$\|\hat{z} - \tilde{z}\| \geqq \|\hat{z} - z^*\| - \|z^* - \tilde{z}\| > \bar{\varepsilon} - \bar{\varepsilon}/3 = 2\bar{\varepsilon}/3 \quad \text{(by 3.16(b))}.$$

However,

$$\|\bar{z} - \tilde{z}\| \leqq \|\bar{z} - z^*\| + \|\tilde{z} - z^*\| \leqq (r+1)\|\tilde{z} - z^*\| \leqq 2\bar{\varepsilon}/3.$$

Therefore $\bar{z}$ is the closest z-solution of $DQ(\tilde{x}, A)$ to $\tilde{z}$ and the proof is complete. □

Since the quadratic programs $DQ(\tilde{x}, A)$ and $DQ(\tilde{x}, \frac{1}{2}(A + A^T))$ have identical Kuhn–Tucker points, Theorem 3.7 and Theorem 3.10 below are also true for a nonsymmetric A if $\bar{x}$ is generated by $\bar{x} = \tilde{x} - \frac{1}{2}(A + A^T)(\nabla f(\tilde{x}) + \nabla g(\tilde{x})\bar{u} + \nabla h(\tilde{x})\bar{v})$ rather than by (2.2).

We note here that a sequence $\{z^k\}$ converges Q-linearly to a point z^* if $\|z^{k+1} - z^*\| \leqq r\|z^k - z^*\|$ for some r in (0, 1); and it converges Q-superlinearly if $\|z^{k+1} - z^*\| \leqq \theta_k\|z^k - z^*\|$ with $\lim_{k\to\infty}\theta_k = 0$. Therefore, it follows from Theorem 3.7 that if the starting point is close to a solution and the sequence $\{A_k\}$ of matrices remains close to the inverse of the Hessian of the Lagrangian, then our algorithm will generate a sequence of points which converges to the solution with at least Q-linear rate. The following result is an immediate consequence of Theorem 3.7 and of the above remarks.

COROLLARY 3.8. *Let the assumptions of Theorem* 3.7 *hold and let* $\{j_k\}$ *be a subsequence of positive integers with* $j_k \leqq k$. *If* z^0 *is sufficiently close to* z^* *and* $\|A_k - \nabla_{xx}L(z^{j_k})^{-1}\| \leqq \alpha_k$ *where* $\{\alpha_k\}$ *is a sequence of nonnegative numbers bounded by a sufficiently small number, then the sequence* $\{z^k\}$ *generated by the algorithm exists and converges to* z^* *with at least a* Q*-linear rate. Furthermore, if* $j_k \to \infty$ *and* $\alpha_k \to 0$, *then* $\{z^k\}$ *converges to* z^* *with at least a* Q*-superlinear rate.*

A straightforward way to generate the matrices $\{A_k\}$ in the algorithm is by setting $A_k = \nabla_{xx}L(z^{j_k})^{-1} + \alpha_k I$ and $j_k = k$. When $\alpha_k = 0$ we obtain a Newton-type method which can be shown to possess a quadratic rate of convergence; for the equality constraint problem this method turns out to be similar to a method studied by Polyak [34].

Inequality (3.18) in the proof of Theorem 3.7 shows that we actually can get the sharper result $\|\bar{z} - z^*\| \leqq r\|\tilde{x} - x^*\|$, and thus the following corollary.

COROLLARY 3.9. *Let all the assumptions of Theorem* 3.7 *hold. Then for any* $r \in (0, 1)$ *there exist two positive numbers* $\varepsilon(r)$ *and* $\delta(r)$ *such that if* $\|\tilde{x} - x^*\| \leqq \varepsilon(r)$ *and* $\|A - \nabla_{xx}L(z^*)\| \leqq \delta(r)$ *then a unique Kuhn–Tucker point* $(\bar{u}, \bar{v})$ *of* $DQ(\tilde{x}, A)$ *in* $N((u^*, v^*), \varepsilon(r))$ *exists and* $\|(\bar{u}, \bar{v}) - (u^*, v^*)\| \leqq r\|\tilde{x} - x^*\|$. *Moreover,* $\bar{u}_i = 0$ *if* $u_i^* = 0$.

The result of Corollary 3.9 has nothing to do with the way we generate $\bar{x}$, and hence can be applied to establish the convergence theorem for some other methods in which x^{k+1} is generated by another way. Indeed, this corollary has been so used in [23].

To achieve convergence, according to Theorem 3.7 the matrices $\{A_k\}$ in the algorithm are required to remain close to $\nabla_{xx}L(z^*)^{-1}$. In the theorem below we give a sufficient condition which ensures such closeness and at the same time can be satisfied by some variable metric updates. This condition was first studied by Broyden, Dennis and Moré [7] for nonlinear system of equations and some techniques of their proof will be employed here. Throughout this work $\|\cdot\|'$ denotes any fixed matrix norm which may be different from $\|\cdot\|$.

THEOREM 3.10. *Let $z^* = (x^*, u^*, v^*)$ be a Kuhn–Tucker triple of problem* (P) *and let f, g and $h \in LC^2[x^*]$ and $\nabla_{xx}L(z^*)$ be nonsingular. Let the Jacobian uniqueness condition hold at z^* and let there exist two nonnegative numbers α_1 and α_2 such that for an update which generates symmetric matrices the following condition holds:*

$$\|A_{k+1}-\nabla_{xx}L(z^*)^{-1}\|' \leqq (1+\alpha_1\|z^k-z^*\|)\|A_k-\nabla_{xx}L(z^*)^{-1}\|' + \alpha_2\|z^k-z^*\|. \tag{3.28}$$

Then for any $r \in (0, 1)$ there exist two positive numbers $\varepsilon(r)$ and $\delta(r)$ such that if $\|z^0-z^*\| \leqq \varepsilon(r)$ and $\|A_0-\nabla_{xx}L(z^*)^{-1}\| \leqq \delta(r)$ then the sequence $\{z^k\}$ generated by the algorithm is well defined and converges Q-linearly to z^*. Furthermore, $\varepsilon(r)$ and $\delta(r)$ can be chosen small enough to ensure the nonsingularity of all the updated matrices $\{A_k\}$ and the uniform boundedness of $\{A_k^{-1}\}$.

Proof. By the equivalence of matrix norms, there exist two positive numbers d and d' such that for any A in $L(R^n)$ we have

$$d\|A\|' \geqq \|A\| \quad \text{and} \quad d'\|A\| \geqq \|A\|'. \tag{3.29}$$

Let $r \in (0, 1)$ be given. By Theorem 3.7 there exist two positive numbers $\bar{\varepsilon}$ and $\bar{\delta}$ such that if $\|\bar{z}-z^*\| \leqq \bar{\varepsilon}$ and $\|A-\nabla_{xx}L(z^*)^{-1}\| \leqq \bar{\delta}$ then the closest z-solution $\hat{z}$ of $DQ(\bar{x}, A)$ exists and $\|\hat{z}-z^*\| \leqq r\|\bar{z}-z^*\|$. We choose two positive numbers ε and δ such that the following conditions hold:

$$\begin{aligned} &\text{(a)}\ \ \varepsilon \leqq \bar{\varepsilon}, \\ &\text{(b)}\ \ 2dd'\delta \leqq \bar{\delta}, \\ &\text{(c)}\ \ (2\alpha_1\delta d'-\alpha_2)\frac{\varepsilon}{1-r} \leqq d'\delta. \end{aligned} \tag{3.30}$$

If we can show that for each k

$$\|z^k-z^*\| \leqq r^k\varepsilon \tag{3.31}$$

and

$$\|A_k-\nabla_{xx}L(z^*)^{-1}\|' \leqq 2d'\delta, \tag{3.32}$$

then $\|z^k-z^*\| \leqq \bar{\varepsilon}$ and $\|A_k-\nabla_{xx}L(z^*)^{-1}\| \leqq \bar{\delta}$, and the theorem follows immediately from Theorem 3.7.

We prove (3.31) and (3.32) by induction. They are obviously true for $k = 0$. Assume that they are true for j, $0 \leqq j \leqq k$; then it follows from (3.28) that

$$\|A_{j+1}-\nabla_{xx}L(z^*)^{-1}\|' - \|A_j-\nabla_{xx}L(z^*)^{-1}\|' \leqq 2\alpha_1 d'\varepsilon\delta r^j + \alpha_2\varepsilon r^j,$$

and by taking the sum from $j=0$ to $j=k$,

$$\|A_{k+1}-\nabla_{xx}L(z^*)^{-1}\|' \leqq \|A_0-\nabla_{xx}L(z^*)^{-1}\|' + (2\alpha_1 d'\delta+\alpha_2)\frac{\varepsilon}{1-r}$$
$$\leqq d'\delta + d'\delta \leqq 2d'\delta.$$

Therefore, (3.32) is true for $j=k+1$. Moreover, we have $\|A_{k+1}-\nabla_{xx}L(z^*)^{-1}\| \leqq \bar{\delta}$, and by the induction hypothesis and (3.30(c)) we have $\|z^k-z^*\| \leqq r^k\varepsilon \leqq \bar{\varepsilon}$. Thus, it follows from Theorem 3.7 that z^{k+1} exists and $\|z^{k+1}-z^*\| \leqq r\|z^k-z^*\| \leqq r^{k+1}\varepsilon$.

The second part of the theorem follows directly from (3.32) and the perturbation lemma. □

4. Superlinear rate of convergence. In this section sufficient conditions are given which guarantee a superlinear rate of convergence for the sequence of points generated by the algorithm. We first introduce a lemma which is due to Mangasarian [32] and is closely related to a result of Dennis and Moré [13]; its proof can be found in [22].

LEMMA 4.1. *Let z^* be a Kuhn–Tucker triple of problem (P) satisfying the Jacobian uniqueness condition and f, g and $h \in LC^2[x^*]$. A sequence $\{z^k\}$ converges Q-superlinearly to z^* if $\{z^k\}$ converges to z^* and*

$$\lim_{k\to\infty}\frac{\|E(z^{k+1})\|}{\|z^{k+1}-z^k\|}=0 \tag{4.1}$$

where $E: R^{n+m+q} \to R^{n+m+q}$,

$$E(z)=[\nabla_x L(z)^T, u_1 g_1(x), \cdots, u_m g_m(x), h_1(x), \cdots, h_q(x)]^T.$$

In Lemma 4.1 above there is no specification on how the sequence $\{z^k\}$ is generated. Nevertheless, if the sequence is generated by our algorithm then we can derive from Lemma 4.1 some other criteria, which are contained in the following two theorems.

THEOREM 4.2. *Let all the assumptions of Lemma 4.1 hold. If a sequence $\{z^k\}$ constructed by the algorithm converges to z^* and*

$$\lim_{k\to\infty}\frac{\|\nabla_x L(z^{k+1})\|}{\|z^{k+1}-z^k\|}=0, \tag{4.2}$$

then $\{z^k\}$ converges Q-superlinearly to z^*.

Proof. By Lemma 4.1 we need only to establish (4.1). By virtue of (3.6) and (3.7) in the proof of Lemma 3.1 we have

$$\|E(z^{k+1})\| \leqq \|\nabla_x L(x^{k+1}, u^{k+1}, v^{k+1})\| + \|\nabla h(x^{k+1})\| + \sum_{i=1}^{m}|u_i^{k+1}g_i(x^{k+1})|$$
$$\leqq \|\nabla_x L(x^{k+1}, u^{k+1}, v^{k+1})\| + \|h(x^{k+1})-h(x^k)-\nabla h(x^k)^T(x^{k+1}-x^k)\|$$
$$+ \sum_{i=1}^{m}|u_i^{k+1}(g_i(x^{k+1})-g_i(x^k)-\nabla g_i(x^k)^T(x^{k+1}-x^k))|$$
$$\leqq \|\nabla_x L(x^{k+1}, u^{k+1}, v^{k+1})\| + o(\|x^{k+1}-x^k\|).$$

Hence (4.1) follows. □

THEOREM 4.3. *Let all the assumptions of Lemma* 4.1 *hold; furthermore, let* $\nabla_{xx}L(z^*)$ *be nonsingular and let* $\{z^k\}$ *be a sequence of points generated by the algorithm with respect to a sequence of nonsingular symmetric matrices* $\{A_k\}$ *with* $\{A_k^{-1}\}$ *uniformly bounded. If* $\{z_k\}$ *converges to* z^*, *and*

$$\lim_{k\to\infty}\frac{\|(A_k-\nabla_{xx}L(z^*)^{-1})y^k\|}{\|y^k\|}=0 \tag{4.3}$$

where $y^k=\nabla_x L(x^{k+1}, u^{k+1}, v^{k+1})-\nabla_x L(x^k, u^{k+1}, v^{k+1})$, *then* $\{z^k\}$ *converges Q-superlinearly to* z^*.

Proof. By the assumptions x^{k+1} is a z-solution and

$$A_k\nabla_x L(x^k, u^{k+1}, v^{k+1})+(x^{k+1}-x^k)=0,$$

which yields

$$(A_k-\nabla_{xx}L(z^*)^{-1})y^k=A_k\nabla_x L(z^{k+1})+(x^{k+1}-x^k)-\nabla_{xx}L(z^k)^{-1}y^k$$

and in turn implies

$$\nabla_x L(z^{k+1})=A_k^{-1}(A_k-\nabla_{xx}L(z^*)^{-1})y^k+A_k^{-1}(\nabla_{xx}L(z^*)^{-1}y^k-(x^{k+1}-x^k)).$$

By the uniform boundedness of $\{A_k^{-1}\}$ there exists $\lambda>0$ such that

$$\begin{aligned}\|\nabla_x L(z^{k+1})\|\leqq\lambda\|(A_k-\nabla_{xx}L(z^*)^{-1})y^k\|+\lambda\|\nabla_{xx}L(z^*)^{-1}\|\\ \|y^k-\nabla_{xx}L(z^*)(x^{k+1}-x^k)\|.\end{aligned} \tag{4.4}$$

On the other hand, since f, g and $h\in LC^2[x^*]$ and $z^k\to z^*$, there exists some $\alpha>0$ such that

$$\|y^k\|\leqq\alpha\|x^{k+1}-x^k\|\leqq\alpha\|z^{k+1}-z^k\|. \tag{4.5}$$

By (4.3), (4.4) and Lemma 3.4, taking (4.5) into account, we get (4.2) and complete the proof. □

5. Updates. The updates which we consider for generating matrices in the algorithm are of the following form

$$\bar{A}=A+\frac{(s-Ay)d^T+d(s-Ay)^T}{d^Ty}-\frac{y^T(s-Ay)dd^T}{(d^Ty)^2} \tag{5.1}$$

where $\bar{A}=A_{k+1}$, $A=A_k$, $s=x^{x+1}-x^k$ and $y=\nabla_x L(x^{k+1}, u^{k+1}, v^{k+1})-\nabla_x L(x^k, u^{k+1}, v^{k+1})$ and d is any vector in R^n with $y^Td\neq 0$. A particular algorithm is determined once d is specified. The algorithm will be called Algorithm D1 when d = s and Algorithm D2 when $d=y$. Thus, we have the following updates:

$$\bar{A}=A+\frac{(s-Ay)s^T+s(s-Ay)^T}{s^Ty}-\frac{y^T(s-Ay)ss^T}{(s^Ty)^2}, \tag{5.2}$$

$$\bar{A}=A+\frac{(s-Ay)y^T+y(s-Ay)^T}{y^Ty}-\frac{y^T(s-Ay)yy^T}{(y^Ty)^2}. \tag{5.3}$$

These updates are well known in unconstrained optimization where y is defined as $y=\nabla f(x^{k+1})-\nabla f(x^k)$. In this context update (5.2) has been studied by Broyden [6], Fletcher [16], Goldfarb [19] and Shanno [41], and is often referred to as the BFGS update. Update (5.3) is one of Greenstadt's methods [20].

In the sequel we establish superlinear convergence theorems for Algorithms D1 and D2 by utilizing the techniques developed by Broyden, Dennis and Moré [7] and Dennis and Moré [13]. It is noted here that for any nonsingular $n \times n$ matrix M the matrix norm $\|\cdot\|_M$ is defined in such a way that for any $n \times n$ matrix A

$$\|A\|_M = \text{trace}\,[(MAM)^T(MAM)].$$

We state two lemmas below; their proofs are in [13] and [7] respectively.

LEMMA 5.1. *Let $\{a_k\}$ and $\{b_k\}$ be sequences of nonnegative numbers and $\alpha_1 \geqq 0$, $\alpha_2 \geqq 0$ such that*

$$a_{k+1} \leqq (1+\alpha_1 b_k)a_k + \alpha_2 b_k$$

and

$$\sum_{k=1}^{\infty} b_k < \infty;$$

then $\{a_k\}$ converges.

LEMMA 5.2. *Let A be any $n \times n$ symmetric matrix and s, d and y be vectors in R^n with $d^T y \neq 0$ and define $\bar{A}$ by* (5.1). *If M is a nonsingular symmetric $n \times n$ matrix with*

$$\|Md - M^{-1}y\| \leqq \beta \|M^{-1}y\| \tag{5.4}$$

for some $\beta \in [0, 1/3]$, then for any symmetric $n \times n$ matrix B with $B \neq A$ we have

$$\|\bar{A} - B\|_M \leqq \left((1-\lambda\theta^2)^{1/2} + \lambda_1 \frac{\|Md - M^{-1}y\|}{\|M^{-1}y\|}\right)\|A - B\|_M + \lambda_2 \frac{\|s - By\|}{\|y\|} \tag{5.5}$$

where $\lambda \in (0, 1)$, and λ_1 and λ_2 are constants which only depend on M and n, and

$$\theta = \frac{\|M(A-B)y\|}{\|A-B\|_M \|M^{-1}y\|} \tag{5.6}$$

if $A \neq B$ and $\theta = 0$ otherwise.

The following theorem gives a sufficient condition for the superlinear convergence of the algorithm with an update of form (5.1).

THEOREM 5.3. *Let $z^* = (x^*, u^*, v^*)$ be a KuhnTucker triple of problem* (P) *satisfying the Jacobian uniqueness condition and f, g and $h \in LC^2[x^*]$. Suppose that $\nabla_{xx}L(x^*)$ is nonsingular and in the algorithm the matrices $\{A_k\}$ are updated by formula* (5.1) *with any d^k such that for $y^k \neq 0$,*

$$\frac{\|Md^k - M^{-1}y^k\|}{\|M^{-1}y^k\|} \leqq \mu \max\{\|z^k - z^*\|, \|z^{k+1} - z^*\|\} \tag{5.7}$$

for a constant μ and an arbitrary but fixed nonsingular symmetric matrix M. If z^0 and A_0 are sufficiently close to z^* and $\nabla_{xx}L(z^*)^{-1}$ respectively then the sequence $\{z^k\}$ generated by the algorithm is well defined and converges Q-superlinearly to z^*.

Proof. For any $r \in (0, 1)$ let $\varepsilon(r)$ and $\delta(r)$ be defined as in Theorem 3.10 with matrix norm $\|\cdot\|'$ as $\|\cdot\|_M$. Now set $\alpha_1 = \lambda_1$, $\alpha_2 = (\lambda_2/\eta)(\bar{K} + \tilde{K})\|\nabla_{xx}L(z^*)^{-1}\|$ where $\bar{K}$ and $\tilde{K}$ are the constants defined in Lemma 3.4 and λ_1 and λ_2 are as in Lemma 5.2 and η is as in Corollary 3.5. We further require $\varepsilon(r)$ to satisfy

$$\varepsilon(r) \leqq \tfrac{1}{3}\mu. \tag{5.8}$$

We first show by induction that if $\|z^0 - z^*\| \leqq \varepsilon(r)$ and $\|A_0 - \nabla_{xx} L(z^*)^{-1}\| \leqq \delta(r)$ then the generated sequence $\{z^k\}$ exists and converges Q-linearly to z^*; that is,

$$\|z^{j+1} - z^*\| \leqq r\|z^j - z^*\|. \tag{5.9}$$

When $j = 0$ the existence of z^{j+1} and (5.9) follows directly from Theorem 3.7 and the choice of $\varepsilon(r)$ and $\delta(r)$. Assume z^{j+1} exists for all $j \leqq k$ and that (5.9) holds. We show that z^{k+2} exists and that (5.9) is also true for $j = k+1$. Assume $y^k \neq 0$, for if $y^k = 0$ then Corollary 3.5 implies that $s^k = 0$ which by Corollary 3.2 in turn implies that $z^{k+1} = (x^{k+1}, u^{k+1}, v^{k+1})$ is a Kuhn–Tucker triple of (P). On the other hand the Jacobian uniqueness condition yields that z^* is the unique Kuhn–Tucker triple in $N(z^*, \varepsilon(r))$ and hence we have $z^{k+1} = z^*$. Therefore, in case $y^k = 0$ the sequence $\{z^k\}$ converges to z^* in a finite number of steps. When $y^k \neq 0$ it follows from (5.7) and (5.8) that

$$\frac{\|Md^k - M^{-1}y^k\|}{\|M^{-1}y^k\|} \leqq \mu \max\{\|z^k - z^*\|, \|z^{k+1} - z^*\|\}$$

$$< \mu\varepsilon(r) \leqq \frac{1}{3}.$$

Let $B = \nabla_{xx} L(z^*)^{-1}$; then by Lemma 5.2 we have

$$\begin{aligned} \|A_{k+1} - B\|_M \leqq &((1 - \lambda\theta_k^2)^{1/2} + \lambda_1\mu\|z^k - z^*\|)\|A_k - B\|_M \\ &+ \lambda_2 \frac{\|s^k - By^k\|}{\|y^k\|} \end{aligned} \tag{5.10}$$

where

$$\theta_k = \frac{\|M(A_k - B)y^k\|}{\|A_k - B\|_M \|M^{-1}y^k\|}.$$

Lemma 3.4 yields

$$\|s^k - By^k\| \leqq \|B\|(\bar{K} + \tilde{K})\|z^k - z^*\|\,\|s^k\|$$

and Corollary 3.5 implies that for some $\eta > 0$, $\eta\|s^k\| \leqq \|y^k\|$. Therefore,

$$\frac{\|s^k - By^k\|}{\|y^k\|} \leqq \frac{1}{\eta}\|B\|(\bar{K} + \tilde{K})\|z^k - z^*\|,$$

which in conjunction with (5.10) and the fact that $(1 - \lambda\theta^2)^{1/2} \leqq 1 - (\lambda/2)\theta^2$ implies

$$\|A_{k+1} - B\|_M \leqq \left(1 - \frac{\lambda}{2}\theta_k^2 + \alpha_1\|z^k - z^*\|\right)\|A_k - B\|_M + \alpha_2\|z^k - z^*\|. \tag{5.11}$$

Hence the existence of z^{k+2} and $\|z^{k+2} - z^*\| \leqq r\|z^{k+1} - z^*\|$ follow from Theorem 3.10 and (5.11) immediately.

So far we have shown that the sequence $\{z^k\}$ exists and converges to z^* with at least a Q-linear rate; we are going to prove that the rate of convergence is

actually Q-superlinear. By Lemma 5.1 and (5.11) the sequence $\{\|A_k - B\|_M\}$ has a limit, say p. If $p = 0$ then the desired result follows directly from Theorem 4.3. Assume $p \neq 0$. It follows from (5.11) that

$$\frac{\lambda}{2}\theta_k^2 p_k \leqq p_k - p_{k+1} + (\alpha_1 p_k + \alpha_2)\|z^k - z^*\|$$

where $p_k = \|A_k - B\|_M$. By taking the sum of both sides over $k = 0, 1, 2, \cdots$ and taking the Q-linear convergence of $\{\|z^k - z^*\|\}$ and the boundedness of $\{p_k\}$ into consideration we have $\sum_{i=0}^{\infty} \theta_k^2 p_k < \infty$. Since $\theta \in (0, 1)$ and $p_k \to p$ with $p \neq 0$, we must have $\lim_{k\to\infty}\theta_k = 0$ which implies

$$\lim_{k\to\infty} \frac{\|(A_k - \nabla_{xx}L(z^*)^{-1})y^k\|}{\|y^k\|} = 0.$$

Hence the result also follows from Theorem 4.3. □

Our main results are contained in the following theorem which shows that Algorithms D1 and D2 possess local superlinear convergence properties.

THEOREM 5.4. *Let $z^* = (x^*, u^*, v^*)$ be a Kuhn–Tucker triple of* (P) *satisfying the Jacobian uniqueness condition and f, g and $h \in LC^2[x^*]$. If $\nabla_{xx}L(z^*)$ is nonsingular and the starting point z^0 and the starting matrix A_0 are sufficiently close to z^* and $\nabla_{xx}L(z^*)^{-1}$ respectively then the sequence $\{z^k\}$ generated by Algorithm* D2 *exists and converges Q-superlinearly to z^*. If $\nabla_{xx}L(z^*)$ is further assumed to be positive definite then the conclusion is also true for Algorithm* D2.

Proof. By Theorem 5.3 it is sufficient to establish (5.7) for some suitable matrix M and constant μ. Since it is obviously true for Algorithm D2, we only need to verify it for Algorithm D1. With $\nabla_{xx}L(z^*)$ positive definite we can set $M = (\nabla_{xx}L(z^*))^{1/2}$. By Lemma 3.4 we have

$$\begin{aligned}\|Ms^k - M^{-1}y^k\| &\leqq \|M^{-1}\|\,\|y^k - \nabla_{xx}L(z^*)s^k\| \\ &\leqq \|M^{-1}\|(\bar{K} + \tilde{K}) \max\{\|z^k - z^*\|,\ \|z^{k+1} - z^*\|\}\|s^k\|.\end{aligned}$$

and by Corollary 3.5 we have that for some $\xi > 0$, $\|s^k\| \leqq \xi\|y^k\|$. Therefore,

$$\|Ms^k - M^{-1}y^k\| \leqq \|M^{-1}\|(\bar{K} + \tilde{K}) \max\{\|z^k - z^*\|,\ \|z^{k+1} - z^*\|\}\xi\|M\|\,\|M^{-1}y^k\|.$$

Thus (5.7) is true with $\mu = \xi\|M\|\,\|M^{-1}\|(\bar{K} + \tilde{K})$. □

We note here that local superlinear convergence can also be achieved for the algorithm if the following updates are used [21]:

$$\bar{A} = A - \frac{Ayy^TA}{y^TAy} + \frac{ss^T}{s^Ty}, \tag{5.12}$$

$$\bar{A} = A + \frac{(s - Ay)s^T + s(s - Ay)^T}{2x^Ty}, \tag{5.13}$$

$$\bar{A} = A + \frac{(s - Ay)y^T + y(s - Ay)^T}{2y^Ty}. \tag{5.14}$$

The nonsymmetric updates such as

$$A = A + \frac{(s - Ay)s^T}{s^T y}, \tag{5.15}$$

$$\bar{A} = A + \frac{(s - Ay)y^T}{y^T y} \tag{5.16}$$

can also be shown to possess Q-linear rates of convergence; however, we have not succeeded in establishing Q-superlinear rates for them, though such results are predicted. In unconstrained optimization (5.12) is the famous Daviden–Fletcher–Powell update and (5.15) and (5.16) have been studied by Broyden [5] and McCormick [33] respectively.

6. Modification via a penalty Lagrangian. Considerable attention has been given recently to a penalty Lagrangian developed by Hestenes [25], Powell [35] and Rockafellar [39]. This function, $F: R^{n+m+q+1} \to R$, is defined by

$$\begin{aligned} &F(x, u, v, \alpha) \\ &= f(x) + \frac{1}{2\alpha} \sum_{i=1}^{m} ((\alpha g_i(x) + u_i)_+^2 - u_i^2) + v^T h(x) + \frac{\alpha}{2} h(x)^T h(x) \end{aligned} \tag{6.1}$$

where $(\alpha g_i(x) + u_i)_+ = \max\{0, \alpha g_i(x) + u_i\}$. A very attractive feature of this function is that a local convexification procedure can be carried out by choosing a sufficiently large penalty parameter α. We state this result in the following lemma which is due to Arrow, Gould and Howe [1].

LEMMA 6.1. *Let f, g and $h \in LC^2[x^*]$ and $z^* = (x^*, u^*, v^*)$ be a Kuhn–Tucker triple which satisfies the Jacobian uniqueness condition. Then there exists an $\bar{\alpha} > 0$ such that if $\alpha \geqq \bar{\alpha}$ then $\nabla_{xx} F(x^*, u^*, v^*, \alpha)$ is positive definite.*

With this result the assumptions on the Lagrangian L in Theorems 5.3 and 5.4 can be relaxed if the function F replaces the Lagrangian L. Moreover, with a large α the function F has the property of penalizing infeasible points, so the domain of convergence is likely to be enlarged.

For reasons which will become clear later on we consider separately the equality constraint problem

$$\begin{aligned} &\min_x \quad f(x) \\ &\text{subject to} \quad h(x) = 0 \end{aligned} \tag{6.2}$$

and the inequality constraint problem

$$\begin{aligned} &\min_x \quad f(x) \\ &\text{subject to} \quad g(x) \leqq 0; \end{aligned} \tag{6.3}$$

the functions F and the Jacobian uniqueness condition are also defined accordingly. We note that the modified algorithms are applicable to the general problem with constraints of mixed type.

For problem (6.2), Algorithms D1 and D2 can be modified as follows.

ALGORITHM M.

Step 1. Start with a penalty parameter α, an estimate $z^0=(x^0, v^0)$ of a Kuhn–Tucker pair $z^*=(x^*, v^*)$ of (6.2) and an estimate A_0 of $\nabla_{xx}F(x^*, v^*, \alpha)^{-1}$.

Step 2. Set $k=0$.

Step 3. Solve the system of linear equations

$$Bv=b \tag{6.4}$$

where

$$B=\nabla h(x^k)^T A_k \nabla h(x^k),$$

$$b=h(x^k)-\nabla h(x^k)^T A_k(\nabla f(x^k)+\alpha \nabla h(x^k)h(x^k)),$$

and let the solution be v^{k+1}. Set

$$x^{k+1}=x^k-A_k\nabla_x F(x^k, v^{k+1}, \alpha). \tag{6.5}$$

Step 4. Check convergence; if not, generate A_{k+1} from A_k, $s^k=x^{k+1}-x^k$ and $y^k=\nabla_x F(x^{k+1}, v^{k+1}, \alpha)-\nabla_x F(x^k, v^{k+1}, \alpha)$ either by (5.2) or by (5.3). Set $k=k+1$ and go to Step 3.

To show the superlinear convergence of Algorithm M, we consider the following auxiliary problem:

$$\begin{aligned} &\min_x && f(x)+\frac{\alpha}{2}h(x)^T h(x) \\ &\text{subject to} && h(x)=0. \end{aligned} \tag{6.6}$$

It is evident that problems (6.2) and (6.6) have the same Kuhn–Tucker pairs and furthermore, the function $F(x, v, \alpha)$ is the Lagrangian of problem (6.6). When Algorithm D1 or D2 is adopted to solve (6.6), the resulting algorithm is just Algorithm M. Therefore, taking Lemma 6.1 into consideration, the following results follows from Theorem 5.4.

THEOREM 6.2. *Let $z^*=(x^*, v^*)$ be a Kuhn–Tucker pair of* (6.2) *satisfying the Jacobian uniqueness condition and let f, g and $h\in LC^2[x^*]$. If the penalty parameter α is sufficiently large and if the starting point $z^0=(x^0, v^0)$ and the starting matrix A_0 are sufficiently close to z^* and $\nabla_{xx}F(x^*, v^*, \alpha)^{-1}$ respectively then the sequence $\{z^k\}$ generated by Algorithm* M *exists and converges Q-superlinearly to z^*.*

For the inequality constraint problem (6.3) the modification is the following.

ALGORITHM M′.

Step 1. Start with a positive number α and an estimate $z^0=(x^0, u^0)$ of a Kuhn–Tucker pair $z^*=(x^*, u^*)$ of (6.3) and an estimate A_0 of $\nabla_{xx}F(x^*, u^*, \alpha)^{-1}$.

Step 2. Set $k=0$.

Step 3. Solve the following quadratic programming subproblem

$$\begin{aligned} &\min_u && \tfrac{1}{2}\phi_k(u)^T A_k \phi_k(u)-u^T g(x^k) \\ &\text{subject to} && u\geqq 0 \end{aligned} \tag{6.7}$$

where $\phi_k(u)=\nabla f(x^k)+\alpha\sum_{i\in I_k} g_i(x^k)\nabla g_i(x^k)+\nabla g(x^k)u$ and $I_k=\{i: g_i(x^k)\geqq -u_i^k/\alpha\}$: let its solution be u^{k+1} and set

$$x^{k+1}=x^k-A_k\nabla_x F(x^k, u^{k+1}, \alpha). \tag{6.8}$$

Step 4. Check convergence; if not, generate matrix A_{k+1} from A_k, $s^k = x^{k+1} - x^k$ and $y^k = \nabla_x F(x^{k+1}, u^{k+1}, \alpha) - \nabla_x F(x^k, y^{k+1}, \alpha)$ either by (5.2) or by (5.3). Set $k = k+1$ and go to Step 3.

THEOREM 6.3. *Let $z^* = (x^*, u^*)$ be a Kuhn–Tucker pair of* (6.3) *satisfying the Jacobian uniqueness condition and f, g and $h \in LC^2[x^*]$. If the penalty parameter α is sufficiently large and the starting point $z^0 = (x^0, u^0)$ and the starting matrix A_0 are sufficiently close to z^* and $\nabla_{xx} F(x^*, u^*, \alpha)^{-1}$ respectively, then the sequence $\{z^k\}$ generated by Algorithm* M′ *exists and converges Q-superlinearly to z^*.*

Proof. Consider the following auxiliary problem

$$\begin{aligned} &\min_x && f(x) + \frac{\alpha}{2} \sum_{i \in I^*} g_i(x)^2 \\ &\text{subject to} && g(x) \leqq 0 \end{aligned} \tag{6.9}$$

where $I^* = \{i : g_i(x^*) = 0\}$. If $z = (x, u)$ and $z^k = (x^k, u^k)$ are sufficiently close to z^* and α is sufficiently large, then it is easy to check that $I_k = I^*$. If we further assume that $u_i = 0$ for all $i \notin I^*$, then $F(x, u, \alpha)$ turns out to be the Lagrangian of (6.9). Therefore, with the second part of Corollary 3.9 taken into account, Algorithm M′ is equivalent to Algorithm D1 or D2 applied to problem (6.9). Since z^* is also a Kuhn–Tucker pair of (6.9), the theorem follows immediately from Theorem 5.4 and Lemma 6.1. □

We note that problem (6.9) is only used in the proof of convergence for Algorithm M′. Actually, it is not obtainable because the set I^* is not known a priori. We also point out that the assumptions of nonsingularity and positive definiteness on the Hessian of the Lagrangian L are not needed in Theorem 6.3 because of the local convexification property of the function F.

We would like to compare our modified algorithms with the recently developed multiplier method in which we generate $z^{k+1} = (x^{k+1}, u^{k+1}, v^{k+1})$ from $z^k = (x^k, u^k, v^k)$ by

$$u_i^{k+1} = \max\{0, u_i^k + \alpha g_i(x^k)\} \quad \text{for } i = 1, \cdots, m, \tag{6.10}$$

$$v_j^{k+1} = v_j^k + \alpha h_j(x^k) \quad \text{for } j = 1, \cdots, q, \tag{6.11}$$

and

$$F(x^{k+1}, u^{k+1}, v^{k+1}, \alpha) = \min_x F(x, u^{k+1}, v^{k+1}, \alpha). \tag{6.12}$$

This method has been shown superlinearly convergent if α is replaced by a sequence $\{\alpha_k\}$ required to go to infinity [3]. However, this usually causes numerical instability. It has been shown [4] that (6.10) and (6.11) are a steepest ascent step for finding a maximum point of the function $\psi_\alpha(u, v) = \min_x F(x, u, v, \alpha)$. With α bounded the multiplier method has only a linear rate of convergence. To avoid the numerical instability caused by large α and at the same time to achieve a superlinear rate we need a more accurate scheme for updating (u^k, v^k) than (6.10) and (6.11). An appropriate candidate is (6.4) and (6.7). This approach results in our modified algorithms. Moreover, to find x^{k+1} we need only take one step of a variable metric method to minimize $F(\cdot, u^{k+1}, v^{k+1}, \alpha)$ with an updated matrix A_k which has already been obtained in the stage of finding multiplier

vector (u^{k+1}, v^{k+1}); however, in the multiplier method we need to do a whole unconstrained minimization process. It is noted that a similar approach can be found in [23].

7. Comments and computational experiences.

1. The algorithm based on the subproblem

$$\begin{aligned} \min_x \quad & \nabla f(x^k)^T(x-x^k)+\tfrac{1}{2}(x-x^k)^T H_k(x-x^k) \\ \text{subject to} \quad & g(x^k)+\nabla g(x^k)^T(x-x^k)\leqq 0, \\ & h(x^k)+\nabla h(x^k)^T(x-x^k)=0 \end{aligned}$$

can be viewed as primal to the algorithm discussed in this paper because this subproblem is primal to subproblem (2.1). To achieve local superlinear convergence for the primal algorithm the matrices $\{H_k\}$ need to be good estimates to the Hessian of the Lagrangian rather than to the inverse of the Hessian. In [22] local superlinear convergence has been established for the primal algorithm with $\{H_k\}$ updated by the following schemes

$$\bar{H}=H+\frac{(y-Hs)y^T+y(y-Hs)^T}{y^Ts}-\frac{s^T(y-Hs)yy^T}{(y^Ts)^2} \tag{7.1}$$

(unconstrained case: Davidon–Fletcher–Powell [12], [15])

$$\bar{H}=H+\frac{(y-Hs)s^T+S(y-Hs)^T}{s^Ts}-\frac{s^T(y-Hs)ss^T}{(s^Ts)^2} \tag{7.2}$$

(unconstrained case: Powell [36], [37])

where s and y are defined as in (5.1). It is noted that updates (7.1) and (7.2) are dual to updates (5.2) and (5.3) respectively in the sense of Fletcher [16]. The duality of updating schemes and the duality of mathematical programming have been defined and used in two different contexts. It is very interesting that in our approach they are coincidentally connected to each other. We also note that though some theorems in this paper are analogous to those in [22], there is no direct implication among them.

2. Our algorithm is in a sense a natural extension of variable metric algorithms to general nonlinear programming and this extension provides a fruitful field of future research. A lot of results in the extensive literature of variable metric algorithms need to be investigated and developed for nonlinear programming and the whole theory can be treated in a unified way in both constrained and unconstrained optimization.

3. All the results in this paper are local. One approach studied by this author for achieving global convergence is to determine a stepsize in each iteration which maintains a monotone decrease of an exact penalty function or the penalty Lagrangian defined in (6.1). Some global convergence results have already been established [21], [24].

Computational tests of the algorithms in this paper have been performed and are still going on. A report on the tests results is expected to be published in the

near future. However, it would be unfair to finish without at least giving some idea of the power of these algorithms in practice. We state in the table below the test results of Algorithm D1 and D2 for Colville's test problems 1 and 2. The computations were done on the UNIVAC 1110 system at the University of Wisconsin—Madison. The principal pivoting method [10], [11] was used in solving the quadratic programming subproblems.

TABLE 1

Prob.	Algorithm	Obj. Fct. value	Standard time ratio
1	D1	−32.3487	.00448*
	D2	−32.3486	.00906
2†	D1	−32.3488	.2133
	D2	−32.3488	.6311

* This result is better than any one reported in Colville's report [9].
† Infeasible starting point.

Acknowledgment. This work is an extension of a portion of the author's doctoral thesis under the supervision of Professor O. L. Mangasarian at the Department of Computer Sciences, University of Wisconsin—Madison. The author is also indebted to Professors J. E. Dennis, Jr., and J. J. Moré for reading the manuscript and making many valuable suggestions.

[1] K. J. ARROW, F. H. GOULD AND S. M. HOWE, *A general saddle point result for constrained optimization*, Math. Programming, 5 (1973), pp. 225–234.
[2] E. M. L. BEALE, *On quadratic programming*, Naval Res. Logist. Quart., 6 (1959), pp. 227–243.
[3] D. P. BERTSEKAS, *On penalty and multiplier methods for constrained minimization*, this Journal, 14 (1976), pp. 216–235.
[4] ———, *Combined primal-dual and penalty methods for constrained minimization*, this Journal, 13 (1975), pp. 521–544.
[5] C. G. BROYDEN, *A class of methods for solving nonlinear simultaneous equations*, Math. Comput., 19 (1965), pp. 577–593.
[6] ———, *A new double-rank minimization algorithm*, Abstract, Notices Amer. Math. Soc., 760 (1969), no. 16, p. 670.
[7] C. G. BROYDEN, J. E. DENNIS AND J. J. MORÉ, *On the local and superlinear convergence of quasi-Newton methods*, J. Inst. Math. Appl., 12 (1973), pp. 223–245.
[8] J. D. BUYS, *Dual algorithms for constrained optimization*, Ph.D. thesis, University of Leiden, Leiden, the Netherlands, 1972.
[9] A. R. COLVILLE, *A comparative study on nonlinear programming codes*, IBM New York Scientific Center Report, 320–2949, New York, 1968.
[10] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory of mathematical programming*, Linear Algebra and Appl., 1 (1968), pp. 103–125.
[11] R. W. COTTLE, *The principal pivoting method of quadratic programming*, Mathematics of the Decision Sciences, G. B. Dantzig and A. F. Veinott, eds., vol. 1, American Mathematical Society, Providence, R.I., 1968, pp. 144–162.
[12] W. C. DAVIDON, *Variable metric method for minimization*, Argonne Nat. Lab. Rep. #ANL-5990, Argonne, Ill., 1959.

[13] J. E. DENNIS AND J. J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comput., 28 (1974), pp. 549–560.
[14] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
[15] R. FLETCHER AND M. J. D. POWELL, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.
[16] R. FLETCHER, *A new approach to variable metric algorithms*, Ibid., 13 (1970), pp. 317–322.
[17] U. M. GARCIA-PALOMARE, *Superlinearly convergent quasi-Newton method for nonlinear programming*, Ph.D. dissertation, Comput. Sci. Dept., Univ. of Wisconsin, Madison, 1973.
[18] U. M. GARCIA-PALOMARE AND O. L. MANGASARIAN, *Superlinearly convergent quasi-Newton algorithms for nonlinearly constrained optimization problems*, Computer Sciences Tech. Rep. 195, Univ. of Wisconsin, Madison, 1974.
[19] D. GOLDFARB, *A family of variable metric methods derived by variational means*, Math. Comput., 24 (1970), pp. 23–26.
[20] J. GREENSTADT, *Variations on variable metric methods*, Ibid., 24 (1970), pp. 1–18.
[21] S. P. HAN, *Superlinearly Convergent variable metric algorithms for general nonlinear programming problems*, Ph.D. dissertation, Comput. Sci. Dept., Univ. of Wisconsin, Madison, 1974.
[22] ———, *Superlinearly convergent variable metric algorithms for general nonlinear programming problems*, Math. Programming, 11 (1976).
[23] ———, *Penalty Lagrangian methods via a quasi-Newton approach*, Computer Science Tech. Rep. 75–252, Cornell Univ., Ithaca, N.Y., 1975.
[24] ———, *A globally convergent method for nonlinear programming*, J. Optimization Theory Appl., to appear.
[25] M. R. HESTENES, *Multiplier and gradient methods*, Ibid., 4 (1969), pp. 303–320.
[26] E. ISAACSON AND H. B. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.
[27] C. E. LEMKE, *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11 (1965), pp. 681–689.
[28] L. A. LIUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Frederick Ungar, New York, 1961.
[29] F. A. LOOTSMA, *A survey of methods for solving constrained minimization problems via unconstrained minimization*, Numerical Methods for Nonlinear Optimization, F. A. Lootsma, ed., Academic Press, London, 1972, pp. 313–347.
[30] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
[31] ———, *Dual, feasible direction algorithms*, Techniques of Optimization, A. V. Balakvishrian, ed., Academic Press, New York, 1972, pp. 67–88.
[32] ———, Private communication.
[33] J. D. PEARSON, *Variable metric methods of minimization*, Comput. J., 12 (1969), pp. 171–178.
[34] B. T. POLYAK, *Iterative methods using Lagrange multipliers for solving extremal problems with constraints of the equation type*, U.S.S.R. Computational Math. and Math. Phys., 10 (1970), pp. 42–52.
[35] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, London, 1969, pp. 283–298.
[36] ———, *A new algorithm for unconstrained optimization*, Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1970.
[37] ———, *A FORTRAN subroutine for unconstrained minimization, requiring first derivatives of the objective functions*, A.E.R.E. Harwell Rep. R64-69, 1970.
[38] K. RITTER, *A method for solving maximum problems with a nonconcave quadratic objective function*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 4 (1966), pp. 340–351.
[39] R. T. ROCKAFELLAR, *New applications of duality in nonlinear programming*, Symp. on Mathematical Programming, The Hague, the Netherlands, September 1970.
[40] ———, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optimization Theory Appl. 12 (1973), pp. 555–562.
[41] D. F. SHANNO, *Conditions of quasi-Newton methods for function minimization*, Math. Comput., 24 (1970), pp. 647–656.
[42] P. WOLFE, *The simplex method for quadratic programming*, Econometrica, 27 (1959), pp. 382–398.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

AN ABSTRACT THEORY FOR UNBOUNDED CONTROL ACTION FOR DISTRIBUTED PARAMETER SYSTEMS*

RUTH F. CURTAIN AND A. J. PRITCHARD†

Abstract. In *The infinite dimensional Riccati equations for systems defined by evolution operators* [Ruth F. Curtain and A. J. Pritchard, this Journal, 14 (1976), pp. 951–983], we have examined the linear quadratic control problem for systems described by abstract input-output relationships on Hilbert spaces, but the application of our results to distributed systems governed by partial differential equations requires that the control operators are bounded. This is a severe restriction, since for most systems of practical interest the controls will act on the boundary or on submanifolds of the system region and so unbounded operators are involved. In this paper we generalize the above work to include such control action.

Introduction. The linear, quadratic control problem for distributed systems has been studied by Lions [7] for operators satisfying a coercivity condition, and by Curtain and Pritchard [2] using a semigroup or evolution operator approach. Most of the work in these references is concerned with bounded control action. However, in practice it is very difficult to implement bounded control action because of the severe limitations that are inherent in distributed systems. For example control action is usually confined to regions on the boundary of the system or to a manifold of lower dimension and interior to the system region. Another way that unbounded control action can arise is if the implementation of the control on the system involves an unbounded operator. For example, in a heat conduction process the control may be related to the temperature; however, the only way of effecting the system may be by heat flow considerations, and this will require the gradient of the control, which could be an unbounded operator. While it is sometimes possible to formulate these problems in the bounded control theory of [2], [7], by changing the state and control spaces, usually this is unsatisfactory and it is necessary to consider unbounded control action.

In [7], [8], Lions has developed a very general theory for boundary control action for distributed systems, where the operators satisfy a coercivity condition. Balakrishnan has used a semigroup approach in [1] to study the special case of boundary control problems for the diffusion equation. However, our approach differs from both these authors and we develop a unified approach to the quadratic cost control problem for a wider class of distributed systems with unbounded control action using the semigroup or evolution operator approach introduced in [2].

There is of course a strong duality between the filtering and control problems, and this is often exploited in solving the filtering problem (see [3]). Our approach also enables us to consider the filtering problem for the cases where the observations are limited to regions of the boundary or manifolds of lower dimension interior to the system, or indeed points. For example, the dual of boundary control action is boundary observations for the filtering problem, and in fact in order to formulate the boundary problem we first formulate the filtering problem with

* Received by the editors July 8, 1976.

† Control Theory Centre, University of Warwick, Coventry, Warwickshire CV4 7AL, England.

boundary observations, and examine its mathematical dual. So for this case the duality between the filtering and control problems is particularly important.

As our approach to the unbounded control problem is abstract, in § 1 we discuss the motivation for our analysis in some detail with reference to specific examples. Paralleling the treatment in [2], in § 2 we develop an unbounded perturbation theory for mild evolution operators $\mathcal{U}(t, s)$ on a Hilbert space H, and in § 3 we develop the abstract theory for the quadratic cost unbounded control problem, obtaining the optimal control in feedback form. The feedback operator is again the unique solution for an integral Riccati equation, but this time with unbounded operators. As in the bounded case we show that if $\mathcal{U}(t, s)x$ is differentiable with respect to s almost everywhere for x in a dense set in H (that is $\mathcal{U}(t, s)$ is a quasi-evolution operator), then the integral Riccati equation may be differentiated to obtain a differential version. If further $\mathcal{U}(t, s)x$ can be differentiated with respect to t for x in a dense set in H (that is, $\mathcal{U}(t, s)$ is a strong evolution operator) we can show that the differential Riccati equation has a unique solution. As the application of these abstract results is not straightforward, and depends on duality concepts, this aspect is examined in some detail in § 5. Finally in § 6 the theory is applied to several classes of distributed systems, including the parabolic and hyperbolic systems with different types of unbounded control action.

1. Motivation. Our aim is to consider the control system

$$\dot{z}(t) = A(t)z(t) + B(t)u(t), \qquad z(0) = z_0, \tag{1.1}$$

where $A(t)$, $B(t)$ are linear unbounded operators. However, before we give the precise conditions imposed on these operators we will motivate our considerations by some examples of autonomous systems of the form

$$\dot{z} = Az + Bu, \qquad z(0) = z_0. \tag{1.2}$$

We assume that A is the infinitesimal generator of a strongly continuous semigroup $\mathcal{T}_t$ on a Hilbert space H and we will consider two different assumptions on the operator B. First it is necessary to give some interpretation of a solution of (1.2). One way is to consider the integral equation

$$z(t) = \mathcal{T}_t z_o + \int_0^t \mathcal{T}_{t-s} Bu(s)\, ds. \tag{1.3}$$

If $B \in \mathcal{L}(U, H)$, where U is a Hilbert space, and $u \in L_2(0, T; U)$ such a solution is known as a "mild solution" and we know $z \in C[0, T; H]$ (see [2]), although we are not in general able to differentiate (1.3) to obtain (1.2). We want to generalize the concept of a mild solution to the case where B is an unbounded operator. In order to see how this may be achieved let us consider the simple example

Example 1.1. Let

$$Az = z_{xxxx}, \qquad z \in \mathcal{D}(A),$$

$$\mathcal{D}(A) = \{z : z \in L_2(0, 1), Az \in L_2(0, 1); z(0) = 0 = z(1), z_{xx}(0) = 0 = z_{xx}(1)\},$$

$$Bu = u_x, \qquad u \in \mathcal{D}(B),$$

$$\mathcal{D}(B) = \{u : u \in L_2(0, 1), Bu \in L_2(0, 1)\}.$$

Then (1.2) is an evolution equation which results from abstracting the partial differential equation

$$z_t = z_{xxxx} + u_x, \qquad z(0,t) = z(1,t) = z_{xx}(0,t) = z_{xx}(1,t) = 0. \tag{1.4}$$

Now A generates the semigroup $\mathcal{T}_t$, where

$$\mathcal{T}_t z = \sum_{n=1}^{\infty} 2c^{-n^4\pi^4 t} \sin n\pi x \int_0^1 \sin n\pi y x(y)\, dy$$

and it is easy to show that

$$\|\mathcal{T}_t B u\|_{L_2(0,1)} \leqq \frac{M}{t^{1/4}} \|u\|_{L_2(0,1)}, \qquad u \in \mathcal{D}(B), \quad t > 0.$$

Thus for each $t > 0$, $\mathcal{T}_t B \in \mathcal{L}(\mathcal{D}(B), L_2(0,1))$ and since $\overline{\mathcal{D}(B)} = L_2(0,1)$, $\mathcal{T}_t B$ has a unique extension, which we will denote by $\overline{\mathcal{T}_t B}$, and $\overline{\mathcal{T}_t B} \in \mathcal{L}(L_2(0,1))$, $t > 0$. We have

$$\overline{\mathcal{T}_t B} u = \lim_{n\to\infty} \mathcal{T}_t B u_n, \qquad u_n \in \mathcal{D}(B), \quad u_n \to u \text{ in } L_2(0,1).$$

Furthermore for all $t > 0$, $s \geqq 0$, $u \in L_2(0,1)$,

$$\begin{aligned} \overline{\mathcal{T}_{t+s} B} u &= \lim_{n\to\infty} \mathcal{T}_{t+s} B u_n \\ &= \mathcal{T}_s \lim_{n\to\infty} \mathcal{T}_t B u_n \\ &= \mathcal{T}_s \overline{\mathcal{T}_t B} u. \end{aligned}$$

We define a solution of (1.4) by

$$z(t) = \mathcal{T}_t z_0 + \lim_{\varepsilon\to 0} \int_0^{t-\varepsilon} \overline{\mathcal{T}_{t-s} B} u(s)\, ds. \tag{1.5}$$

It is a simple matter to show that $z(\cdot)$ given by (1.5) is well defined for $u \in L_2(0, T; L_2(0,1))$ and $z \in C[0, T; L_2(0,1)]$ (we will prove this in a more general setting in § 2).

From Example 1.1 we see that the kinds of conditions it is necessary to impose are

$$\overline{\mathcal{D}(B)} = U, \tag{1.6}$$

$$\|\mathcal{T}_t B u\|_H \leqq \frac{M}{t^\alpha} \|u\|_u, \qquad u \in \mathcal{D}(B), \quad \alpha < \tfrac{1}{2}. \tag{1.7}$$

The $\alpha < \frac{1}{2}$ arises from the estimate

$$\begin{aligned} \left\| \int_0^{t-\varepsilon} \overline{\mathcal{T}_{t-s} B} u(s)\, ds \right\|_H &\leqq \int_0^{t-\varepsilon} \|\overline{\mathcal{T}_{t-s} B} u(s)\|\, ds \\ &\leqq \int_0^{t-\varepsilon} \frac{M}{(t-s)^\alpha} \|u(s)\|\, ds \\ &\leqq \left(\int_0^t \frac{M}{(t-s)^{2\alpha}}\, ds \right)^{1/2} \|u\|_{L_2(0,T;U)}. \end{aligned}$$

We will also consider the filtering problem

$$z(t) = \mathcal{T}_t z_0 + \int_0^t \mathcal{T}_{t-s} D\, dw(s), \tag{1.8}$$

$$y(t) = \int_0^t Cz(s)\, ds + \int_0^t F\, dv(s), \tag{1.9}$$

where $w(t)$, $v(t)$ are noise processes which will be specified precisely in § 5. It turns out that the filtering and control problems can be developed in parallel so that unbounded operators C may be considered if for example we make the identification $B^* = C$. We note that this implies $C: H \to U^*$ is a closed linear operator [11], and if we further assume B is closed, we have $C^* = B$ and C is densely defined.

The above considerations do not enable us to consider control or observations from boundaries or from lower dimensional manifolds. For these problems C will not be closed, although it will be densely defined, whereas in general B will not be densely defined and so we do not have condition (1.6). In order to see how we can treat such operators let us consider the following simple example.

Example 1.2.

$$\begin{aligned} &z_t = z_{xx}, \\ &z_x(0, t) = u, \quad z_x(1, t) = 0, \quad z(x, 0) = z_0(x). \end{aligned} \tag{1.10}$$

We examine this problem by first determining an operator B such that the mild solution of

$$\begin{aligned} &z_t = z_{xx} + Bu, \\ &z_x(0, t) = 0, \quad z_x(1, t) = 0; \qquad z(x, 0) = z_0(x) \end{aligned} \tag{1.11}$$

is a weak solution of (1.10). Abstracting the problem (1.11) we obtain

$$\dot{z} = Az + Bu; \qquad z(0) = z_0,$$

where A generates the following semigroup $\mathcal{J}_t$ on $L_2(0, 1)$:

$$(\mathcal{T}_t h)(x) = \sum_{n=1}^{\infty} 2\, e^{-n^2\pi^2 t} \cos \pi x \int_0^1 h(y) \cos n\pi y\, dy.$$

We will show in § 5 that this approach leads to an operator $B = C^*$, where

$$(Cz)(t) = -z(0, t).$$

Now C is densely defined on $L_2(0, 1)$, but is not closed; so, B will be a closed operator but its domain is trivial (i.e., B has domain 0). In fact $Bu = -\delta(x)u$, the delta function, so $Bu \notin L_2(0, 1)$ for any $u \neq 0$. However we note that for $t > 0$, $\mathcal{T}_t Bu \in L_2(0, 1)$ and

$$\|\mathcal{T}_t By\|_{L_2(0,1)} \leqq \frac{M}{t^{1/4}} |u|.$$

Thus we are able to define a mild solution

$$z(t) = \mathcal{T}_t z_0 = \int_0^t \mathcal{T}_{t-s} B u(s)\, ds$$

for all $u \in L_2(0, T)$.

This second example indicates that in order to incorporate this type of problem into our general theory we will need to assume that $\mathcal{T}_t$ can be extended to act on rays of a larger space than H (a typical element being Bu), in such a way that for $t > 0$,

$$\|\mathcal{T}_t B u\|_H \leqq \frac{M}{t^\alpha} \|u\|_U; \qquad u \in U, \quad t > 0, \quad \alpha < \tfrac{1}{2}.$$

In order to present a unified theory for the types of problems associated with Examples 1.1, 1.2, we introduce an operator $\mathcal{T}(t; B)$ which satisfies

$$\mathcal{T}(t; B) \in \mathcal{L}(U, H) \quad \text{for } t > 0, \tag{1.12}$$

$$\mathcal{T}_s \mathcal{T}(t; B) = \mathcal{T}(t+s; B) \quad \text{for } t > 0, \quad s \geqq 0, \tag{1.13}$$

$$\|\mathcal{T}(t; B)\| \leqq \frac{M}{t^\alpha}; \qquad \alpha < \tfrac{1}{2}, \quad t > 0. \tag{1.14}$$

In fact in the next section we generalize the operator $\mathcal{T}(t; B)$ to an evolution type operator $\mathcal{U}(t, s; B)$ in order to include nonautonomous evolution equations.

2. Unbounded perturbations of evolution operators. We recall the definition of a mild evolution operator.

DEFINITION 2.1. Let H be a real Hilbert space, $T = [0, T]$ a real time interval, and $\Delta(T) = \{(t, s): 0 \leqq s < t \leqq T\}$. Then $\mathcal{U}(t, s): \Delta(T) \to \mathcal{L}(H)$ is a *mild evolution operator* if

$$\mathcal{U}(t, r)\mathcal{U}(r, s) = \mathcal{U}(t, s) \quad \text{for } 0 \leqq s \leqq r \leqq t \leqq T, \qquad \mathcal{U}(t, t) = I. \tag{2.1}$$

$$\begin{aligned} &\mathcal{U}(t, \cdot) \quad \text{is weakly continuous on } [0, t) \quad \text{and} \\ &\mathcal{U}(\cdot, s) \quad \text{is weakly continuous on } (s, T]. \end{aligned} \tag{2.2}$$

In this paper all mild evolution operators will satisfy the stronger hypotheses

$$\mathcal{U}(\cdot, \cdot) \quad \text{is jointly strongly continuous on } \Delta(T). \tag{2.2}'$$

This is a mathematical convenience to simplify some of the proofs, and is not a crucial assumption. (All the results remain true under (2.2).)

We take our system model to be

$$z(t) = \mathcal{U}(t, 0) z_0 + \int_{t_0}^t \mathcal{U}(t, r; B) u(r)\, dr, \tag{2.3}$$

where $u \in L_2(T; U)$ and we make the following assumptions on $\mathcal{U}(\cdot, \cdot; B)$:

$$\|\mathcal{U}(t, s; B)u\|_H \leqq g(t-s)\|u\|_U, \qquad 0 \leqq s < t \leqq T, \quad g \in L_2(T). \tag{2.4}$$

$$\mathcal{U}(t, r)\mathcal{U}(r, s; B) = \mathcal{U}(t, s; B), \qquad 0 \leqq s < r \leqq t \leqq T. \tag{2.5}$$

We note that (2.5) implies

$$\|\mathcal{U}(t, s; B)u\|_H \leqq Mg(r-s)\|u\|_U; \qquad 0 \leqq s < r \leqq t \leqq T,$$

where $M = \sup_{\Delta(T)} \|\mathcal{U}(t, s)\|$ and so without loss of generality we may assume

$$g(t+s) \leqq Mg(t), \qquad s \geqq 0, \quad t > 0. \tag{2.6}$$

The interpretation to be given to (2.4) is

$$\int_0^t \|\mathcal{U}(t, s; B)u(s)\|_H \, ds \leqq \int_0^t g(t-s)\|u(s)\|_U \, ds \quad \text{for } u \in L_2(T; U).$$

By using (2.4) it is easy to show that $z(t)$ is a well defined H-valued function for each $t \in T$ and in fact using (2.5) we can show that $z(\cdot) \in C[t_0, T; H]$. If $h > 0$, we have

$$\begin{aligned} z(t+h) - z(t) = {} & (\mathcal{U}(t+h, t_0) - \mathcal{U}(t, t_0))z_0 \\ & + \int_{t_0}^{t} [\mathcal{U}(t+h, r; B) - \mathcal{U}(t, r; B)]u(r) \, dr \\ & + \int_{t}^{t+h} \mathcal{U}(t+h, r; B)u(r) \, dr. \end{aligned}$$

Thus

$$\begin{aligned} \|z(t+h) - z(t)\| \leqq {} & \|(\mathcal{U}(t+h, t_0) - \mathcal{U}(t, t_0))z_0\| \\ & + \left\|(\mathcal{U}(t+h, t) - I) \int_{t_0}^{t} \mathcal{U}(t, r; B)u(r) \, dr\right\| \\ & + \int_{t}^{t+h} \|\mathcal{U}(t+h, r; B)u(r)\| \, dr. \end{aligned}$$

Hence by (2.2)′ and (2.4) we are able to conclude continuity on the right. For $h < 0$, we have

$$\begin{aligned} z(t) - z(t-h) = {} & (\mathcal{U}(t, t_0) - \mathcal{U}(t-h, t_0))z_0 \\ & + \int_{t_0}^{t-h} [\mathcal{U}(t, r; B) - \mathcal{U}(t-h, r; B)]u(r) \, dr \\ & + \int_{t-h}^{t} \mathcal{U}(t, r; B)u(r) \, dr. \end{aligned}$$

Thus

$$\begin{aligned}\|z(t)-z(t-h)\| &\leqq \|(\mathcal{U}(t,t_0)-\mathcal{U}(t-h,t_0))z_0\| \\ &\quad + \left\|(\mathcal{U}(t,t-h)-I)\int_{t_0}^{t-h}\mathcal{U}(t-h,r;B)u(r)\,dr\right\| \\ &\quad + \int_{t-h}^{t}\|\mathcal{U}(t,r;B)u(r)\|\,dr \\ &\leqq \|(\mathcal{U}(t,t_0)-\mathcal{U}(t-h,t_0))z_0\| \\ &\quad + \|(\mathcal{U}(t,t-\varepsilon)-\mathcal{U}(t-h,t-\varepsilon))z(t+\varepsilon)\| \\ &\quad + \int_{t-\varepsilon}^{t-h}\|\mathcal{U}(t-h,r;B)u(r)\|\,dr \\ &\quad + \int_{t-\varepsilon}^{t}\|\mathcal{U}(t,r;B)u(r)\|\,dr\end{aligned}$$

for any $t>\varepsilon>h$. And so $z(\cdot)\in C(t_0, T; H)$ by using (2.4).

Since we wish to consider feedback controls of the form $u(t)=F(t)z(t)$, we are led to considering perturbations of $\mathcal{U}(t, s)$ defined by

$$\mathcal{U}_{BF}(t,s)x=\mathcal{U}(t,s)x+\int_s^t \mathcal{U}(t,r;B)F(r)\mathcal{U}_{BF}(r,s)x\,dr, \qquad x\in H, \tag{2.7}$$

for a special case of feedback gain operators $F(t)$ satisfying

$$\|F(t)h\|_U \leqq f(T-t)\|h\|_H \tag{2.8}$$

for any $t_2\in(t, T]$ and some $f\in L_2(T)$, which satisfies (2.6).

Our main perturbation result is the following.

THEOREM 2.1. *Let $\mathcal{U}(t, s)$ be a mild evolution operator satisfying* (2.1), (2.2)', *$\mathcal{U}(t, s; B)$ satisfying* (2.4)–(2.6) *and $F(t)$ satisfying* (2.8). *Then* (2.7) *has a unique solution $\mathcal{U}_{BF}(t, s)$ which is a mild evolution operator on H, satisfying* (2.1) *and* (2.2).

Proof. (a) *Existence and uniqueness.* The proof is the usual constructive approach for Volterra integral equations, where we consider $V_n(t, s)\in\mathcal{L}(H)$ given by

$$\begin{aligned}V_0(t,s)&=\mathcal{U}(t,s),\\ V_n(t,s)x&=\int_s^t \mathcal{U}(t,r;B)F(r)V_{n-1}(r,s)x\,dr, \qquad x\in H.\end{aligned}$$

Note that (2.4), (2.8) imply

$$\begin{aligned}\|\mathcal{U}(t,r;B)F(r)h\|_H &\leqq Cg(t-r)f(t-r)\|h\|_H \\ &= G(t-r)\|h\|_H\end{aligned} \tag{2.9}$$

for all $t\in(r, T]$ and almost all $r\in[0, t)$. By induction we prove that

$$\|V_n(t,s)x\|\leqq M\int_s^t G_n(t-\alpha)\,d\alpha\,\|x\| \tag{2.10}$$

for $n \geqq 1$, where

$$G_1(t) = G(t),$$

$$G_n(t) = \int_0^t G(t-r)G_{n-1}(r)\,dr$$

and

$$\|U(t, s)\| \leqq M.$$

We have

$$\|V_1(t, s)x\| \leqq \int_s^t G(t-r)\|V_0(r, s)x\|\,dr \quad \text{(by (2.9))}$$

$$\leqq M\int_s^t G(t-r)\,dr\,\|x\|.$$

If we assume (2.10) holds for $n = k-1$, we have

$$\|V_k(t, s)x\| \leqq \int_s^t G(t-r)\|V_{k-1}(r, s)x\|\,dr$$

$$\leqq M\int_s^t G(t-r)\int_s^t G_{k-1}(r-\alpha)\,d\alpha\,dr\,\|x\|$$

$$= M\int_s^t \int_\alpha^t G(t-r)G_{k-1}(r-\alpha)\,dr\,d\alpha\,\|x\|$$

(interchanging the order of integration)

$$= M\int_s^t G_k(t-\alpha)\,d\alpha\,\|x\|.$$

Thus (2.10) is established.

From Appendix A, we see that the equation

$$v(t-s) = M + \int_s^t G(t-\alpha)v(\alpha-s)\,d\alpha \tag{2.11}$$

can be transformed to the form (A.1), where G satisfies (A.2) by virtue of (2.6) and (2.9). So (2.11) has a unique continuous solution

$$v(t-s) = M + M\sum_{n=1}^{\infty}\int_s^t G_n(t-\alpha)\,d\alpha$$

and thus

$$\sum_{n=0}^{\infty}\|V_n(t, s)\| \leqq M + M\sum_{n=1}^{\infty}\int_s^t G_n(t-\alpha)\,d\alpha$$

$$< \infty \quad \text{(by Corollary (A.4)).}$$

Thus $\mathcal{U}_{BF}(t, s) = \sum_{n=0}^{\infty} V_n(t, s)$ converges absolutely in the uniform topology, and the convergence is uniform in s and t. Clearly this expression satisfies (2.7) with $\mathcal{U}_{BF}(t, t) = I$.

For the uniqueness we suppose that $\mathcal{U}_2(t, s)$ is another solution, and let

$$V(t, s) = \mathcal{U}_{BF}(t, s) - \mathcal{U}_2(t, s).$$

Then

$$V(t, s)x = \int_s^t \mathcal{U}(t, r; B)F(r)V(r, s)x\, dr$$

and

$$\|V(t, s)x\| \leqq \int_s^t G(t-r)\|V(r, s)x\|\, dr.$$

So $V(t, s)x = 0$ by the generalized Gronwall's inequality (Appendix A, (A.5)).

(b) *Semigroup property.*

$$\mathcal{U}_{BF}(t, r)\mathcal{U}_{BF}(r, s)x = \mathcal{U}(t, r)\mathcal{U}(r, s)x + \int_s^r \mathcal{U}(t, r)\mathcal{U}(r, \alpha; B)F(\alpha)\mathcal{U}_{BF}(\alpha, s)x\, d\alpha + \int_r^t \mathcal{U}(t, \alpha; B)F(\alpha)\mathcal{U}_{BF}(\alpha, r)\mathcal{U}_B(r, s)x\, d\alpha.$$

Therefore

$$(\mathcal{U}_{BF}(t, r)\mathcal{U}_{BF}(r, s) - \mathcal{U}_{BF}(t, s)) = \int_r^t \mathcal{U}(t, \alpha; B)F(\alpha) \cdot [\mathcal{U}_{BF}(\alpha, r)\mathcal{U}_{BF}(r, s) - \mathcal{U}_{BF}(\alpha, s)]x\, d\alpha.$$

Denoting the left-hand side by $V(t, r, s)x$ we have

$$\|V(t, r, s)x\| \leqq \int_r^t G(t-\alpha)\|V(\alpha, r, s)x\|\, d\alpha.$$

So $V(t, r, s)x = 0$ by the generalized Gronwall's inequality (Appendix A, (A.5)).

(c) *Continuity.* Since the convergence of $\sum V_n(t, s)x$ is uniform in s and t on T it suffices to prove that each $V_n(t, s)$ is jointly continuous for each n and $x \in H$. In fact since $\sup_{\Delta(T)} \|\mathcal{U}_{BF}(t, s)\| < \infty$, we need only prove continuity in each variable separately (see [10]). The proof is by induction starting with the fact that $\mathcal{U}(\cdot, \cdot)$ is jointly continuous on $\Delta(T)$.

Suppose $V_{k-1}(\cdot, \cdot)$ is jointly continuous; we have

$$V_k(t, s)x = \int_s^t \mathcal{U}(t, r; B)F(r)V_{k-1}(r, s)x\, dr.$$

The continuity of $V_k(t, s)$ in t is proved in exactly the same way as the continuity of $z(t)$ in t for (2.3). For $h > 0$, we have

$$V_k(t, s+h)x - V_k(t, s)x = \int_{s+h}^t \mathcal{U}(t, r; B)F(r)V_{k-1}(r, s+h) - V_{k-1}(r, s)x\, dr - \int_s^{s+h} \mathcal{U}(t, r; B)F(r)V_{k-1}(r, s)x\, dr.$$

Hence

$$\|V(t, s+h)x - V_k(t, s)x\| \leqq \int_{s+h}^{t} G(t-r)\|V_{k-1}(r, s+h)x - V_{k-1}(r, s)x\|\, dr + \int_{s}^{s+h} G(t-r)\|V_{k-1}(r, s)x\|\, dr.$$

Since $\|V_{k-1}(r, s)\|$ is uniformly bounded and $V_{k-1}(r, s)x$ is jointly continuous we see that

$$\|V_k(t, s+h)x - V_k(t, s)x\| \to 0 \quad \text{as } h \to 0.$$

A similar argument holds for $h < 0$.

In order to give meaning to a sequence of feedback controllers in § 3 we need to define terms like

$$h(t) = \int_0^t \mathcal{U}_{BF}(t, s; B)u(s)\, ds \quad \text{for } u \in L_2(T; U),$$

where $\mathcal{U}_{BF}(t, s; B)$ satisfies

$$\mathcal{U}_{BF}(t, s; B)u(s) = \mathcal{U}(t, s; B)u(s) + \int_s^t \mathcal{U}(t, r; B)F(r)\mathcal{U}_{BF}(r, s; B)u(s)\, dr \tag{2.12}$$

The above equation is to be interpreted in $L_1(0, t; H)$ for each $t \in [0, T]$, and $u \in L_2(T; U)$.

THEOREM 2.2. *Let $\mathcal{U}(t, s; B)$ satisfy assumptions* (2.4)–(2.6) *and f satisfy* (2.8); *then for each $u \in L_2(T; U)$,* (2.12) *has a unique solution*

$$\mathcal{U}_{BF}(t, \cdot\,; B) \in \mathcal{L}(L_2(T; U), L_1(0, t; H))$$

such that

$$h(t) = \int_0^t \mathcal{U}_{BF}(t, s; B)u(s)\, ds \in C(T; H).$$

Proof. We construct a sequence of operators $V_n(t, s; B)$,

$$V_0(t, s; B)u(s) = \mathcal{U}(t, s; B)u(s),$$

$$V_n(t, s; B)u(s) = \int_s^t \mathcal{U}(t, r; B)F(r)V_{n-1}(r, s; B)u(s)\, dr.$$

We have

$$\|V_0(t, s; B)u(s)\| \leqq g(t-s)\|u(s)\| \quad \text{by (2.5).}$$

We prove by induction that

$$\|V_n(t, s; B)u(s)\| \leqq \int_0^t G_n(t-r)g(r-s)\, dr\, \|u(s)\| \quad \text{for } n \geqq 1, \tag{2.13}$$

where

$$G_n(t-r)=\int_r^t G(t-\alpha)G_{n-1}(\alpha-r)\,d\alpha \quad \text{for } n\geqq 2,$$

$$G_1(t-\alpha)=Cg(t-\alpha)f(t-\alpha).$$

For $n=1$, we have

$$\|V_1(t,s;B)u(s)\|=\left\|\int_s^t \mathcal{U}(t,r;B)F(r)V_0(r,s;B)u(s)\,dr\right\|$$

$$\leqq \int_s^t G(t-r)g(r-s)\,dr\,\|u(s)\| \quad \text{by (2.9)}.$$

If we assume (2.13) holds for $n=k-1$, we have

$$\|V_k(t,s;B)u(s)\|\leqq \int_s^t G(t-r)\|V_{k-1}(r,s;B)u(s)\|\,dr$$

$$\leqq \int_s^t G(t-r)\int_s^r G_{k-1}(r-\alpha)g(\alpha-s)\,d\alpha\,\|u(s)\|\,dr$$

$$=\int_s^t\int_\alpha^t G(t-r)G_{k-1}(r-\alpha)\,dr\,g(\alpha-r)\,d\alpha\,\|u(s)\|$$

$$=\int_s^t G_k(t-r)g(r-s)\,dr\,\|u(s)\|.$$

Hence

$$\int_0^t \|V_n(t,s;B)u(s)\|\,ds\leqq \int_0^t\int_s^t G_n(t-r)g(r-s)\,dr\,\|u(s)\|\,ds$$

$$\leqq \int_0^t G_n(t-r)\int_0^r g(r-s)\|u(s)\|\,ds\,dr$$

$$\leqq \sup_{r\in T}\left(\int_0^r g^2(r-s)\,ds\right)^{1/2}$$

$$\cdot\int_0^t G_n(t-r)\,dr\,\|u\|_{L_2(T;U)} \quad \text{by (2.5)}.$$

Now since $\sum_{n=1}^{\infty}\int_0^t G_n(t-\alpha)\,d\alpha<\infty$ (Appendix A, (A.4)) we see that

$$\mathcal{U}_{BF}(t,s;B)u(s)=\sum_{n=0}^{\infty}V_n(t,s;B)u(s)$$

converges absolutely in $L_1(0,t;H)$ for each $t\in[0,T]$ and

$$\mathcal{U}_{BF}(t,\cdot\,;B)\in\mathcal{L}(L_2(T;U),L_1(0,t;H)).$$

Moreover

$$\|\mathcal{U}_{BF}(t, s; B)u(s)\| \leqq \bar{g}(t-s)\|u(s)\|,$$

where

$$\bar{g}(t-s) = g(t-s) + \sum_{n=1}^{\infty} \int_s^t G_n(t-r)g(r-s)\, dr$$

or

$$\bar{g}(t-s) = g(t-s) \int_0^{t-s} G(t-s-r)\bar{g}(r)\, dr$$

from Appendix A and $\bar{g} \in L_2(T)$.

For the uniqueness let $\mathcal{U}_2(t, s; B)$ be another solution and

$$V(t, s) = \mathcal{U}_{BF}(t, s; B) - \mathcal{U}_2(t, s; B);$$

then

$$V(t, s)u(s) = \int_s^t \mathcal{U}(t, r; B)F(r)V(r, s)u(s)\, dr.$$

Hence

$$\|V(t, s)u(s)\| \leqq \int_s^t G(t-r)\|V(r, s)u(s)\|\, dr.$$

So $V(t, s)u(s) = 0$ by the generalized Gronwall's inequality (Appendix A, (A.3)).

For the semigroup property we consider

$$\begin{aligned}\mathcal{U}_{BF}(t, r)\mathcal{U}_{BF}(r, s; B)v = {} & \mathcal{U}(t, s; B)v \\ & + \int_s^r \mathcal{U}(t, \alpha; B)F(\alpha)\mathcal{U}_{BF}(\alpha, s; B)v\, d\alpha \\ & + \int_r^t \mathcal{U}(t, \alpha; B)F(\alpha)\mathcal{U}_{BF}(\alpha, r)\mathcal{U}_{BF}(r, s; B)v\, d\alpha.\end{aligned}$$

Thus

$$\begin{aligned}&[\mathcal{U}_{BF}(t, r)\mathcal{U}_{BF}(r, s; B) - \mathcal{U}_{BF}(t, s; B)]v \\ &\quad = \int_r^t \mathcal{U}(t, \alpha; B)F(\alpha)[\mathcal{U}_{BF}(\alpha, r)\mathcal{U}_{BF}(r, s; B) - \mathcal{U}_{BF}(\alpha, s; B)]v\, d\alpha.\end{aligned}$$

Denoting the left-hand side by $V(t, r, s; B)v$, we find

$$V(t, r, s; B)v = \int_r^t \mathcal{U}(t, \alpha; B)F(\alpha)V(\alpha, r, s; B)v\, d\alpha.$$

Hence $V(t, r, s; B)v = 0$ by the generalized Gronwall's inequality (Appendix A, (A.5)).

Thus $\mathcal{U}_{BF}(t, s; B)$ satisfies the same assumptions (2.4)–(2.5) as $\mathcal{U}(t, s; B)$ with $g(t-s)$ being replaced by $\bar{g}(t-s)$ and $\bar{g}$ satisfies (2.6).

We recall the concept of a quasi-evolution operator introduced in [2].

DEFINITION 2.2. Let H be a Hilbert space, T a real time interval. A *quasi-evolution operator* is a mild evolution operator $\mathcal{U}(t, s)$ such that there exists

a nonzero $x \in H$ and a closed linear operator $A(s)$ on H for $0 \leqq s \leqq T$ satisfying

$$\langle y, \mathcal{U}(t, s)x - x\rangle = \int_s^t \langle y, \mathcal{U}(t, \alpha)A(\alpha)x\rangle \, d\alpha \tag{2.14}$$

for all $y \in H$. The set of $x \in H$ for which (2.14) is valid is denoted by $\mathcal{D}_A$ and $A(\cdot)$ is called the *generator of* $\mathcal{U}(\cdot, \cdot)$.

In line with assumption (2.2)′, we assume the stronger condition

$$\mathcal{U}(t, s)x - x = \int_s^t \mathcal{U}(t, \alpha)A(\alpha)x \, d\alpha \tag{2.14}'$$

although it is in no way essential. We now show that if $\mathcal{U}(t, s)$ is a quasi-evolution operator satisfying (2.14)′, then the perturbed evolution operator $\mathcal{U}_{BF}(t, s)$ is also a quasi-evolution operator.

THEOREM 2.3. *If $\mathcal{U}(t, s)$ is a quasi-evolution operator with generator $A(t)$, $\mathcal{U}(t, s; B)$ satisfies* (2.4), (2.6), *and $F(t)$ satisfies* (2.8), *then $\mathcal{U}_{BF}(t, s)$ defined by* (2.7) *is a quasi-evolution operator, with*

$$\mathcal{U}_{BF}(t, s)x - x = \int_s^t (\mathcal{U}_{BF}(t, \alpha)A(\alpha)x + \mathcal{U}_{BF}(t, \alpha; B)F(\alpha)x) \, d\alpha \tag{2.15}$$

for $x \in \mathcal{D}_A$ and $\mathcal{U}_{BF}(t, s; B)$ defined by (2.12). (2.15) *implies that*

$$\frac{\partial}{\partial s}(\mathcal{U}_{BF}(t, s)x) = -\mathcal{U}_{BF}(t, s)A(s)x - \mathcal{U}_{BF}(t, s; B)F(s)x \tag{2.15}'$$

for almost all $s \in [0, t)$ and $x \in \mathcal{D}_A$.

Proof. For $x \in \mathcal{D}_A$, from (2.7), (2.14)′ we have

$$\begin{aligned}\int_s^t \mathcal{U}_{BF}(t, \alpha)A(\alpha)\, s\alpha &= \mathcal{U}(t, s)x - x + \int_s^t \int_\alpha^t \mathcal{U}(t, r; B)F(r) \\ &\qquad \cdot \mathcal{U}_{BF}(r, \alpha)A(\alpha)x \, dr \, d\alpha \\ &= \mathcal{U}(t, s)x - x + \int_s^t \mathcal{U}(t, r; B)F(r) \\ &\qquad \cdot \int_s^r \mathcal{U}_{BF}(r, \alpha)A(\alpha)x \, d\alpha \, dr.\end{aligned}$$

From (2.12)

$$\begin{aligned}\int_s^t \mathcal{U}_{BF}(t, \alpha; B)F(\alpha)x \, d\alpha &= \int_s^t \mathcal{U}(t, \alpha; B)F(\alpha)x \, d\alpha \\ &\quad + \int_s^t \int_\alpha^t \mathcal{U}(t, \alpha; B)F(\alpha)\mathcal{U}_{BF}(\alpha, \beta; B)F(\alpha)x \, d\alpha \, d\beta \\ &= \int_s^t \mathcal{U}(t, \alpha; B)F(\alpha)x \, d\alpha \\ &\quad + \int_s^t \mathcal{U}(t, \alpha; B)F(\alpha) \int_s^\alpha \mathcal{U}_{BF}(\alpha, \beta; B)F(\alpha)x \, d\beta \, d\alpha.\end{aligned}$$

Let

$$f(t, s)x = \int_s^t (\mathcal{U}_{BF}(t, \alpha)A(\alpha)x + \mathcal{U}_{BF}(t, \alpha; B)F(\alpha)x)\, d\alpha.$$

Then

$$f(t, s)x = \mathcal{U}(t, s)x - x + \int_s^t \mathcal{U}(t, \alpha; B)F(\alpha)x\, d\alpha + \int_s^t \mathcal{U}(t, r; B)F(r)f(r, s)x\, dr$$

$$= \mathcal{U}_{BF}(t, s)x - x - \int_s^t \mathcal{U}(t, r; B)F(r)\mathcal{U}_{BF}(r, s)x\, dr$$

$$+ \int_s^t \mathcal{U}(t, \alpha; B)F(\alpha)x\, d\alpha + \int_s^t \mathcal{U}(t, r; B)F(r)f(r, s)x\, dr.$$

Hence

$$f(t, s)x - \mathcal{U}_{BF}(t, s)x + x = \int_s^t \mathcal{U}(t, r; B)F(r)[f(r, s)x - \mathcal{U}_{BF}(t, s)x - x]\, dr;$$

letting

$$R(r, s)x = f(t, s)x - \mathcal{U}_{BF}(t, s)x + x,$$

we have

$$\|R(t, s)x\| \leqq \int_s^t G(t-r)\|R(t, s)x\|\, dr.$$

Thus $R(t, s)x = 0$ by the generalized Gronwall's inequality (Appendix A, (A.3)) and (2.15) is established.

3. The infinite dimensional Riccati equation. We consider the following generalized control problem:

$$z(t) = \mathcal{U}(t, t_0)z_0 + \int_{t_0}^t \mathcal{U}(t, s; B)u(s)\, ds, \qquad 0 \leqq t_0 \leqq t \leqq T, \tag{3.1}$$

where U, H are real Hilbert spaces and the admissible controls $u \in L_2(t_0, T; U)$. $\mathcal{U}(t, s)$ is a mild evolution operator satisfying (2.1), (2.2) and $\mathcal{U}(t, s; B)$ satisfies (2.4)–(2.6).

For the cost functional, we take

$$\mathcal{C}(u; t_0, z_0) = \langle z(T), Gz(T)\rangle_H + \int_{t_0}^T \langle z(s), W(s)z(s)\rangle_H\, ds + \int_{t_0}^T \langle u(s), R(s)u(s)\rangle_{\mathcal{U}}\, ds, \tag{3.2}$$

where $G \in \mathcal{L}(H)$, $W \in \mathcal{B}_\infty(T; \mathcal{L}(H))$, R, $R^{-1} \in \mathcal{B}_\infty(T; \mathcal{L}(U))$ and G, W and R are self adjoint and nonnegative definite with

$$\langle u, R(t)u\rangle_U \geqq \beta\|u\|_U^2 \quad \text{for almost all } t \in T \text{ and all } u \in U.$$

The quadratic cost control problem is then to find the optimal control $u \in L_2(T; U)$ which minimizes $\mathscr{C}(u; t_0, z_0)$. We proceed along similar lines to [2] allowing for a wider class of controls which in some sense may be considered unbounded (cf. § 1).

Consider a sequence of controls $\{u_k\}$ given by

$$u_k(t) = F_k(t)z(t),$$

where $F_k(t)$ is defined recursively by

$$F_k(t) = -R^{-1}(t)Q_{k-1}^*(t; B), \qquad F_0 = 0, \tag{3.3}$$

$$\begin{aligned} Q_k(t)x &= \mathscr{U}_k^*(T, t)G\mathscr{U}_k(T, t)x \\ &\quad + \int_t^T \mathscr{U}_k^*(s, t)(W(s) + F_k^*(s)R(s)F_k(s))\mathscr{U}_k(s, t)x\, ds, \end{aligned} \tag{3.4}$$

$$\begin{aligned} Q_k(t; B)u(t) &= \mathscr{U}_k^*(T, t)G\mathscr{U}_k(T, t; B)u(t) \\ &\quad + \int_t^T \mathscr{U}_k^*(s, t)(W(s) + F_k^*(s)R(s)F_k(s))\mathscr{U}_k(s, t; B)u(t)\, ds, \end{aligned} \tag{3.5}$$

$$\begin{aligned} Q_k^*(t; B)h(t) &= \mathscr{U}_k^*(T, t; B)G\mathscr{U}_k(T, t)h(t) \\ &\quad + \int_t^T \mathscr{U}_k^*(s, t; B)(W(s) + F_k^*(s)R(s)F_k(s))\mathscr{U}_k(s, t)h(t)\, ds, \end{aligned} \tag{3.6}$$

$$\begin{aligned} Q_k(t; B, B^*)u(t) &= \mathscr{U}_k^*(T, t; B)G\mathscr{U}_k(T, t; B)u(t) \\ &\quad + \int_t^T \mathscr{U}_k^*(s, t; B)(W(s) + F_k^*(s)R(s)F_k(s)) \\ &\qquad \cdot \mathscr{U}_k(s, t; B)u(t)\, ds, \end{aligned} \tag{3.7}$$

$$\mathscr{U}_k(t, s)x = \mathscr{U}(t, s)x + \int_s^t \mathscr{U}(t, \alpha; B)F_k(\alpha)\mathscr{U}_k(\alpha, s)x\, d\alpha, \tag{3.8}$$

$$\begin{aligned} \mathscr{U}_k(t, s; B)u(s) &= \mathscr{U}(t, s; B)u(s) \\ &\quad + \int_s^t \mathscr{U}(t, \alpha; B)F_k(\alpha)\mathscr{U}_k(\alpha, s; B)u(s)\, d\alpha, \end{aligned} \tag{3.9}$$

where $x \in H$, $h \in L_2(T; H)$ or $C(T; H)$ and $u \in L_2(T; U)$ or $C(T; U)$.

To establish that this sequence is well defined, and to interpret (3.5), (3.6) we need the following lemmas.

LEMMA 3.1.

$$\begin{aligned} &Q_k(\cdot\,; B) \in \mathscr{L}(C(T; U), L_2(T; H)) \cap \mathscr{L}(L_2(T; U), L_1(T; H)), \\ &Q_k^*(\cdot\,; B) \in \mathscr{L}(C(T; H), L_2(T; U)) \cap \mathscr{L}(L_2(T; H), L_1(T; U)), \\ &Q_k(\cdot\,; B, B^*) \in \mathscr{L}(C(T; U), L_1(T; H)). \end{aligned} \tag{3.10}$$

$F_k(t)$ satisfies an estimate of the form

$$\|F_k(t)h\| \leqq f_k(t_2 - t)\|h\|, \quad \textit{where } f_k \in L_2(T) \quad \textit{and} \quad t_2 \in (t, T]. \tag{3.11}$$

$\mathcal{U}_k(t, s)$ is well defined by (3.8) *satisfying an estimate of the form*

$$\|\mathcal{U}_k(t, s)\| \leqq M_k. \tag{3.12}$$

$\mathcal{U}_k(t, s; B)$ is well defined by (3.9) *in the sense of* (2.12) *and has the same properties as $\mathcal{U}(t, s; B)$* (2.4)–(2.6), *satisfying an estimate of the form*

$$\|\mathcal{U}_k(t, s; B)u\| \leqq g_k(t-s)\|u\|, \tag{3.13}$$

where $g_k \in L_2(T)$.

Proof. See Appendix B.

LEMMA 3.2. *$Q_k(t)$ is a linear, self adjoint, bounded operator on H, which is strongly measurable. $Q_k(t)x$ is weakly continuous on T under the stronger assumption* (2.2)′.

Proof. See Appendix C.

By Lemma 3.1 and Theorems 2.1, 2.2 we see that the sequence of control problems

$$z_k(t) = \mathcal{U}_k(t, t_0)z_0 + \int_{t_0}^{t} \mathcal{U}_k(t, s; B)\bar{u}(s)\, ds \tag{3.14}$$

is well defined for $\bar{u} \in L_2(T; U)$, and $z_k \in C(T; H)$.

Just as in [2], the following lemmas are easily established by replacing $Q_k(t)B(t)$ by $Q_k(t; B)$.

LEMMA 3.3.

$$\begin{aligned}\langle z_k(t), Q_k(t)z_k(t)\rangle = {} & \langle z_k(T), Gz_k(T)\rangle \\ & + \int_t^T \langle z_k(s), (W(s) + F_k(s)R(s)F_k(s))z(s)\rangle\, ds \\ & - 2\int_t^T \langle z_k(s), Q_k(s; B)\bar{u}(s)\rangle\, ds.\end{aligned}$$

LEMMA 3.4. *With the feedback control $u_k(t) = -R^{-1}(t)Q_{k-1}^{*}(t; B)z_k(t)$, the cost is given by*

$$J(u_k) = \langle z_0, Q_k(t_0)z_0\rangle.$$

Furthermore, $\langle z_0, Q_k(t_0)z_0\rangle$ is monotonically decreasing in k for each $t \in T$, $z_0 \in H$, with

$$\sup_{t\in T} \|Q_k(t)\| \leqq K_1. \tag{3.15}$$

LEMMA 3.5. *$Q_k(t)$ defined by* (3.4)–(3.8) *converges strongly as $k \to \infty$ to a self-adjoint nonnegative definite bounded linear operator $Q(t)$ on H with $\sup_{t\in T} \|Q(t)\| \leqq K_1$.*

In order to prove the convergence of $\mathcal{U}_k(t, s)$, $\mathcal{U}_k(t, s; B)$ and hence $Q_k(t; B)$ we need the following lemmas.

LEMMA 3.6. (a) *$\mathcal{U}_k(t, s+\varepsilon)\mathcal{U}(s+\varepsilon, s; B)$ converges to $\mathcal{U}_k(t, s; B)$ in*

$$\mathcal{L}(C(t_0, t; U), L_2(t_0, t; H)) \cap \mathcal{L}(L_2(t_0, t; U), (L_2(t_0, t; H))$$

as $\varepsilon \to 0$.

(b) $\mathcal{U}^*(t+\varepsilon, t; B)Q_k(t+\varepsilon)$ *converges to* $Q_k^\#(t; B)$ *in*

$$\mathcal{L}(C(T; H), L_2(T; U)) \cap \mathcal{L}(L_2(T; H), L_1(T; U))$$

as $\varepsilon \to 0$.

(c) $\mathcal{U}^*(t+\varepsilon, t; B)Q_k(t+\varepsilon)\mathcal{U}(t+\varepsilon, t; B)$ *converges to* $Q_k(t; B, B^*)$ *in*

$$\mathcal{L}(C(T; U), L_1(T; U))$$

as $\varepsilon \to 0$.

Proof. See Appendix D.

LEMMA 3.7. $\langle Q_k(t; B, B^*)u_0, u_0\rangle$ *is a positive decreasing sequence in* k *for almost all* $t \in T$, *and all* $u_0 \in U$, *and so is uniformly bounded for almost all* $t \in T$.

Proof. See Appendix E.

LEMMA 3.8. $\mathcal{U}_k(t, s)$, $\mathcal{U}_k(t, s; B)$, $Q_k(t; B)$, $Q_k^\#(t; B)$, $Q_k(t; B, B^*)$ *are uniformly bounded in* k *by estimates of the same form and the convergence in Lemma* 3 *is uniform in* k.

Proof. See Appendix F.

THEOREM 3.1. $U_k(t, s)$ *converges strongly in* H *to a mild evolution operator* $\mathcal{U}_\infty(t, s)$; $\mathcal{U}_k(t, s; B)$ *converges strongly in* $\mathcal{L}(C(t_0, t; U), L_2(t_0, t; H)) \cap \mathcal{L}(L_2(t_0, t; U), L_1(t_0, t; H))$ *to* $\mathcal{U}_\infty(t, s; B)$ *and* $Q_k(t; B)$ *converges strongly to* $Q_\infty(t; B)$ *in* $\mathcal{L}(C(T; U), L_2(T; H)) \cap \mathcal{L}(L_2(T; U), L_1(T; H))$. *Furthermore* $Q_\infty(t)$ *and* $Q_\infty(t; B)$ *satisfy the integral equations*

$$\begin{aligned} Q_\infty(t)x &= \mathcal{U}_\infty^*(T, t)G\mathcal{U}_\infty(T, t)x \\ &\quad + \int_t^T \mathcal{U}_\infty^*(s, t)(W(s) + Q_\infty(s; B)R^{-1}(s)Q_\infty^\#(s; B))\mathcal{U}_\infty(s, t)x\, ds, \end{aligned} \tag{3.16}$$

$$\begin{aligned} Q_\infty(t; B)u(t) &= \mathcal{U}_\infty^*(T, t)G\mathcal{U}_\infty(T, t; B)u(t) \\ &\quad + \int_t^T \mathcal{U}_\infty^*(s, t)[W(s) + Q_\infty(s; B)R^{-1}(s)Q_\infty^\#(s; B)] \\ &\qquad \cdot \mathcal{U}_\infty(s, t; B)x\, ds, \end{aligned} \tag{3.17}$$

where $\mathcal{U}_\infty(t, s)$ *and* $\mathcal{U}_\infty(t, s; B)$ *are the unique solutions of*

$$\begin{aligned} \mathcal{U}_\infty(t, s)x &= \mathcal{U}(t, s)x \\ &\quad - \int_s^t \mathcal{U}(t, \alpha; B)R^{-1}(\alpha)Q_\infty^\#(\alpha; B)\mathcal{U}_\infty(\alpha, s)x\, d\alpha, \end{aligned} \tag{3.18}$$

$$\begin{aligned} \mathcal{U}_\infty(t, s; B)u(s) &= \mathcal{U}(t, s; B)u(s) \\ &\quad - \int_s^t \mathcal{U}(r, \alpha; B)R^{-1}(\alpha)Q_\infty^\#(\alpha; B)\mathcal{U}_\infty(\alpha, s; B)u(s)\, d\alpha. \end{aligned} \tag{3.19}$$

(Of course (3.17) and (3.19) must be interpreted in the appropriate spaces.)

Proof. (a) Now from (3.4)

$$\begin{aligned}\langle Q_k(t)x, x\rangle &= \langle \mathcal{U}_k(T,t)x, G\mathcal{U}_k(T,t)x\rangle \\ &\quad + \int_t^T \langle \mathcal{U}_k(s,t)x, W(s)\mathcal{U}_k(s,t)x\rangle\, ds \\ &\quad + \int_t^T \langle \mathcal{U}_k(s,t)x, F_k^*(s)R(s)F_k(s)\mathcal{U}_k(s,t)x\rangle\, ds \\ &\leqq K_1\|x\|^2.\end{aligned}$$

Hence since $\langle u, R(t)u\rangle \geqq \beta\|u\|^2$, we have

$$\int_t^T \|F_k(s)\mathcal{U}_k(s,t)x\|^2\, ds \leqq C\|x\|^2. \tag{3.20}$$

Then from (3.7),

$$\begin{aligned}\|\mathcal{U}_k(t,s)x\| &\leqq M\|x\| + \int_s^t g(t-\alpha)\|F_k(\alpha)\mathcal{U}_k(\alpha,s)x\|\, d\alpha \\ &\leqq M\|x\| + \left(\int_s^t g^2(t-\alpha)\, d\alpha\right)^{1/2}\left(\int_s^t \|F_k(\alpha)\mathcal{U}_k(\alpha,s)x\|^2\, d\alpha\right)^{1/2} \\ &\leqq K_2\|x\| \quad \text{by (2.5) and (3.16).}\end{aligned}$$

From Lemma 3.3 with $\bar{u}(s) = -\langle F_{k+1}(s) - F_k(s))z(s)$, $z(t) = z_0$, we obtain

$$\begin{aligned}\langle z_0, Q_k(t)z_0\rangle &= \langle z_0, Q_{k+1}(t)z_0\rangle \\ &\quad + \int_t^T \langle (F_{k+1}(s) - F_k(s))\mathcal{U}_{k+1}(s,t)z_0, \\ &\qquad R(s)(F_{k+1})(s) - F_k(s))\mathcal{U}_{k+1}(s,t)z_0\rangle\, ds.\end{aligned}$$

Hence using the positivity of R and the convergence of $Q_k(t)$, we find

$$\int_t^T \|(F_{k+1}(s) - F_k(s))\mathcal{U}_{k+1}(s,t)z_0\|^2\, ds \to 0 \quad \text{as } k \to \infty. \tag{3.21}$$

Now

$$\begin{aligned}&\|\mathcal{U}_{k+1}(\alpha,t)z_0 - \mathcal{U}_k(\alpha,t)z_0\| \\ &\qquad = \left\| \int_t^{\vartheta} (\mathcal{U}(\alpha,s;B)F_{k+1}(s)\mathcal{U}_{k+1}(s,t)z_0 - F_k(s)\mathcal{U}_k(s,t)z_0)\, ds \right\|.\end{aligned}$$

Hence

$$\begin{aligned}\|\mathcal{U}_{k+1}(\alpha,t)z_0 - \mathcal{U}_k(\alpha,t)z_0\| &\leqq \int_t^{\alpha} g(\alpha-s)\|F_{k+1}(s)\mathcal{U}_{k+1}(s,t)z_0 \\ &\qquad - F_k(s)\mathcal{U}_k(s,t)z_0\|\, ds\end{aligned}$$

$$\leqq \text{const.} \left(\int_t^\alpha \|(F_{k+1}(s) - F_k(s))\mathcal{U}_{k+1}(s, t)z_0\|^2 \, ds \right)^{1/2}$$

$$+ \int_t^\alpha g(\alpha - s)\|F_k(s)(\mathcal{U}_{k+1}(s, t) - \mathcal{U}_k(s, t))z_0\| \, ds$$

$$\leqq \text{const.} \left(\int_t^\alpha \|(F_{k+1}(s) - F_k(s))\mathcal{U}_{k+1}(s, t)z_0\|^2 \, ds \right)^{1/2}$$

$$+ \int_t^\alpha g(\alpha - s)h(\alpha - s)\|\mathcal{U}_{k+1}(s, t) - \mathcal{U}_k(s, t)z_0\| \, ds.$$

For some $h \in L_2(T)$ by Lemma 3.8 and using (3.21) and the generalized Gronwall's inequality (Appendix A, (A.3)) we see that $\mathcal{U}_k(t, s)z_0$ converges in H as $k \to \infty$.

(b) Next we show that $\mathcal{U}_k(t, s; B)$ has a strong limit by proving that for $u \in U$,

$$a_{kn}(t) = \int_{t_0}^t \|\mathcal{U}_k(t, s; B)u - \mathcal{U}_n(t, s; B)u\|^2 \, ds \to 0 \quad \text{as } k, n \to \infty.$$

Now

$$a_{kn}(t) \leqq 3 \int_{t_0}^t \|\mathcal{U}_k(t, s; B)u - \mathcal{U}_k(t, s + \varepsilon)\mathcal{U}(s + \varepsilon, s; B)u\|^2 \, ds$$

$$+ 3 \int_{t_0}^t \|\mathcal{U}_k(t, s + \varepsilon) - \mathcal{U}_n(t, s + \varepsilon)\mathcal{U}(s + \varepsilon, s; B)u\|^2 \, ds$$

$$+ 3 \int_{t_0}^t \|\mathcal{U}_n(t, s + \varepsilon)\mathcal{U}(s + \varepsilon, s; B)u - \mathcal{U}_n(t, s; B)u\|^2 \, ds$$

$$\to 0 \quad \text{as } k, n \to \infty$$

by (a), Lemma 3.6, and since (2.4) justifies the use of the Lebesgue dominated convergence theorem on the middle term.

Hence $\mathcal{U}_k(t, s; B)$ is Cauchy in both $\mathcal{L}(C(t_0, t; U), L_2(t_0, t; H))$ and $\mathcal{L}(L_2(t_0, t; U), L_1(t_0, t; H))$ and since these are complete $\mathcal{U}_\infty(t, s; B)$ exists. Similarly $Q_k(t; B)$ has a strong limit since

$$\int_0^T \|(Q_k(t; B) - Q_n(t; B))u\|^2 \, dt \leqq 3 \int_0^T \|Q_k(t; B)u - Q_k(t + \varepsilon)\mathcal{U}(t + \varepsilon, t; B)u\|^2 \, dt$$

$$+ 3 \int_0^T \|(Q_k(t + \varepsilon) - Q_n(t + \varepsilon))\mathcal{U}(t + \varepsilon, t; B)u\|^2 \, dt$$

$$+ 3 \int_0^T \|(Q_n(t + \varepsilon)\mathcal{U}(t + \varepsilon, t; B) - Q_n(t; B))u\|^2 \, dt$$

$$\to 0 \quad \text{as } k, n \to \infty$$

by Lemmas 3.5 and 3.6.

(c) By Theorem 2.1, and (b), (3.18) has a unique solution which is a mild evolution operator. Denoting this solution by $V(t, s)$, we have

$$\begin{aligned}\|V(t,s)x-\mathcal{U}_k(t,s)x\| &\leqq \beta\int_s^t g(t-\alpha)\|Q_\infty^*(\alpha;B)-Q_k^*(\alpha;B))V(\alpha,s)x\|\,d\alpha\\ &\quad+\beta\int_s^t g(t-\alpha)\|Q_k^*(\alpha;B)(V(\alpha,s)-\mathcal{U}_k(\alpha,s))x\|\,d\alpha\\ &\leqq \beta\Big(\int_s^t g^2(t-\alpha)\,d\alpha\Big)^{1/2}\\ &\quad\cdot\Big(\int_s^t \|(Q_\infty^*(\alpha;B)-Q_k^*(\alpha;B))V(\alpha,s)x\|^2\,d\alpha\Big)^{1/2}\\ &\quad+\text{const.}\int_s^t g(t-\alpha)h(t-\alpha)\|V(\alpha,s)x-\mathcal{U}_k(\alpha,s)x\|\,d\alpha\end{aligned}$$

for some $h\in L_2(T)$ by Lemma 3.8. Now the first term on the right-hand side $\to 0$ as $k\to\infty$ by the convergence of $Q_k^*(t;B)$ in $\mathscr{L}(C(T;H),L_2(T;U))$. So for k sufficiently large

$$\|V(t,s)x-\mathcal{U}_k(t,s)x\|\leqq\varepsilon\|x\|+\int_s^t g(t-\alpha)g(t-\alpha)\|V(\alpha,s)x-\mathcal{U}_k(\alpha,s)x\|\,d\alpha$$

where ε is arbitrarily small. Hence by the generalized Gronwall's inequality (Appendix A, (A.3))

$$\|V(t,s)x-\mathcal{U}_k(t,s)x\|\to 0\quad\text{as } k\to\infty.$$

So $\mathcal{U}_k(t,s)$ converges strongly to the unique solution of (3.18).

(d) By Theorem 2.2 and (b), (3.19) has a unique solution which we denote by $V(t,s;B)$, so

$$\begin{aligned}\|V(t,s;B)u-\mathcal{U}_k(t,s;B)u\| &\leqq \beta\int_s^t g(t-\alpha)\|(Q_\infty^*(\alpha;B)\\ &\qquad -Q_k^*(\alpha;B))V(\alpha,s;B)u\|\,d\alpha\\ &\quad+\beta\int_s^t g(t-\alpha)\|Q_k^*(\alpha;B)\\ &\qquad\cdot(V(\alpha,s;B)u-\mathcal{U}_k(\alpha,s;B)u)\|\,d\alpha.\end{aligned}$$

Hence

$$\begin{aligned}&\int_{t_0}^t \|V(t,s;B)u-\mathcal{U}_k(t,s;B)u\|^2\,ds\\ &\quad\leqq 2\beta^2\int_{t_0}^t\int_s^t g^2(t-\alpha)\,d\alpha\int_s^t\|(Q_\infty^*(\alpha;B)-Q_k^*(\alpha;B))V(\alpha,s;B)u\|^2\,d\alpha\,ds\\ &\qquad+2\beta^2\int_{t_0}^t\int_s^t g^2(t-\alpha)\|Q_k^*(\alpha;B)(V(\alpha,s;B)-\mathcal{U}_k(\alpha,s;B)u\|\,d\alpha)^2\,ds\end{aligned}$$

$$\leqq \text{const.} \int_{t_0}^{t} \int_{t_0}^{\alpha} \|(Q_\infty^*(\alpha; B) - Q_K^*(\alpha; B)) V(\alpha, s; B)u\|^2 \, ds \, d\alpha$$

$$+ \text{const.} \int_{t_0}^{t} h^2(t-\alpha) \int_{t_0}^{\alpha} \|V(\alpha, s; B)u - \mathcal{U}_k(\alpha, s; B)u\|^2 \, ds \, d\alpha.$$

If we can show that the first term $\to 0$ as $k \to \infty$, then using the generalized Gronwall's lemma (A.3), we see that $\int_{t_0}^{t} \|(V(t, s; B) - \mathcal{U}_k(t, s; B))u\|^2 \, ds \to 0$ as $k \to \infty$ and $\mathcal{U}_k(t, s; B)$ converges to the unique solution of (3.19) $\mathcal{U}_\infty(t, s; B)$ in $\mathcal{L}(C(t_0, t; U), L_2(t_0, t; H)) \cap \mathcal{L}(L_2(t_0, t; U), L_1(t_0, t; H))$,

$$\int_{t_0}^{t} \int_{t_0}^{\alpha} \|(Q_k^*(\alpha; B) - Q_n^*(\alpha; B)) V(\alpha, s; B)u\|^2 \, ds \, d\alpha$$

$$\leqq 3 \int_{t_0}^{t} \int_{t_0}^{\alpha} \|(Q_k^*(\alpha; B) - \mathcal{U}^*(\alpha + \varepsilon, \alpha; B) Q_k^*(\alpha + \varepsilon)) V(\alpha, s; B)u\|^2 \, ds \, d\alpha$$

$$+ 3 \int_{t_0}^{t} \int_{t_0}^{\alpha} \|\mathcal{U}^*(\alpha + \varepsilon, \alpha; B) Q_n^*(\alpha + \varepsilon) - Q_n^*(\alpha; B) V(\alpha, s; B)u\|^2 \, ds \, d\alpha$$

$$+ 3 \int_{t_0}^{t} \int_{t_0}^{\alpha} \|\mathcal{U}^*(\alpha + \varepsilon, \alpha; B) Q_n^*(\alpha + \varepsilon) - Q_n^*(\alpha; B) V(\alpha, s; B)u\|^2 \, ds \, d\alpha$$

$$\leqq 3 \int_{t_0}^{t} \int_{t_0}^{\alpha} \|(Q_k^*(\alpha; B) - \mathcal{U}^*(\alpha + \varepsilon, \alpha; B) Q_k^*(\alpha + \varepsilon)) V(\alpha, s; B)u\|^2 \, ds \, d\alpha$$

$$+ 3 \int_{t_0}^{t} g^2(\varepsilon) \int_{t_0}^{\alpha} \|(Q_k^*(\alpha + \varepsilon) - Q_n^*(\alpha + \varepsilon)) V(\alpha, s; B)u\|^2 \, ds \, d\alpha$$

$$+ 3 \int_{t_0}^{t} \int_{t_0}^{\alpha} \|(\mathcal{U}^*(\alpha + \varepsilon, \alpha; B) Q_n^*(\alpha + \varepsilon) - Q_n^*(\alpha; B)) V(\alpha, s; B)u\|^2 \, ds \, d\alpha.$$

Since $Q_k(\alpha)$ converges by Theorem 3.1 and $\|V(\alpha, s; B)\| \leqq h(\alpha, s)$ with $h \in L_2(T)$, we have $\int_{t_0}^{t} \int_{t_0}^{\alpha} g^2(\varepsilon) h^2(\alpha - s) \, ds \, d\alpha < \infty$; and the Lebesgue dominated convergence theorem shows that the middle term $\to 0$ as $k \to \infty$. The other terms $\to 0$ as $\varepsilon \to 0$ by Lemma 3.7.

(e) From (3.4), we have

$$Q_k(t)x = \mathcal{U}_k^*(T, t) G \mathcal{U}_k(T, t)x$$

$$+ \int_{t}^{T} \mathcal{U}_k^*(s, t)[W(s) + Q_k(s; B) R^{-1}(s) Q_k^*(s; B)] \mathcal{U}_k(s, t)x \, ds,$$

and we may take limits to obtain (3.16) from the convergence of $\mathcal{U}_k(t, s)$, $Q_k(s; B)$ using the Lebesgue dominated convergence theorem. A typical term to

be estimated is

$$\int_t^T \|\mathcal{U}_k^*(s,t)Q_k^*(s;B)R^{-1}(s)(Q_k^*(s;B)-Q_\infty^*(s;B))\mathcal{U}_\infty(s,t)x\|\,ds$$

$$\leqq \text{const.}\int_t^T f_k(T-s)\|(Q_K^*(s;B)-Q_\infty^*(s;B))\mathcal{U}_\infty(s,t)x\|\,ds$$

$$\leqq \text{const.}\int_t^T \|(Q_k^*(s;B)-Q_\infty^*(s;B))\mathcal{U}_\infty(s,t)x\|^2\,ds)^{1/2} \qquad \text{(by Lemma 3.8)}$$

as $k\to\infty$ since $\mathcal{U}_\infty(s,t)x$ is continuous in s.

(f) From (3.5),

$$Q_k(t;B)u = \mathcal{U}_k^*(T,t)G\mathcal{U}_k(T,t;B)u + \int_t^T \mathcal{U}_k^*(s,t)(W(s)+Q_k(s;B)R^{-1}(s)Q_k^*(s;B))\mathcal{U}_k(s,t;B)u\,ds.$$

From (b)

$$\int_0^T \|Q_k(t;B)u - Q_\infty(t;B)u\|^2\,dt \to 0 \quad \text{as } k\to\infty$$

and we can show

$$\int_0^T \|\mathcal{U}_k^*(T,t)G\mathcal{U}_k(T,t)u - \mathcal{U}_\infty^*(T,t)G\mathcal{U}_\infty(T,t;B)u\|^2\,dt \to 0 \quad \text{as } k\to\infty.$$

Also

$$\int_0^T \Big\| \int_t^T \mathcal{U}_k^*(s,t)(W(s)+Q_k(s;B)R^{-1}(s)Q_k^*(s;B))\mathcal{U}_k(s,t;B)u - \int_t^T \mathcal{U}_\infty^*(s,t)(W(s)+Q_\infty(s;B)R^{-1}(s)Q_\infty^*(s;B))\mathcal{U}_\infty(s,t;B)u \Big\|^2 dt \to 0$$

as $k\to\infty$ using our convergence results for $\mathcal{U}_k(t,s)$, $\mathcal{U}_k(t,s;B)$, $Q_k(t,s;B)$; and the uniform estimates of Appendix D ensure the applicability of the Lebesgue dominated convergence theorem. So $Q_\infty(t;B)$ satisfies (3.19) in the space

$$\mathcal{L}(C(T;U),L_2(T;H))\cap\mathcal{L}(L_2(T;U),L_1(T;H)).$$

As in [2] the following result is proved by a simple application of Lemma 3.3.

COROLLARY 3.1. The optimal control which minimizes $\mathcal{C}(u;t_0,z_0)$ is the feedback control

$$u_\infty(t) = -R^{-1}(t)Q_\infty^*(t;B)z(t).$$

It is possible to convert (3.16), (3.17) into simpler equivalent forms which we will find useful.

THEOREM 3.2. *The equations* (3.16), (3.17) *are equivalent to*

$$Q(t)x = \mathcal{U}_\infty^*(T,t)G\mathcal{U}(T,t)x + \int_t^T \mathcal{U}_\infty^*(s,t)W(s)\mathcal{U}(s,t)x\,ds, \tag{3.22}$$

$$(3.23)\quad Q_\infty(t;B)u = \mathcal{U}_\infty^*(T,t)G\mathcal{U}(T,t;B)u + \int_t^T \mathcal{U}_\infty^*(s,t)W(s)\mathcal{U}(s,t;B)u\,ds,$$

where $\mathcal{U}_\infty(t,s)$ *is the unique solution of* (3.22).

Proof. Let

$$(3.24)\quad P(t;B)u = \mathcal{U}_p^*(T,t)G\mathcal{U}(T,t;B)u + \int_t^T \mathcal{U}_p^*(s,t)W(s)\mathcal{U}(s,t;B)u\,ds,$$

where

$$(3.25)\quad \mathcal{U}_p(t,s;B)u = \mathcal{U}(t,s;B)u - \int_s^t \mathcal{U}(t,\alpha;B)R^{-1}(\alpha)P^*(\alpha;B)\mathcal{U}_p(\alpha,s;B)u\,d\alpha$$

and

$$(3.26)\quad \mathcal{U}_p(t,s)x = \mathcal{U}(t,s)x - \int_s^t \mathcal{U}(t,\alpha;B)R^{-1}(\alpha)P^*(\alpha;B)\mathcal{U}_p(\alpha,s)x\,d\alpha.$$

Then

$$\begin{aligned}P(t;B)u = {}& \mathcal{U}_p^*(T,t)G\mathcal{U}_p(T,t;B)\\ &+ \mathcal{U}_p^*(T,t)G\int_t^T \mathcal{U}(T,\alpha;B)R^{-1}(\alpha)P^*(\alpha;B)\mathcal{U}_p(\alpha,t;B)u\,d\alpha\\ &+ \int_t^T \mathcal{U}_p^*(s,t)W(s)\mathcal{U}_p(s,t;B)u\,ds\\ &+ \int_t^T\int_\alpha^T \mathcal{U}_p^*(s,t)W(s)\mathcal{U}(s,\alpha;B)R^{-1}(\alpha)P^*(\alpha;B)\mathcal{U}_p(\alpha,t;B)u\,ds\,d\alpha.\end{aligned}$$

Hence using (3.24),

$$\begin{aligned}P(t;B)u = {}& \mathcal{U}_p^*(T,t)G\mathcal{U}_p(T,t;B)u + \int_t^T \mathcal{U}_p^*(s,t)W(s)\mathcal{U}_p(s,t;B)u\,ds\\ &+ \int_t^T \mathcal{U}_p^*(\alpha,t)P(\alpha;B)R^{-1}(\alpha)P^*(\alpha;B)\mathcal{U}_p(\alpha,t;B)u\,d\alpha.\end{aligned}$$

Similarly

$$\begin{aligned}P(t)x &= \mathcal{U}_p^*(T,t)G\mathcal{U}(T,t)x + \int_t^T \mathcal{U}_p^*(s,t)W(s)\mathcal{U}(s,t)x\,ds\\ &= \mathcal{U}_p(T,t)G\mathcal{U}(T,t)x + \int_t^T \mathcal{U}_p^*(s,t)W(s)\mathcal{U}_p(s,t)x\,ds\\ &\quad \int_t^T \mathcal{U}_p^*(\alpha,t)P(\alpha;B)R^{-1}(\alpha)P^*(\alpha;B)\mathcal{U}_p(\alpha,t)x\,d\alpha.\end{aligned}\quad (3.27)$$

So the systems (3.16), (3.17) are equivalent to (3.22), (3.23) if either system yields a unique solution. But as in Theorem 2.3 [2], you can prove that (3.18)–(3.21) yields a unique solution $Q(t)$ for (3.18) in the class of self adjoint bounded linear

operator in $\mathscr{B}_\infty(T; \mathscr{L}(H))$, and from Lemma 3.2 we obtain $Q(t)$ is weakly continuous on T. As in Lemma 3.6 it follows that $Q(t+\varepsilon)\mathscr{U}(t+\varepsilon, t; B)$ converges to $Q_\infty^*(t$ in $\mathscr{L}(C(T; U), L_2(T; H)) \cap \mathscr{L}(L_2(T; U), L_1(T; H))$ as $\varepsilon \to 0$ and this yields the uniqueness of solutions of (3.19).

COROLLARY 3.2. *$Q(t)$, $Q_\infty(t; B)$ are the unique solutions of the equivalent systems* (3.18)–(3.19) *or* (3.22), (3.23), (3.17), (3.19), *in the class of bounded self adjoint weakly continuous operators on H and the class $\mathscr{L}(C(T; U), L_2(T; H)) \cap \mathscr{L}(L_2(T; U), L_1(T; H))$ respectively.*

In [2] we obtained a differential Riccati equation by differentiating (3.16) when $\mathscr{U}(t, s)$ was a quasi-evolution operator, and here we obtain a similar result using the following lemma from [2].

LEMMA 3.9. *If $g_i(t)$, $i = 1, 2$, are weakly absolutely continuous H-valued functions, such that $\langle g_i(t), x\rangle = \langle g_i(0), x\rangle + \int_0^t \partial/\partial s \langle g_i(s), x\rangle\, ds$ for $x \in H$, $i = 1, 2$, then $f(t) = \langle Wg_1(t), g_2(t)\rangle$ is absolutely continuous with*

$$\langle Wg_1(s), g_2(t)\rangle = \langle Wg_1(0), g_2(0)\rangle + \int_0^t \frac{\partial}{\partial s}\langle Wg_1(s), g_2(s)\rangle\, ds$$

THEOREM 3.3. *Under the additional assumption that $\mathscr{U}(t, s)$ is a quasi-evolution operator with generator $A(t)$, $Q(t)$ also satisfies the differential Riccati equation*

$$\frac{d}{dt}\langle Q(t)x, y\rangle + \langle Q(t)x, A(t)y\rangle + \langle A(t)x, Q(t)y\rangle$$

$$-\langle R^{-1}(t)Q^*(t; B)x, Q^*(t; B)y\rangle + \langle W(t)x, y\rangle = 0 \quad a.e., \tag{3.28}$$

$$Q(T) = G \quad for\ x, y \in \mathscr{D}_A.$$

Furthermore if $\mathscr{U}(t, s)$ is a strong evolution operator and $\mathscr{D}_A = \bigcap_{s\in T} \mathscr{D}(A(s))$ is dense in H, $Q(t)$ is the unique solution of (3.28) *in the following class of operators on $\mathscr{L}(H)$:*

(3.29) (a) *$Q(t)$ is weakly continuous on T,*

(b) *$\langle Q(t)x, y\rangle$ is absolutely continuous for $x, y \in \mathscr{D}_A$,*

(c) *$\mathscr{U}^*(t+\varepsilon, t; B)Q(t+\varepsilon)$ has the strong limit $Q^*(t; B)$ as $\varepsilon \to 0$ in $\mathscr{L}(C(T; H), L_2(T; U)) \cap \mathscr{L}(L_2(T; H), L_1(T; U))$ with*

$$\|Q^*(t; B)x\| \leqq \bar{f}(t_2 - t)\|x\|, \qquad \bar{f} \in L_2(T), \quad t_2 \in (t, T].$$

Proof. (a) From Theorem 2.3, we have

$$\frac{\partial}{\partial s}\mathscr{U}_\infty(t, s)x = -\mathscr{U}_\infty(t, s)A(s)x - \mathscr{U}_\infty(t, s; B)R^{-1}(s)Q_\infty^*(s; B)x \quad \text{a.e.}$$

for $x \in \mathscr{D}_A$. Using this expression it is readily verified that the formal differentiation of (3.16) yields (3.28) and this formal differentiation "under the integral procedure" is justified by appealing to Lemma 3.9.

(b) *Uniqueness.* Initially this follows along similar lines to the proof in [2] (Theorem 2.5) so we outline this briefly.

Let $P_i(t)$, $P_i(t; B)$, $i = 1, 2$, be solutions of (3.28) in the specified class and let $Q(t) = P_1(t) - P_2(t)$, $Q^*(t; B) = P_1^*(t; B) - P_2^*(t; B)$. Then it is easily verified that

$$\begin{aligned}\frac{d}{dt}\langle Q(t)x, x\rangle &= -2\langle A(t)x, Q(t)x\rangle + 2\langle R^{-1}(t)P_1^*(t; B)x, Q^*(t; B)x\rangle \\ &\quad -\langle Q(t; B)R^{-1}(t)Q^*(t; B)x, x\rangle \quad \text{a.e.}\end{aligned} \tag{3.30}$$

$$\begin{aligned}\frac{d}{dt}\langle Q(t)x, x\rangle &= -2\langle A(t)x, Q(t)x\rangle + 2\langle R^{-1}(t)P_2^*(t; B)x, Q^*(t; B)x\rangle \\ &\quad +\langle Q(t; B)R^{-1}(t)Q^*(t; B)x, x\rangle \quad \text{a.e.}\end{aligned} \tag{3.31}$$

Let

$$F(t)x = \int_t^T \mathcal{U}_1^*(s, t)Q(s; B)R^{-1}(s)Q^*(s; B)\mathcal{U}_1(s, t)x\, ds$$

and

$$F(t; B)x = \int_t^T \mathcal{U}_1^*(s, t)Q(s; B)R^{-1}(s)Q^*(s; B)\mathcal{U}_1(s, t; B)x\, ds,$$

where $\mathcal{U}_1(t, s)$, $\mathcal{U}_1(t, s; B)$ are the perturbations corresponding to $P_1^*(t; B)$ (cf. (3.18), (3.19)). Then we note that $F(t)$ has properties (3.29) (a), (b), (c), and by Lemma 3.9, we may differentiate $\langle F(t)x, x\rangle$ for $x \in \mathcal{D}_A$ and subtracting this from (3.30) yields

$$\begin{aligned}&\frac{d}{dt}\langle (Q(t)-F(t))x, x\rangle - 2\langle (Q(t)-F(t))x, A(t)x\rangle \\ &\qquad +2\langle (Q(t; B)-F(t; B))x, R^{-1}(t)P_1^*(t; B)x\rangle = 0 \quad \text{a.e.}, \\ &Q(T) = F(T).\end{aligned} \tag{3.32}$$

In (c) we shall prove that (3.32) has the unique solution zero and so $Q(t) = F(t)$, with

$$\begin{aligned}\langle Q(t)x, x\rangle &= \int_t^T \langle \mathcal{U}_1^*(s, t)Q(s; B)R^{-1}(s)Q^*(s; B)\mathcal{U}_1(s, t)x, x\rangle\, ds \\ &\geqq 0.\end{aligned}$$

Using the same arguments for P_2 perturbations, we find $\langle Q(t)x, x\rangle \leqq 0$ and so $Q(t)x = 0$ for $x \in \mathcal{D}_A$. Since $\mathcal{D}_A$ is dense in H, $Q(t) = 0$ and so (3.30) has a unique solution in the specified class.

(c) We consider the equation

$$\begin{aligned}\frac{d}{dt}\langle P(t)x, y\rangle &= -\langle P(t)x, A(t)y\rangle - \langle A(t)x, P(t)y\rangle \\ &\quad +\langle P^*(t; B)x, R^{-1}(t)P_1(t; B)^*y\rangle \\ &\quad +\langle P_1^*(t; B)x, R^{-1}(t)P^*(t; B)y\rangle \quad \text{a.e.}, \\ P(T) &= 0,\end{aligned} \tag{3.33}$$

where $P_1(t)$, $P_1^*(t; B)$ is any solution of (3.28) in the specified class. We show that (3.33) has the unique solution in the class of operators with properties (3.29) (a), (b) and (c).

Let $S(t) = \mathcal{U}^*(t, s)P(t)\mathcal{U}(t, s)$. Then since $\mathcal{U}(t, s)$ is a strong evolution operator, we may apply Lemma 3.9 to show that $\langle S(t)x, y\rangle$ is absolutely continuous for $x, y \in \mathcal{D}_A$ with

$$\frac{d}{dt}\langle S(t)x, y\rangle = \langle P^*(t; B)\mathcal{U}(t, s)x, R^{-1}(t)P_1^*(t; B)\mathcal{U}(t, s)y\rangle$$

$$+\langle P_1^*(t; B)\mathcal{U}(t, s)x, R^{-1}(t)P^*(t; B)\mathcal{U}(t, s)y\rangle \quad \text{a.e.}$$

So

$$\langle S(t)x, y\rangle = -\int_t^T \langle P^*(\alpha; B)\mathcal{U}(\alpha, s)x, R^{-1}(\alpha)P_1^*(\alpha; B)\mathcal{U}(\alpha, s)y\rangle$$

$$+\langle P_1^*(\alpha; B)\mathcal{U}(\alpha, s)x, R^{-1}(\alpha)P^*(\alpha; B)\mathcal{U}(\alpha, s)y\rangle;$$

letting $s \to t$, we obtain

$$\langle P(t)x, y\rangle = -\int_t^T [\langle P^*(\alpha; B)\mathcal{U}(\alpha, t)x, R^{-1}(\alpha)P_1^*(\alpha; B)\mathcal{U}(\alpha, t)y\rangle$$

$$+\langle P_1^*(\alpha; B)\mathcal{U}(\alpha, t)x, R^{-1}(\alpha)P^*(\alpha; B)\mathcal{U}(\alpha, t)y\rangle]\, d\alpha. \tag{3.34}$$

Since $P^*(t; B)$ exists by property (3.29) (c), we see that

$$\langle P^*(t; B)x, u\rangle = -\int_t^T [\langle P^*(\alpha; B)\mathcal{U}(\alpha, t)x, R^{-1}(\alpha)P_1^*(\alpha; B)\mathcal{U}(\alpha, t; B)u\rangle$$

$$+\langle P_1^*(\alpha; B)\mathcal{U}(\alpha, t)x, R^{-1}(\alpha)P^*(\alpha; B)\mathcal{U}(\alpha, t; B)u\rangle]\, d\alpha.$$

Let $u = P^*(t; B)x$; then

$$\|P^*(t; B)x\|^2 \leqq \text{const.} \int_t^T \|P^*(\alpha; B)\|\, \bar{f}(t-\alpha)g(\alpha-t)\|P^*(t; B)x\|\, \|x\|\, d\alpha.$$

Hence

$$\|P^*(t; B)x\| \leqq \text{const.} \int_t^T \bar{f}(t-\alpha)\bar{g}(\alpha-t)\|P^*(\alpha; B)\|\, \|x\|\, d\alpha.$$

So

$$\|P^*(t; B)\| \leqq \text{const.} \int_t^T \bar{f}(t-\alpha)\bar{g}(\alpha-t)\|P^*(\alpha; B)\|\, d\alpha.$$

From the generalized Gronwall's inequality (Appendix A, (A.3)) and from (3.34), $P(t) = 0$.

4. Extensions to more general cost functionals. The motivation for extending the results of § 3 to a more general cost functional is that one may wish to allow for penalties of the state values on a particular lower dimensional manifold Γ of

the region Ω by including a term

$$|Cz(T)|^2+\int_{t_0}^{T}|Cz(s)|^2\,ds,$$

where $C: L_2(\Omega)\to R^K$ represents the evaluation map on Γ. (The dual for this in the filtering problem is boundary noise in the state equation, an important practical problem.) To allow for this, we suppose there exists a map $\mathcal{U}(C; t, s): H\to K$ defined on $\Delta(T)$, where K is another Hilbert space and

$$\|\mathcal{U}(C; r, s)h\|_K \leqq f(t-s)\|h\|_H \quad \text{for } t>s, \tag{4.1}$$

where

$$\mathcal{U}(C; t, s)\mathcal{U}(r, s)=\mathcal{U}(C; t, s) \quad \text{for } 0\leqq s<r\leqq t\leqq T. \tag{4.2}$$

This means we may assume

$$f(t+s)\leqq Mf(t), \qquad t>0, \quad s\geqq 0 \quad \text{(cf. (2.6))}. \tag{4.3}$$

Now we define a map $\mathcal{U}(C; t, s; B): U\to K$ by

$$\mathcal{U}(C; t, s; B)=\mathcal{U}(C; t, r)\mathcal{U}(r, s; B), \qquad 0\leqq s<r<t\leqq T, \tag{4.4}$$

which implies that

$$\begin{aligned}\|\mathcal{U}(C; t, s; B)u\|_K &\leqq f(t-r)g(r-s)\|u\|_U \quad (0\leqq s<r<t\leqq T)\\ &= f(\mu(t-s))g((1-\mu)(t-s))\|u\|_U,\end{aligned} \tag{4.5}$$

where $0<\mu<1$, μ is arbitrary.

We now consider the more general cost function

$$\begin{aligned}\mathcal{C}(u; t_0, z_0) &= \langle z(T), Gz(T)\rangle_K+\int_{t_0}^{T}\langle Cz(s), Cz(s)\rangle_K\,ds\\ &\quad+\int_{t_0}^{T}\langle u(s), R(s)u(s)\rangle_U\,ds,\end{aligned} \tag{4.6}$$

where

$$(Cz)(t)=\mathcal{U}(C; t, t_0)z_0+\int_{t_0}^{t}\mathcal{U}(C; t, s; B)u(s)\,ds$$

and $\mathcal{U}(C; t, s)$, $\mathcal{U}(C; t, s; B)$ satisfies (4.1)–(4.5). Now it is possible to form a recursive scheme analogous to (3.3)–(3.9), but replacing $\mathcal{U}_k^*(s, t)W(s)\mathcal{U}_k(s, t)$ by $\mathcal{U}_k^*(C; s, t)\mathcal{U}_k(C; s, t)$ and $\mathcal{U}_k^*(s, t)W(s)\mathcal{U}_k(s, t; B)$ by $\mathcal{U}_k^*(C; s, t)\mathcal{U}_k(C; s, t; B)$ and so on. If we make the additional assumption

$$f(\cdot)g(\cdot)\in L_2(T), \tag{4.7}$$

then it is readily verified that by modifying the f_k estimates, the whole argument of § 3 goes through as before and so we state the final result.

THEOREM 4.1. *Consider the controlled system*

$$z(t)=\mathcal{U}(t, t_0)z_0+\int_{t_0}^{t}\mathcal{U}(t, s; B)u(s)\,ds \tag{4.8}$$

under the cost (4.6) *where H, K are real Hilbert spaces,* $z_0 \in H$, $\mathcal{U}(t, s)$ *is a mild evolution operator on H,* $\mathcal{U}(t, s; B)$ *satisfies* (2.4)–(2.6), $\mathcal{U}(C; t, s)$, $\mathcal{U}(C; t, s; B)$ *and* (4.7), *and R,* $R^{-1} \in \mathcal{B}_\infty(T: \mathcal{L}(U))$ *satisfy* (4.1)–(4.5). *Then there exists a unique optimal control* $u^* \in L_2(T; U)$ *given by*

$$u^*(t) = -R^{-1}(t)Q_\infty^C(t; B)z(t), \tag{4.9}$$

$$\begin{aligned} Q_\infty^C(t; B)u &= \mathcal{U}_\infty^*(T, t)G\mathcal{U}_\infty(T, t; B)u \\ &\quad + \int_t^T \mathcal{U}_\infty^*(C; s, t)\mathcal{U}(C; s, t; B)u \, ds, \end{aligned} \tag{4.10}$$

$$\begin{aligned} \mathcal{U}_\infty(C; t, s)x &= \mathcal{U}(C; t, s)x \\ &\quad - \int_s^t \mathcal{U}(C; t, \alpha; B)R^{-1}(\alpha)Q_\infty^C(\alpha; B)\mathcal{U}_\infty(\alpha, s)x \, d\alpha. \end{aligned} \tag{4.11}$$

The minimum cost is $\langle Q^C(T)z_0, z_0\rangle$, *where*

$$Q^C(t)x = \mathcal{U}_\infty^*(T, t)G\mathcal{U}(T, t)x + \int_t^T \mathcal{U}_\infty^*(C; s, t)\mathcal{U}(C; s, t)x \, ds. \tag{4.12}$$

There are also analogues to (3.16)–(3.17), but there is no analogue to the differential form of $Q(t)$, as in Theorem 3.3.

5. The filtering and control problems and their duality. In this section we show how the filtering problem with unbounded observations, and the control problem with unbounded control action can be formulated, and explore the duality introduced by the infinite dimensional Riccati equation.

For the general formulation of the filtering problem for infinite dimensional linear systems with Gaussian white noise disturbance we follow [3], and consider

$$z(t) = V(t, 0)z_0 + \int_0^t V(t, s)D(s) \, dw(s), \tag{5.1}$$

$$y(t) = \int_0^t C(s)z(s) \, ds + \int_0^t F(s) \, dv(s), \tag{5.2}$$

where $V(t, s)$ is a mild evolution operator on a real separable Hilbert space H, $(\Omega, \mathcal{P}, \mu)$ is a complete probability space, z_0 is a Gaussian random variable $\in L_2(\Omega; H)$ with zero expectation and covariance P_0, $w(t)$ is an H-valued Wiener process with covariance matrix $R_1 \in \mathcal{L}(H)$, $D \in \mathcal{B}_\infty(T; \mathcal{L}(H))$, F, $F^{-1} \in L_\infty(T; \mathcal{L}(R^k))$, v is an R^k-valued Wiener process with covariance R_2, and R_2^{-1} exists, w, v and z_0 are mutually independent. The filtering problem is to find the best global estimate $\hat{z}(t)$ of the signal process $z(t)$ based on the observation process $y(s)$; $0 \leqq s \leqq t$. In [3] it is shown that

$$\hat{z}(t) = \int_0^t \mathcal{Y}(t, s)P(s)C^*(s)(F(s)R_2F^*(s))^{-1} \, dy(s), \tag{5.3}$$

where $\mathcal{Y}(t, s)$ is the perturbation of $\mathcal{U}(t, s)$ by $-P(t)C^*(t)(F(t)R_2F^*(t))^{-1}C(t)$ satisfying

$$\begin{aligned}\mathcal{Y}(t, s)x &= V(t, s)x - \int_s^t \mathcal{Y}(t, r)(C(r)P(r))^*(F(r)R_2F^*(r))^{-1}C(r)V(r, s)x\,dr \\ &= V(t, s)x - \int_s^t V(t, r)(C(r)P(r))^*(F(r)R_2F^*(r))^{-1}C(r)\mathcal{Y}(r, s)x\,dr\end{aligned} \tag{5.4}$$

(see [2]). $P(t)$ is the unique solution of the integral Riccati equation

$$P(t)x = V(t, 0)P_0\mathcal{Y}^*(t, 0)x + \int_0^t V(t, s)D(s)R_1D^*(s)\mathcal{Y}^*(t, s)x\,ds. \tag{5.5}$$

Existence and uniqueness for (5.5) are obtained by transforming it to the dual Riccati equation (5.6) and appealing to known results for the quadratic cost control problem of [2].

$$Q(t)x = \mathcal{U}_\infty^*(T, t)P_0V(T, t)x + \int_t^T \mathcal{U}_\infty^*(s, t)W_1(s)\mathcal{U}(s, t)x\,ds, \tag{5.6}$$

where

$$P(t) = Q(T-t), \quad \mathcal{U}_\infty(t, s) = \mathcal{Y}^*(T-s, T-t), \quad W_1(s) = D(T-s)R_1D^*(T-s).$$

So $\mathcal{U}_\infty(t, s)$ is the perturbation of the dual mild evolution operator $\mathcal{U}(t, s) = V^*(T-s, T-t)$ by $-C^*(T-t)(F(T-t)R_2F^*(T-t))^{-1}C(T-t)Q(t) = G(t)$ satisfying

$$\begin{aligned}\mathcal{U}_\infty(t, s)x &= \mathcal{U}(t, s)x - \int_s^t \mathcal{U}(t, r)G(r)\mathcal{U}_\infty(r, s)x\,ds \\ &= \mathcal{U}(t, s)x - \int_s^t \mathcal{U}_\infty(t, r)G(r)\mathcal{U}(r, s)x\,ds.\end{aligned} \tag{5.7}$$

If the observation is of the value of the state at certain points or manifolds, C will be linear, but unbounded with dense domain. So assuming that $C(t)V(t, s)$ exists as a bounded extension, (5.3), (5.4) and (5.5) will still make sense. To establish existence and uniqueness of the integral Riccati equation (5.6) with unbounded C, we are led to considering the following generalization of the dual Riccati system (5.6), (5.7):

$$Q(t)x = \mathcal{U}_\infty^*(T, t)P_0\mathcal{U}(T, t)x + \int_t^T \mathcal{U}_\infty^*(s, t)W_1(s)\mathcal{U}(s, t)x\,ds, \tag{5.8}$$

$$Q(t; C)x = \mathcal{U}_\infty^*(T, t)P_0\mathcal{U}(T, t; C)x + \int_t^T \mathcal{U}_\infty^*(s, t)W_1(s)\mathcal{U}(s, t; C)x\,ds, \tag{5.9}$$

$$\begin{aligned}\mathcal{U}_\infty(t, s)x = \mathcal{U}(t, s)x - \int_s^t \mathcal{U}(t, \rho; C)(F(T-\rho)R_2F^*(T-\rho))^{-1} \\ \cdot Q^*(\rho; C)\mathcal{U}_\infty(\rho, s)x\,d\rho,\end{aligned} \tag{5.10}$$

where we have introduced the duality

$$\mathcal{U}(t, s) = V^*(T-s, T-t), \tag{5.11}$$

$$\mathcal{U}(t, s; C) = [C(T-s)V(T-s, T-t)]^* \tag{5.12}$$

(assuming this can be well-defined).

If $\mathcal{U}(t, s)$, $\mathcal{U}(t, s; C)$ satisfy the assumptions (2.1), (2.2)', (2.4), (2.5), (2.6), then from the theory of § 3, we know that (5.9)–(5.10) has a unique solution.

Consequently for both the filtering and control problems we need to show how it is possible to impose conditions on the operators $V(t, s)$, $C(t)$, or $\mathcal{U}(t, s)$, $B(t)$ which enable us to construct an operator $V(t, s; C)$ or $\mathcal{U}(t, s; B)$ which satisfies the assumptions (2.4), (2.5), (2.6). There will be two types of conditions corresponding to the Examples 1.1 and 1.2 introduced in § 1.

Assumption (5.13). (a) For each $t \in [0, T]$, $C(t)$ is a closed densely defined linear operator $C(t): H \to K$, where H, K are Hilbert spaces. (b) For almost all t, s, $\exists g \in L_2(T)$ and a set $\chi_{t,s}$ dense in H such that

$$\|C(t)V(t, s)x\|_K \leqq g(t-s)\|x\|_H \quad \forall x \in \chi_{t,s}.$$

We note (see [11]) that $C^*(t)$ is a closed, densely defined linear operator on K^* and since $\bar{\chi}_{t,s} = H$, there is an extension $\overline{C(t)V(t, s)} \in \mathcal{L}(H, K)$. This extension will depend on the particular version taken for g from the equivalence classes of $g \in L_2(T)$. However all such extensions will give the same value for

$$y(t) = \int_0^t \overline{C(t)V(t, s)}h(s)\, ds,$$

and we have

$$\|y(t)\|_K \leqq \int_0^t g(t-s)\|h(s)\|\, ds$$

for $y \in L_2(T; H)$. So we will not distinguish between the extensions, and take

$$\|\overline{C(t)V(t, s)}\|_{\mathcal{L}(H,K)} \leqq g(t-s).$$

If we identify H with its dual, since $\mathcal{U}(t, s) \in \mathcal{L}(H)$, we have (see [11])

$$\overline{C(t)V(t, s)}^* = [C(t)V(t, s)]^*$$

and

$$V^*(t, s)C^*(t) \subset [C(t)V(t, s)]^*.$$

So

$$\overline{C(t)V(t, s)}^* = \overline{V^*(t, s)C^*(t)}$$

with

$$\|V^*(t, s)C^*(t)\| \leqq g(t-s)$$

and

$$\overline{V^*(t, s)C^*(t)} = V^*(r, s)\overline{V^*(t, r)C^*(t)}.$$

If we set

$$\mathcal{U}(t, s) = V^*(T-s, T-t) \tag{5.14}$$

and

$$\mathcal{U}(t, s; C) = \overline{V^*(T-s, T-t)C^*(T-s)} \tag{5.15}$$

with

$$K^* = U$$

we see that $\mathcal{U}(t, s; C)$ satisfies the assumptions (2.4), (2.5), (2.6). Conversely for the control problem we assume $B(t): U \to H$ is a closed, densely defined linear operator, such that

$$\|\mathcal{U}(t, s)B(s)u\|_H \leqq g(t-s)\|u\|_U. \tag{5.16}$$

Then

$$\mathcal{U}(t, s; B) = \overline{\mathcal{U}(t, s)B(s)}$$

and

$$\mathcal{U}^*(t, s; B) = B^*(s)\mathcal{U}^*(t, s).$$

The above analysis is not relevant for the case of control or observation from boundaries or lower dimension manifolds, since although in general $C(t)$ will be densely defined it will not be closed, and $C^*(t)$, $B(t)$ will not have dense domain (Example 1.2). For these cases we need to consider a different approach and assume

Assumption (5.17). Let W be a Banach space such that $\bar{W} = H$, and assume

(a) $H \supset \mathcal{D}(C(t)) \supset W, \quad t \in T,$

(b) $C \in L_\infty(T; \mathcal{L}(W, K)),$

(c) $V(t, s) \in \mathcal{L}(H, W), \qquad t > s,$

(d) $\|V(t, s)x\|_W \leqq g(t-s)\|x\|_H \quad$ for all $x \in H.$

The above assumptions imply

$$\|C(t)V(t, s)x\|_K \leqq g(t-s)\|C\|_{L_\infty(T;\mathcal{L}(W,K))}\|x\|_H.$$

Also for $f \in L_2(T; H)$, $V(t, s)f(s)$ is in W and is Bochner integrable with respect to W. Moreover since $C \in L_\infty(T; \mathcal{L}(W, K))$,

$$C(t)\int_0^t V(t, s)f(s)\, ds = \int_0^t C(t)V(t, s)f(s)\, ds.$$

Now for $t > s$,

$$V^*(t, s) \in \mathcal{L}(W^*, H)$$

and

$$C^*(t) \in \mathcal{L}(U, W^*).$$

Hence

$$V^*(t, s)C^*(t) = (C(t)V(t, s))^*$$

and

$$\|V^*(t, s)C^*(t)\| \leqq g(t-s).$$

Thus if we set

$$\mathcal{U}(t, s) = V^*(T-s, T-t), \tag{5.18}$$

$$\mathcal{U}(t, s; C) = V^*(T-s, T-t)C^*(T-s), \qquad K^* = U, \tag{5.19}$$

we see that $\mathcal{U}(t, s; C)$ satisfies the assumptions (2.4), (2.5), (2.6), and the infinite dimensional Riccati equation associated with the filtering problem is well defined. Alternatively the dual conditions for the control problem take the form

$$\begin{aligned} &\text{(a)} \quad \mathcal{U}(t, s) \in \mathcal{L}(W^*, H), \qquad t > s, \\ &\text{(b)} \quad \|\mathcal{U}(t, s)x\|_H \leqq g(t-s)\|x\|_{W^*}, \quad x \in W^*, \\ &\text{(c)} \quad B \in L_\infty(T; \mathcal{L}(U, W^*)), \\ &\text{(d)} \quad \mathcal{U}(t, s; B) = \mathcal{U}(t, s)B(s). \end{aligned} \tag{5.20}$$

Then $\mathcal{U}(t, s; B)$ satisfies the assumptions (2.4), (2.5), (2.6) and the control problem is well defined. Assumption (a) implies that the evolution operator is smoothing as is the case for analytic semigroups, for example. If this is not the case, it is necessary to identify W with H and then to choose spaces U and H so that (b) holds (see § 6). In some control problems finding a suitable abstract formulation satisfying (5.20) is rather indirect, namely one finds first a suitable "dual" operator $C(t)$ and then $B(t)$ is defined as its transpose with respect to certain spaces (see § 6).

6. Applications. In the preceding theory we have assumed that the dynamics are given as input-output relations in terms of an evolution operator $\mathcal{U}(t, s)$:

$$z(t) = \mathcal{U}(t, t_0)z_0 + \int_{t_0}^{t} \mathcal{U}(t, s; B)u(s)\, ds \tag{6.1}$$

with $u \in L_2(T; U)$, $z_0 \in H$. More usually in applications the system will be described in terms of abstract evolution equations and so first we must examine the connection between these two formulations.

If the control is distributed, but unbounded, the appropriate abstract evolution equation is

$$\begin{aligned} &\dot{z}(t) = A(t)z(t) + B(t)u(t) \\ &z(t_0) = z_0, \qquad u(t) \in \mathcal{D}(B(t)). \end{aligned} \tag{6.2}$$

Then if $\mathcal{U}(t, s)$, $B(t)$ satisfy assumption (5.16) we define (6.1) to be a mild solution of (6.2) and there is no difficulty in interpreting the operator $B(t)$. However if the control is constrained to a submanifold, the boundary or even points then it is not clear how we should choose $B(t)$, and what is the exact

relationship of (6.1) to the evolution equation. In order to show how this can be done let us consider the following.

$$\begin{aligned} &\dot{z}(t) = A(t)z(t), \qquad z(t_0) = z_0 \quad \text{in } \Omega, \\ &[\gamma(t)z(t)]_\Gamma = u(t) \quad \text{on } \Gamma, \end{aligned} \tag{6.3}$$

where Ω is an open bounded region in R^n with boundary $\partial\Omega$ and Γ is a submanifold of Ω or of the boundary $\partial\Omega$. The symbol $[\,\cdot\,]_\Gamma$ denotes the change in $\gamma(t)z(t)$ across Γ. Assume that U is a Hilbert space based on Γ, and there is a Green's formula of the form

$$\begin{aligned} &\langle A(t)z_1, z_2\rangle_H - \langle z_1, A^*(t)z_2\rangle_H \\ &\quad = \langle [\gamma(t)z_1]_\Gamma, \delta(t)z_2\rangle_{U,U^*} - \langle [\delta_*(g)z_1]_\Gamma, \gamma_*(t)z_2\rangle_{U,U^*}. \end{aligned} \tag{6.4}$$

For $z_2 \in \mathscr{D}(A^*(t))$ in $\bar{\Omega}$, $z_1 \in \mathscr{D}(A(t))$ on $\bar{\Omega}\backslash\Gamma$ and almost all $t \in T$, where $\langle\,\cdot\,,\,\cdot\,\rangle_{U,U^*}$ denotes the duality pairing between U and U^*.

This Green's formula (6.4) is somewhat unusual although versions for $\Gamma \subset \partial\Omega$ have been established by Lions (see [7] and [8]). Further assume there are Hilbert spaces E_1, E_2, F_1, F_2 with continuous and dense injections into H, such that

$$\begin{aligned} &\gamma(t) \in \mathscr{L}(E_1, U), \quad \gamma_*(t) \in \mathscr{L}(E_2, U^*), \quad \delta(t) \in \mathscr{L}(F_1, U^*), \\ &\delta_*(t) \in \mathscr{L}(F_2, U) \quad \text{for almost all } t \in T. \end{aligned} \tag{6.5}$$

If we choose $W = F_1$ and let $B(t) = \delta(t)^T$, the transpose of δ, then $B(t) \in \mathscr{L}(U, F_1^*)$, and is given by

$$\langle B(t)z_1, z_2\rangle_{W^*W} = \langle z_1, \delta(t)z_2\rangle_{U,U^*}. \tag{6.6}$$

Now assume $A(t)$ generates an evolution operator $\mathscr{U}(t, s)$ on H with $\mathscr{U}(t, s) \in \mathscr{L}(F_1^*, H)$, $t > s$,

$$\|\mathscr{U}(t, s)z\| \leqq g(t-s)\|z\|_{F_1^*} \quad \forall z \in H, \quad g \in L_2(T).$$

The conditions (5.20) are satisfied so that

$$z(t) = \mathscr{U}(t, t_0)z_0 + \int_{t_0}^{t} \mathscr{U}(t, s)B(s)u(s)\,ds \tag{6.7}$$

is well defined for $u \in L_2(T; U)$. If $V(t, s) = \mathscr{U}^*(T-s, T-t)$ is given with generator $A^*(T-t)$, then z given by (6.7) is a weak solution of

$$\dot{z}(t) = A(t)z(t) + B(t)u(t), \qquad z(t_0) = z_0, \tag{6.8}$$

and

$$\begin{aligned} &\dot{z}(t) = A(t)z(t), \qquad z(t_0) = z_0, \\ &[\gamma(t)z(t)]_\Gamma = u, \qquad [\delta_*(t)z_1(t)]_\Gamma = 0. \end{aligned} \tag{6.9}$$

Here by a weak solution we mean there exists

$$\xi \in C[T; W], \quad \dot{\xi} \in H, \quad \xi(s) \in \mathscr{D}(A^*(s)),$$

$\dot{\xi}$, $A^*(s)\xi(s)$ integrable, and $\xi(T)=0$, such that

$$\int_{t_0}^{T}\langle z(s),\dot{\xi}(s)+A^*(s)\xi(s)\rangle_H\,ds+\int_{t_0}^{T}\langle B(s)u(s),\xi(s)\rangle_{W^*W}\,ds + \langle z_0,\xi(t_0)\rangle_H=0. \tag{6.10}$$

To see this we compute for z satisfying (6.7),

$$\begin{aligned}
&\int_{t_0}^{T}\langle z(s),\dot{\xi}(s)+A^*(s)\xi(s)\rangle_H\,ds\\
&\quad=\int_{t_0}^{T}\langle \mathcal{U}(s,t_0)z_0,\dot{\xi}(s)+A^*(s)\xi(s)\rangle_H\,ds\\
&\qquad+\int_{t_0}^{T}\int_{t_0}^{s}\langle \mathcal{U}(s,\alpha)B(\alpha)u(\alpha),\dot{\xi}(s)+A^*(s)\xi(s)\rangle_H\,ds\,d\alpha\\
&\quad=\int_{t_0}^{T}\langle z_0,\mathcal{U}^*(s,t_0)\dot{\xi}(s)+\mathcal{U}^*(s,t_0)A^*(s)\xi(s)\rangle_H\,ds\\
&\qquad+\int_{t_0}^{T}\int_{\alpha}^{T}\langle B(\alpha)u(\alpha),\mathcal{U}^*(s,\alpha)\dot{\xi}(s)+\mathcal{U}^*(s,\alpha)A^*(s)\xi(s)\rangle_{W^*W}\,d\alpha\,ds\\
&\qquad\qquad\text{(interchanging the order of integration and using (5.20))}\\
&\quad=\int_{t_0}^{T}\frac{d}{ds}\langle z_0,\mathcal{U}^*(s,t_0)\xi(s)\rangle\,ds\\
&\qquad+\int_{t_0}^{T}\int_{\alpha}^{T}\frac{d}{ds}\langle B(\alpha)u(\alpha),\mathcal{U}^*(s,\alpha)\xi(\alpha)\rangle_{W^*W}\,ds\,d\alpha\\
&\qquad\qquad\text{(since } V(t,s)\text{ is quasi)}\\
&\quad=-\langle z_0,\xi(t_0)\rangle_H-\int_{t_0}^{T}\langle B(\alpha)u(\alpha),\xi(\alpha)\rangle_{W^*W}\,d\alpha.
\end{aligned}$$

Clearly the weak solution of (6.8) satisfies (6.10), and since by Green's formula (6.4)

$$\begin{aligned}
\langle[\gamma(t)z(t)]_\Gamma,\delta(t)z\rangle_{UU^*}&=\langle u,\delta(t)z\rangle_{UU^*}\\
&=\langle B(t)u,z\rangle_{WW^*},
\end{aligned}$$

then by (6.6) we see that the weak solution of (6.9) also satisfies (6.10).

We remark that a sufficient condition for $V(t,s)$ to be quasi is that $\mathcal{U}(t,s)$ is a strong evolution operator and $A(t)\mathcal{U}(t,s)z$ is Bochner integrable on $(s,T]$ for all $z\in\mathcal{D}(A(s))$ (see [2]).

We now apply the theory to some examples introduced in §1.

Example 6.1 (Example 1.2). $A=A^*=\partial^2/\partial x^2$ and we take $H=L_2(0,1)$, $U=U^*=R^1$,

$$\mathcal{D}(A)=\mathcal{D}(A^*)=\{z\in H:\ z_{xx}\in H,\ z_x(1)=0=z_x(0)\},\qquad \gamma z=z_x(0).$$

The Green's formula for A is

$$\int_0^1 (z_{1xx}z_2 - z_1 z_{2xx})\,dx = -[z_{1x}(0)]_0 z_2(0) + [z_1(0)]_0 z_{2x}(0) \tag{6.11}$$

for $z_1, z_2 \in \mathscr{D}(A)$ on $(0, 1]$. So

$$[\gamma z] = [z_x(0)], \qquad \gamma_* z = z_x(0),$$
$$\delta z = -z(0), \qquad \delta_* z = -z(0),$$

and

$$E_1 = E_2 = H^{3/2}(0, 1); \qquad F_1 = F_2 = H^{1/2}(0, 1) = W.$$

Hence $-B = \delta$, the Dirac delta function, and $B \in \mathscr{L}(R, W^*)$.

To verify that $\mathscr{T}_t \in \mathscr{L}(W^*, H)$ is straightforward and yields

$$\|\mathscr{T}_t x\|_H \leqq \frac{M}{t^{1/4}} \|x\|_{W^*}.$$

So assumption (5.20) is satisfied and the boundary control problem is well-defined. Since the evolution operator is a semigroup, we know that the differential Riccati equation has a unique solution.

Example 6.2 (Example 1.1).

$$A = A^* = \frac{\partial^2}{\partial x^2} \quad \text{and} \quad H = L_2(0, 1), \quad U = U^* = R.$$

$$\mathscr{D}(A) = \mathscr{D}(A^*) = \{z \in H: z_{xx} \in H, z(1) = 0 = z(0)\}, \qquad \gamma z = z(0).$$

We have the Green's formula similar to (6.11) and so

$$\int_0^1 (z_{1xx}z_2 - z_1 z_{2xx})\,dx = [z_1(0)]z_{2x}(0) - [z_{1x}(0)]z_2(0) \quad \text{for } z_1, z_2 \in \mathscr{D}(A) \text{ on } (0, 1].$$

$$\gamma z = z(0), \quad \gamma_* z = z(0), \quad \delta_* z = z_x(0), \quad \gamma_* z = z_x(0)$$

with $E_1 = E_2 = H^{1/2}(0, 1)$, $F_1 = F_2 = H^{3/2}(0, 1)$. As before we let $W = F_2 = H^{3/2}(0, 1)$ and $-B = \delta'$, the negative derivative of the Dirac delta function. However, this time we have

$$\|\mathscr{T}_t x\|_H \leqq \frac{M}{t^{3/4}} \|x\|_{W^*}$$

and so assumption (5.20) does not hold with this choice of spaces. However, if instead we choose $H = H^{(-1/2)-\varepsilon}(0, 1)$, $U = R$ and $F = H^{(1/2)-2\varepsilon}(0, 1)$, we obtain the estimate

$$\|\mathscr{T}_t x\|_H \leqq \frac{M}{t^{1/2-(\varepsilon/2)}} \|x\|_{W^*}$$

and assumption (5.20) will be satisfied.

So we see that the control and filtering problems are well posed for this control and state space pairing. In general if we calculate $g \in L_p(T)$ with $1 \leqq p < 2$

for a particular pairing (U, H), it may be possible to enlarge the state space or smooth the control space U to obtain a new $g \in L_p(T)$ with $p \geqq 2$. Alternatively if it is decided to work with the spaces (U, H) which yield $1 \leqq p < 2$, we will not have z given by (3.1) in $C(T; H)$ but in $L_{2p/(2-p)}(T; H)$. Thus the control problem is well defined if there is no penalty on the final state, i.e., $G = 0$ in (3.2).

From time-dependent parabolic partial differential equations a similar analysis must be made for each particular problem; however, we can make the following general remarks concerning these systems.

For the Kato and Tanabe class for each t, the operator $A(t)$ generates an analytic semigroup and so the evolution operator $\mathcal{U}(t, s)$ maps H into $\mathcal{D}(A(t))$ for each $t > s$ (see [5], [6]). So one of the essential assumptions of (5.17) that $\mathcal{U}(t, s)$ be smoothing is satisfied. For assumption (5.13) we can deduce some information from the following perturbation result in [6]. If $C(t)$ is a closed linear operator whose domain contains that of $A(t)$ and

$$\|C(t)(\lambda I - A(t)^{-1}\| \leqq \frac{M}{|\lambda|^{1-\gamma}}, \qquad M > 0, \quad \gamma > 0$$

for each λ in a closed sector of the resolvent set of $A(t)$, then $A(t) + C(t)$ generates a strong evolution operator and

$$\|C(t)\mathcal{U}(t, s)\| \leqq \frac{C}{(t-s)^{\gamma}}.$$

So for $\gamma < \frac{1}{2}$, assumption (5.13) is satisfied and the first type of unbounded control problem has a unique solution. Moreover, under the extra assumption $\sup_{t \in T} \|A(t)x\| < \infty$ for each $x \in \mathcal{D}(A(t))$, $A(t)$ generates a strong evolution operator and so the differential Riccati equation has a unique solution (see [2]).

In Lions' theory [7], the operator $A(t)$ is associated with a bilinear form $a(t; \varphi, \psi)$ on a Hilbert space V which has continuous injection into H, and is dense in H. Identifying H with its dual, we have

$$V \subset H \subset V^*$$

and under coercivity, boundedness and measurability conditions on the bilinear form, Lions shows that there is a unique solution in $W(0, T)$ of

$$\dot{z}(t) = A(t)z,$$

$$z(0) = z_0 \in H.$$

Here $W(0, T)$ is the Hilbert space

$$W(0, T) = \{z : z \in L_2(T; V), \dot{z} \in L_2(T; V^*)\}.$$

In [2] we have shown that the evolution operator is a quasi-evolution operator, and we note that

$$\int_0^T \|\mathcal{U}(t, s)z_0\|_V^2 \, ds < \infty.$$

This indicates that for assumption (5.20) to be satisfied we should take $V = W$. If the conditions of assumption (5.20) for a particular boundary control problem or

control from manifold problem can be satisfied then we know that we are able to differentiate the Riccati equation. If furthermore the coefficients of $A(t)$ are sufficiently smooth then from Kato and Tanabe [6], we know that $A(t)$ generates a strong evolution operator and so the differential Riccati equation will have a unique solution. However, this is not necessarily the case under the weaker assumptions of Lions [7].

It is useful to note that if we have already established that $\mathcal{U}(t, s)$ and $B(s)$ satisfy (5.16) or (5.20), then we know that any bounded perturbation $\mathcal{U}_D(t, s)$ of $\mathcal{U}(t, s)$ also satisfies (5.16) or (5.20) for $\mathcal{U}_D(t, s)$ is the unique solution of

$$\mathcal{U}_D(t, s)x = \mathcal{U}(t, s)x + \int_s^t \mathcal{U}(t, \rho)D(\rho)\mathcal{U}_D(\rho, s)x \, d\rho.$$

This is particularly useful if $\mathcal{U}(t, s) = \mathcal{T}_{t-s}$, a semigroup with generator A in which case $\mathcal{U}_D(t, s)$ is a quasi evolution operator with generator $A+D(t)$. Moreover the control problem for $z(t) = \mathcal{U}_D(t, t_0)z_0 + \int_{t_0}^t \mathcal{U}_D(t, s; B)u(s)\, ds$ has a unique solution and the differential Riccati equation has a unique solution.

Finally we consider boundary control for a class of hyperbolic systems studied in [9], [10], for which the foregoing approach is not applicable. In [10], Vinter considers both boundary and distributed control obtaining a unique optimal control in feedback form, although he needs to impose extra assumptions for the boundary control case. In [2] we treated the distributed case using an evolution operator approach, so here we just consider the following boundary control problems:

$$\begin{aligned} &\frac{\partial z}{\partial t} = \mathcal{A}(t)z = \sum_{i=1}^{m} A_i \frac{\partial z}{\partial x_i} + Kz, \\ &Mz|_{\partial\Omega} = u, \quad z(0) = z_0, \end{aligned} \tag{6.12}$$

on the spatial domain $\Omega = \{x \in R^m; x_1 > 0\}$, $0 \leqq t \leqq T$, $m > 1$, where A_i, K, M are C^∞ matrix valued functions on $Q = [0, T] \times \Omega$ and $\Sigma = [0, T] \times \Omega$, respectively; each of which may be expressed by a constant function plus a function of compact support. Denote by $C^\infty_{(0)}(Q)$ the restriction of $C_0^\infty(R^{m+1})$ to the closure of $[0, T] \times \Omega$. Then we define a strong solution to (6.12) if given $f \in L_2(Q)$, $u \in L_2(\Sigma)$ and $z_0 \in L_2(\Omega)$, there exists a sequence $\{z_n\}$ with $z_n \in C^\infty_{(0)}(Q)$ such that

$$\begin{aligned} &\|z_n - z\|_{L_2(Q)} \to 0, \\ &\|z_n - z\|_{L_2(\Sigma)} \to 0, \\ &\left\| \frac{\partial z_n}{\partial t} - \mathcal{A}z_n \right\|_{L_2(Q)} \to 0, \\ &\|Mz_n - Mz\|_{L_2(\Sigma)} \to 0, \\ &\|z_n(t_0) - z_0\|_{L_2(\Omega)} \to 0. \end{aligned}$$

Then under technical assumptions on A_i, which ensure that the system is strictly hyperbolic with noncharacteristic boundary, (6.12) has a unique strong

solution $z \in L_2(Q)$. Furthermore, the map $t \to z(t)$ is strongly continuous from $T \to L_2(\Omega)$ and the estimate

$$\|z(t)\|^2_{L_2(\Omega)} + \|z\|^2_{L_2(Q)} + \|z\|^2_{L_2(\Sigma)} \leqq C(\|u\|^2_{L_2(\Sigma)} + \|z_0\|^2_{L_2(\Omega)}) \tag{6.13}$$

holds for all $t \in T$, where C is a constant independent of t.

The technical assumptions on $\mathscr{A}$ and M of interest to us here are

$$A_1(t) = \begin{pmatrix} A^-(t) & 0 \\ 0 & A^+(t) \end{pmatrix},$$

where A^- is a diagonal $r \times r$ matrix with negative entries and A^+ is a diagonal $(n-r) \times (n-r)$ matrix with positive entries.

$M(t) = (I : M^+(t))$, where I is the $r \times r$ identity matrix and M^+ is $r \times (n-r)$ (see [9] for details).

In [2], § 4.4, we showed that (6.5) with $u = 0$ can be rewritten

$$\begin{aligned} \dot{z}(t) &= \mathscr{A}(t) z(t), \\ z(t_0) &= z_0, \end{aligned} \tag{6.14}$$

where $z_0 \in H = L_2(\Omega)$ and $\mathscr{A}(t)$ is the linear operator on H given by

$$(\mathscr{A}(t)h)(x) = \sum_{i=1}^{m} A_i(t, x) \frac{\partial h}{\partial x} + K(t, x) h(x)$$

with domain

$$\mathscr{D}(\mathscr{A}(t)) = \{h \in H : \mathscr{A}(t)h \in H \text{ and } Mh|_{\partial\Omega} = 0\}.$$

Furthermore, $\mathscr{A}(t)$ is the generator of a quasi-evolution operator $\mathscr{U}(t, s)$ on H, and the dual operator $\mathscr{U}^*(T-s, T-t)$ is also quasi.

In order to satisfy assumption (5.20), we need to establish some smoothness properties of $\mathscr{U}(t, s)$, but unfortunately the strongest regularity property from Rauch [9] for $u = 0$ is that if $z_0 \in H_0^s(\Omega)$, then $z \in H^s(Q)$, and an estimate of the form (6.12) holds replacing the L_2 norms by Sobolev norms of order s. So in this case we cannot hope to satisfy (5.20). However, by taking $H = L_2(\Omega)$ and choosing a smooth control space $U = H^{1/2}(\partial\Omega)$, we can reformulate (6.12) as a bounded control problem of the type considered in [2]. As in § 5, we choose $B(t)$ using the Green's formula for $\mathscr{A}(t)$, namely,

$$\langle \mathscr{A}(t)u, v\rangle_H - \langle u, \mathscr{A}^*(t)v\rangle_H = \langle M(t)u, D(t)v\rangle_{L_2(\partial\Omega)} - \langle E(t)u, M^*(t)v\rangle_{L_2(\partial\Omega)}, \tag{6.15}$$

where $M^*(t) = (-(A^+)^{-1}(M^+)^T A^- : I)$; $D(t) = (-A^- - D_1(A^+)^{-1}M^+A^- : D_1)$, $E(t) = [D_1^T : D_1^T M^+ + A^+]$ and D_1 is an arbitrary $r \times (n-r)$ matrix. $\mathscr{A}^*(t)$ is the formal adjoint of $\mathscr{A}(t)$ (see [9]). But the weak solution of (6.12) must satisfy

$$\int_{t_0}^{T} \langle z(t), \mathscr{A}^*(t)\xi(t) + \dot{\xi}(t)\rangle_H \, dt + \int_{t_0}^{T} \langle u(t), D(t)\xi(t)\rangle_{L_2(\partial\Omega)} \, dt + \langle z_0, \xi(t_0)\rangle_H = 0$$

where $\xi \in C^2(T; H)$, $\xi(t) \in \mathscr{D}(\mathscr{A}^*(t))$, $\mathscr{A}^*(t)\xi(t)$ is integrable, $\xi(T) = 0$ and $M^*(t)\xi(t)|_{\partial\Omega} = 0$. Hence we must choose B according to

$$\langle B(y)u, \xi\rangle_H = \langle u, \mathscr{D}(t)\xi\rangle_{UU^*}$$

where $U = H^{1/2}(\partial\Omega)$ and $\mathscr{D}(t)\xi = D(t)\xi|_{\partial\Omega}$, since

$$\mathscr{D}(t) \in \mathscr{L}(H^0(\Omega), H^{-1/2}(\partial\Omega)), \qquad B(t) \in \mathscr{L}(U, H^0(\Omega)),$$
$$z(t) = \mathscr{U}(t, 0)z_0 + \int_0^t \mathscr{U}(t, s)B(s)\, ds \tag{6.16}$$

is the weak solution to (6.12).

We now consider the control problem for (6.9) with the cost function

$$\mathscr{C}(u) = \langle z(T), Gz(T)\rangle_H + \int_0^T [\langle z(s), W(s)z(s)\rangle_H + \langle u(s), R(s)u(s)\rangle_U]\, ds, \tag{6.17}$$

where G, W and R satisfy the usual assumptions as in § 2. With $U = H^{1/2}(\partial\Omega)$, (6.16), (6.17) defines a bounded control problem and from [2], there exists a unique optimal control

$$\bar{u}(t) = -R^{-1}(t)B(t)Q(t)z(t),$$

where

$$Q(t)x = \mathscr{U}^*(T, t)G\mathscr{U}_\infty(T, t)x + \int_t^T \mathscr{U}^*(s, t)W(s)\mathscr{U}_\infty(s, t)x\, ds$$

and $\mathscr{U}_\infty(t, s)$ is the perturbation of $\mathscr{U}(t, s)$ by $-B(t)R^{-1}(t)B^*(t)Q(t)$. Here we have formulated the problem as a bounded control action problem by restricting the control space. Alternatively we could restrict G and W so that $\bar{u}(t)$ is always in $H^{1/2}(\partial\Omega)$. This is so if G and $W(t) \in \mathscr{L}(L_2(\Omega), H_0^1(\Omega))$, for then since $\mathscr{U}(t, s)$ (and $\mathscr{U}^*(t, s)$) maps $H_0^{1/2}(\Omega)$ to $H^{1/2}(\Omega)$, we have $Q(t)$ maps $L_2(\Omega)$ to $H^1(\Omega)$. But $B^*(t) \in \mathscr{L}(H^1(\Omega), H^{1/2}(\partial\Omega))$ and so all feedback controls remain in $H^{1/2}(\partial\Omega)$, provided $R^{-1}(t) \in \mathscr{L}(H^{1/2}(\partial\Omega))$.

Essentially this is what Vinter and Johnson have done in [10] by assuming $W = 0$, R the identity on $L_2(\partial\Omega)$, and $G \in \mathscr{L}(L_2(\Omega), H_0^1(\Omega))$. However, their approach is quite different from ours.

Appendix A.

$$f(t) = h(t) + \int_0^t G(t-s)f(s)\, ds. \tag{A.1}$$

We prove existence and uniqueness of solution for (A.1) for (a) $h \in C(T)$ and (b) $h \in L_2(T)$ under the following assumptions on G:

$$G \text{ is locally integrable, positive and is subadditive:} \quad G(t+s) \leqq MG(t) \quad \text{for } t, s > 0. \tag{A.2}$$

Proof of existence for $h \in C(T)$. Now (A.2) implies that $\int_0^T e^{-\omega t}G(t)\, dt = M_\omega < \infty$ for ω, $T > 0$ and so

$$\int_0^t G(t-s)h(s)\, ds \leqq M_\omega \|h\|_\infty e^{\omega t}. \tag{A.3}$$

Hence the Volterra integral operator G is well defined where

$$(Gh)(t) = \int_0^t G(t-s)h(s)\,ds.$$

The iterates G^n of G are given by

$$(G_n h)(t) = \int_0^t G_n(t-s)h(s)\,ds,$$

where

$$G_n(t) = \int_0^t G(t-r)G_{n-1}(r)\,dr, \qquad n = 2, \cdots,$$

$$G_1(t) = G(t).$$

Now

$$\left\| \sum_{n=1}^{\infty} (G^n h)(t) \right\| \leqq \sum_{n=1}^{\infty} M_\omega^n \|h\|_\infty e^{\omega t}$$

by iterating (A.3).

From (A.2), by choosing ω sufficiently large, we can make $M_\omega < 1$ and so $\sum_{n=1}^{\infty} (G^n h)(\cdot) \in C(T)$. It is then easily verified that $h + \sum_{n=0}^{\infty} G^n h$ is the unique solution to (A.1) in $C(T)$.

COROLLARY A.4.

$$\sum_{n=1}^{\infty} \int_0^t G_n(t-s)\,ds < \text{const. } e^{\omega t}.$$

Proof of existence for $h \in L_2(T)$. Since $(G_n h)(t) = \int_0^t G_n(t-s)h(s)\,ds$ is a convolution, we have $\|G_n h\|_2 \leqq \|G_n\|_1 \|h\|_2$ and so

$$\sum_{n=1}^{\infty} \|G_n h\|_2 \leqq \|h\|_2 \sum_{n=1}^{\infty} \int_0^T G_n(s)\,ds$$

$$< \text{const. } \|h\|_2 \quad \text{(by Corollary A.4).}$$

So $h + \sum_{n=1}^{\infty} G_n h$ is the unique solution of (A.1) in $L_2(T)$.

COROLLARY A.5 (Generalized Gronwall's inequality). *Suppose*

$$g(t) \leqq h(t) + \int_0^t G(t-s)g(s)\,ds,$$

where G satisfies (A.2) *and* $h \in L_2(T)$; *then*

$$g(t) \leqq h(t) + \sum_{n=1}^{\infty} G_n h(t).$$

So if $h = 0$, $g = 0$.

Appendix B. Proof of Lemma 3.1. We prove the results by induction on k. From (2.4)

$$\|\mathcal{U}^*(t, s; B)h\|_U \leqq g(t-s)\|h\|_H.$$

Hence $Q_0^*(t;B)$ is well defined, and

$$\|Q_0^*(t;B)h\|_{\mathcal{U}} \leqq g(T-t)\|G\|M\|h\| + \int_t^T g(s-t)WM\|h\|\,ds,$$

where $W = \text{ess. sup}_{s\in T}\|W(s)\|$, $M = \sup_{\Delta(T)}\|\mathcal{U}(t,s)\|$ by (2.4)–(2.6), so that

$$\|F_1(t)h\|_U \leqq f_1(T-t)\,\|h\|_H,$$

where

$$f_1(t_2-t) = \frac{1}{\beta}\left(g(t_2-t)\|G\|M + MW\int_0^T g(\alpha)\,d\alpha\right) \in L_2(T)$$

since $g \in L_2(T)$ and is subadditive since g is. Also

$$\int_{t_0}^T \|Q_0^*(t;B)h(t)\|_U\,dt \leqq M\|G\|\int_{t_0}^T g(T-t)\|h(t)\|\,dt + MW\int_{t_0}^T\left(\int_t^T g(s-t)\,ds\right)\|h(t)\|\,dt$$

for $h \in L_2(T;H)$. Hence

$$\int_{t_0}^T \|Q_0^*(t;B)h(t)\|\,dt \leqq \left[M\|G\|\left(\int_{t_0}^T g^2(T-t)\,dt\right)^{1/2} + MW\int_{t_0}^T\left(\int_t^T g(s-t)\,ds\right)^2 dt\right]\|h\|_{L_2(T;H)},$$

$$f_{n+1}(T-t) = \frac{1}{\beta}\left(g_n(T-t)\|G\|M_n + WM_n\int_0^T g_n(\alpha)\,d\alpha + M_n\beta\int_0^T g_n(\alpha)f_n^2(T-t-\alpha)\,d\alpha\right)$$

and

$$f_{n+1} \in L_2(T),$$

$$\int_0^T\left(\int_0^T g_n(\alpha)f_n^2(T-t-\alpha)\,d\alpha\right)^2 dt \leqq \int_0^T\left(\int_0^T g_n^2(\alpha)f_n^2(T-t-\alpha)\,d\alpha \cdot \int_0^T f_n^2(T-t-\alpha)\,d\alpha\right)dt$$

$$\leqq \text{const.}\int_0^T g_n^2(\alpha)\int_0^T f_n^2(T-t-\alpha)\,dt\,d\alpha$$

$$< \infty.$$

Then

$$\int_{t_0}^T \|Q_n^*(t;B)h(t)\|\,dt \leqq \beta\int_{t_0}^T f_{n+1}(T-t)\|h(t)\|\,dt \leqq \beta\left(\int_{t_0}^T f_{n+1}^2(T-t)\,dt\right)^{1/2}\|h(\cdot)\|_{L_2(T;H)}$$

and

$$\int_{t_0}^{T} \|Q_n^*(t; B)h(t)\|^2 \, dt \leqq \beta^2 \int_{t_0}^{T} f_{n+1}^2(T-t) \, dt \, \|h\|_{C[T;H]}.$$

Therefore

$$Q_n^*(\,\cdot\,; B) \in \mathscr{L}(C(T; H), L_2(T; U)) \cap \mathscr{L}(L_2(T; H), L_1(T; U)).$$

Similarly we can show the same results for $Q_n(\,\cdot\,; B)$. By Theorem 2.1, $\mathscr{U}_{n+1}(t, s)$ is a well defined mild evolution operator, and by Theorem 2.2, $\mathscr{U}_{n+1}(t, s; B)$ has the properties (2.4)–(2.6) and the estimates (3.12) and (3.13) hold.

Now

$$\begin{aligned}\|Q_n(t; B, B^*)u\| \leqq g_n(T-t)\|G\|g_n(T-t)\|u\|_{C(T;U)} &\\ + W\int_t^T g_n^2(s-t)\|u\|_{C(T;U)} \, ds &\\ + \beta \int_t^T g_n^2(s-t) f_n^2(T-s) \, ds \, \|u\|_{C(T;U)}.&\end{aligned}$$

Thus

$$\begin{aligned}\int_{t_0}^{T} \|Q_n(t; B, B^*)u\| \, dt \leqq \|G\| + \int_{t_0}^{T} g_n^2(T-t) \, dt \, \|u\|_{C(T;U)} &\\ + W\int_{t_0}^{T} g_n^2(s-t) \, ds \, \|u\|_{C(T;U)} &\\ + \beta \int_{t_0}^{T} \left(\int_{t_0}^{s} g_n^2(s-t) \, dt \right) f_n^2(T-s) \, ds \, \|u\|_{C(T;U)}.&\end{aligned}$$

Hence

$$Q_n(\,\cdot\,; B, B^*) \in \mathscr{L}(C(T; U), L_1(T; U)).$$

Appendix C. Proof of Lemma 3.2. Since $F_k \in \mathscr{L}(C(T; H), L_2(T; U))$ and $F_k^* \in \mathscr{L}(L_2(T; U), L_1(T; H))$ we see that the integral in (3.4) is a well defined Bochner integral. Moreover

$$\|Q_k(t)x\| \leqq M_k^2\|G\|\|x\| + \int_t^T M_k^2(W + f_k^2(T-s)\beta) \, ds \, \|x\|$$

so that $Q_k(t)$ is bounded. Clearly $Q_k(t)$ is self adjoint, and the weak continuity of $Q_k(t)x$ is a consequence of the strong continuity of $\mathscr{U}_k(t, s)x$. The weak continuity implies the strong measurability and so the lemma is proved.

Appendix D. Proof of Lemma 3.6. (a) Set

$$\|\mathscr{U}_k(t, s+\varepsilon)\mathscr{U}(s+\varepsilon, s; B)u - \mathscr{U}_k(t, s; B)u\| = h_k(t, s).$$

Then from (3.8),

$$h_k(t,s)=\Big\|\int_{s+\varepsilon}^{t}\mathcal{U}_k(t,\alpha;B)F_k(\alpha)(\mathcal{U}_k(\alpha,s+\varepsilon)\mathcal{U}(s+\varepsilon,s;B)u-\mathcal{U}_k(\alpha,s;B)u)\,d\alpha$$
$$+\int_{s}^{s+\varepsilon}\mathcal{U}(t,\alpha;B)F_k(\alpha)\mathcal{U}_k(\alpha,s;B)t\,d\alpha\Big\|$$

Hence

$$h_k(t,s)\leqq\int_{s+\varepsilon}^{t}g(t-\alpha)f_k(T-\alpha)h_k(\alpha,s)\,d\alpha$$
$$+\int_{s}^{s+\varepsilon}g(t-\alpha)f_k(T-\alpha)g_k(\alpha-s)\,d\alpha\,\|u\|$$
$$\leqq\Big(\int_{s}^{t}g^2(t-\alpha)\,d\alpha\Big)^{1/2}\Big(\int_{s+\varepsilon}^{t}f_k^2(T-\alpha)h_k^2(\alpha,s)\,d\alpha\Big)^{1/2}$$
$$+\Big(\int_{t}^{s+\varepsilon}g^2(t-\alpha)\,d\alpha\Big)^{1/2}\Big(\int_{s}^{s+\varepsilon}f_k^2(T-\alpha)g_k^2(\alpha-s)\,d\alpha\Big)^{1/2}\|u\|.$$

Let $\beta_\varepsilon(t)=\int_{t_0}^{t-\varepsilon}h_k^2(t,s)\,ds$, then

$$\beta_\varepsilon(t)\leqq\text{const.}\int_{t_0+\varepsilon}^{t}f_k^2(T-\alpha)\beta_\varepsilon(\alpha)\,d\alpha$$
$$+\text{const.}\Big(\int_{0}^{\varepsilon}g^2(\alpha)\,d\alpha\Big)\int_{t_0}^{t}f_k^2(T-\alpha)\,d\alpha\,\|u\|$$

So by Gronwall's inequality (A.3) we see that $\beta_\varepsilon(t)\to 0$ as $\varepsilon\to 0$.

The proofs of (b) and (c) are direct consequences of (a), and the bounds on $\mathcal{U}_k(t,s)$, $\mathcal{U}_k(t,s;B)$ and the semigroup properties (2.1) on $\mathcal{U}_k(t,s)$. For example to prove (c) one of the terms is

$$\mathcal{U}^*(t+\varepsilon,t;B)\mathcal{U}_k^*(T,t+\varepsilon)G\mathcal{U}_k(T,t+\varepsilon)\mathcal{U}(t+\varepsilon,t;B)u$$
$$-\mathcal{U}_k^*(T,t;B)G(\mathcal{U}_k(T,t;B))u$$
$$=\mathcal{U}^*(t+\varepsilon,t;B)\mathcal{U}_k^*(T,t+\varepsilon)G(\mathcal{U}_k(T,t+\varepsilon)\mathcal{U}(t+\varepsilon,t;B)-\mathcal{U}_k(T,t;B))u$$
$$+[\mathcal{U}^*(t+\varepsilon,t;B)\mathcal{U}_k^*(T,t+\varepsilon)-\mathcal{U}_k^*(T,t;B)G\mathcal{U}_k(T,t;B)]u.$$

For $u\in C(T;U)$, $\mathcal{U}_k(T;B)u(\cdot)\in L_2(T;H)$ and so the second term tends to zero in $L_1(T;U)$ as $\varepsilon\to 0$ by (a). Now from the convergence in (a) we know that

$$\mathcal{U}^*(t+\varepsilon,t;B)\mathcal{U}_k^*(T,t+\varepsilon)\in\mathcal{L}(L_2(0,T-\varepsilon;H),L_1(0,T-\varepsilon;U))$$

and is uniformly bounded in ε. Hence

$$\int_0^{T-\varepsilon}\|\mathcal{U}^*(t+\varepsilon,t;B)\mathcal{U}_k^*(T,t+\varepsilon)G(\mathcal{U}_k(T,t+\varepsilon)-\mathcal{U}(t+\varepsilon,t;B)$$
$$-\mathcal{U}_k(T,t;B))u\|\,dt$$
$$\leqq\text{const.}\,\|G\|\int\|(\mathcal{U}_k(T,t+\varepsilon)\mathcal{U}(t+\varepsilon,t;B)-\mathcal{U}_k(T,t;B))u\|^2\,dt.$$

But we know from (a) that $\mathcal{U}_k(T, t+\varepsilon)\mathcal{U}(t+\varepsilon, t; B) - \mathcal{U}_k(T, t; B)$ converges as $\varepsilon \to 0$ in $\mathcal{L}(C(T; U), L_2(T; H))$. Similar arguments can be used for the other terms to prove (b) and (c).

Appendix E. Proof of Lemma 3.7. By Theorem 3.1 $\langle Q_k(t+\varepsilon)\mathcal{U}(t+\varepsilon, t; B)w_0$, is decreasing in k for all t, $t+\varepsilon \in T$, $\varepsilon > 0$ and $u_0 \in U$ and

$$\int_{t_0}^{T} \|Q_0^*(t; B)h(t)\|_U^2 \, dt \leqq \frac{3}{2}\Big[M^2\|G\|^2 \int_{t_0}^{T} g^2(T-t)\, dt + W^2M^2 \int_{t_0}^{T} \Big(\int_t^T g(s-t)\, ds\Big)^2 dt\Big]\|h\|_{C(T;H)}^2.$$

Therefore by (2.5),

$$Q_0^*(\,\cdot\,; B) \in \mathcal{L}(C(T; H), L_2(T; U)) \cap \mathcal{L}(L_2(T; H), L_1(T; U)),$$

and similarly we can show

$$Q_0(\,\cdot\,; B) \in \mathcal{L}(C(T; U), L_2(T; H)) \cap \mathcal{L}(L_2(T; U), L_1(T; H)).$$

Also

$$\|Q_0(t; B, B^*)u\| \leqq g(T-t)\|G\|g(T-t)\|u\|_{C(T;U)} + W\int_t^T g^2(s-t)\,\|u\|_{C(T;U)}\, ds$$

and so

$$\int_{t_0}^{T} \|Q_0(t; B, B^*)u\|\, dt \leqq \Big[\|G\| \int_{t_0}^{T} g^2(T-t)\, dt + W\int_{t_0}^{T} g^2(s-t)\, ds\Big]\|u\|_{C(T;U)}$$

and

$$Q_0(\,\cdot\,; B, B^*) \in \mathcal{L}(C(T; U), L_1(T; U)).$$

By Theorem 2.1, $\mathcal{U}_1(t, s)$ is well defined and is a mild evolution operator with

$$\|\mathcal{U}_1(t, s)\| \leqq M_1.$$

By Theorem 2.2, $\mathcal{U}_1(t, s; B)$ is well defined and satisfies an estimate of the form

$$\|\mathcal{U}_1(t, s; B)u\| \leqq g_1(t-s)\|u\|,$$

where $g_1 \in L_2(T)$ and is subadditive.

This establishes the lemma for $k = 1$. Now assume the lemma holds for $k \leqq n-1$; then Theorems 2.1, 2.2 ensure the existence of $\mathcal{U}_n(t, s)$ and $\mathcal{U}_n(t, s; B)$ with the required properties (3.11), (3.12). Also

$$\|Q_n^*(t; B)h\| \leqq \Big(g_n(T-t)\|G\|M_n + \int_t^T g_n(s-t)(W + f_n(T-s)\beta f_n(T-s))M_n\, ds\Big)\|h\|$$

where $\operatorname{ess\,sup}_{s\in T} \|R(s)\| \leqq \beta$. Thus

$$\|F_{n+1}(t)h\| \leqq f_{n+1}(T-t)\|h\|.$$

But $\langle Q_k(t+\varepsilon)\mathcal{U}(t+\varepsilon, t; B)u_0,\ \mathcal{U}(t+\varepsilon, t; B)u_0\rangle$ converges to $\langle Q_k(t; B, B^*)u_0, u_0\rangle$ as $\varepsilon \to 0$ in $L_1(T)$ by Lemma 3.6 (c). So $\langle Q_k(t; B, B^*)u_0, u_0\rangle$ is decreasing in k for almost all $t \in T$, and all $u_0 \in U$ and so is uniformly bounded in k for almost all t. In particular we can find an $L_1(T)$ function such that

$$\langle Q_k(t; B, B^*)u_0, u_0\rangle \leqq q(t)\|u_0\|^2 \quad \text{for almost all } t \in T.$$

Appendix F. Proof of Lemma 3.8. From (3.7) and Lemma 3.7, since G, W are positive operators,

$$\int_t^T \|R^{1/2}(s)F_k(s)\mathcal{U}_k(s, t; B)u_0\|^2\, ds \leqq q(t)\|u_0\|^2.$$

But R is strictly positive, so

$$\int_0^T \|F_k(s)\mathcal{U}_k(s, t; B)u_0\|^2\, ds \leqq \text{const. } q(t)\|u_0\|^2.$$

Hence from (3.9),

$$\begin{aligned}\|\mathcal{U}_k(t, s; B)u_0\| &\leqq g(t-s)\|u_0\| + \int_s^t g(t-\alpha)\|F_k(\alpha)\mathcal{U}_k(\alpha, s; B)u_0\|\, d\alpha \\ &\leqq g(t-s)\|u_0\| + \text{const.} \left(\int_0^{t-s} g^2(\alpha)\, d\alpha\right)^{1/2} \sqrt{q(s)}\, \|u_0\| \\ &= r(t, s)\|u_0\|\end{aligned}$$

and $r \in L_2(T \times T)$ is a uniform bound. This implies that there exists an $r_1 \in L_2(T)$, such that $g_k(t) \leqq r(t)$ for all k and almost all $t \in T$. So from (3.5), $Q_k(t; B)$ has a uniform bound of the form $\|Q_k(t; B)\| \leqq \bar{f}(T-t)$, where $\bar{f} \in L_2(T)$. Similarly since $\langle z_0, Q_k(t)z_0\rangle$ is decreasing in k for each $t \in T$, we have

$$\int_t^T \|F_k(s)\mathcal{U}_k(s, t)z_0\|^2\, ds \leqq \text{const. } \|z_0\|^2.$$

Then from (3.8)

$$\|\mathcal{U}_k(t, s)z_0\| \leqq M\|z_0\| + \left(\int_s^t g^2(t-\alpha)\, d\alpha\right)^{1/2} \text{const. } \|z_0\|.$$

Hence $\mathcal{U}_k(t, s)$ is uniformly bounded. Using these uniform bounds it is easy to show that the convergence as $\varepsilon \to 0$ in Lemma 3.6 is uniform in k.

REFERENCES

[1] A. V. BALAKRISHNAN, *Boundary control of the diffusion equation. A semigroup theoretic approach*, to appear.

[2] RUTH F. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Riccati equation for systems defined by evolution operators*, this Journal, 14 (1976), pp. 951–983.

[3] RUTH F. CURTAIN, *Estimation theory for abstract evolution equations excited by general white noise processes*, this Journal, 14 (1976), pp. 1124–1150.

[4] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloquium Publications, vol. 31, American Mathematical Society, Providence, RI, 1957.

[5] T. KATO, *Abstract evolution equations of parabolic type in Banach and Hilbert spaces*, Nagoya Math. J., 19 (1961), pp. 93–125.

[6] T. KATO AND H. TANABE, *On the abstract evolution equation*, Osaka Math. J., 14 (1962), pp. 107–133.

[7] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[8] J. L. LIONS AND E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, 1972.

[9] J. RAUCH, *L^2 is a continuable initial condition for Kreiss' mixed problems*, Comm. Pure Appl. Math., 25 (1972), pp. 265–285.

[10] R. B. VINTER AND T. L. JOHNSON, *Optimal control of nonsymmetric hyperbolic systems in n variables on the half space*, this Journal, 15 (1977), pp. 129–143.

[11] K. YOSIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1966.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

THE NONLINEAR COMPLEMENTARITY PROBLEM: EXISTENCE AND DETERMINATION OF SOLUTIONS*

M. L. FISHER† AND J. W. TOLLE‡

Abstract. A general simplicial approximation algorithm with a variable initial point is presented for solving the nonlinear complementarity problem. The algorithm leads to a proof of a new existence theorem which unifies and extends previous results. Several previously known existence results are obtained as corollaries of this theorem.

1. Introduction. This paper concerns the nonlinear complementarity problem (hereafter referred to as NLCP): Find an $x \in R^n$ which satisfies

$$x \geqq 0, \qquad f(x) \geqq 0, \qquad \langle x, f(x)\rangle = 0,$$

where f is a continuous map of the nonnegative orthant, R^n_+, into R^n. Among the applications of this problem are included convex programming, computation of economic equilibria, N-person games, saddle point computation, and problems in structural mechanics. The problem of finding the fixed point of some $g: R^n_+ \to R^n_+$ may also be solved as an NLCP problem by taking $f(x) = x - g(x)$. Conversely, the NLCP problem can be formulated as a fixed point problem for $g: R^n_+ \to R^n_+$ defined by $g_j(x) = \max(x_j - f_j(x).\ 0)$, $j = 1, \cdots, n$, and also as the problem of finding solutions to the system of nonlinear equations[1] $h(x) = 0$ where, for instance, $h_j(x) = \min(x_j, f_j(x))$, $j = 1, \cdots, n$.

It appears that the study of NLCP was initiated in 1966 by Cottle [1]. This first paper has been followed by a spate of results formulating conditions on the function f which assure that a solution to NLCP exists. Included among these papers are those of Karamardian [10], [11], Eaves [2], Moré [21], [22], Kojima [12], [13], and Luna [17]. More recently, a number of authors have addressed the question of formulating constructive techniques for finding solutions. These works, including those of Fisher and Gould [4], Kojima [12], Garcia [7], Merrill [20], and Lüthi [18], have all used a variant of a method used by Scarf [23] in constructively proving the Brouwer fixed point theorem. This method of complementary pivoting on a triangulation of R^n is also the basis for this paper. In addition to the above papers, results such as those of Eaves and Saigal [3] and Merrill [20] on fixed point algorithms in unbounded regions and Gould and Tolle [9] and Wolsey [24] on zeros of nonlinear functions can be considered as contributions to the theory of the NLCP problem.

The purpose of this paper is to present a general method for finding solutions to the NLCP problem and to use this method to unify and extend previously known existence theory. The method is based on the complementary pivoting

* Received by the editors September 17, 1975, and in final revised form August 10, 1976.

† Department of Decision Sciences, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19174. The work of this author was supported in part by NSF Grant SOC-74-02516.

‡ Department of Mathematics and Curriculum in Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514.

[1] For other equivalent formulations, see Mangasarian [19].

principles which underlie most of the other algorithms developed for this problem. The method can be initiated from any point in R^n_+ and can employ any of a large family of possible labeling rules. Because the method operates over an unbounded region, finite termination in not guaranteed without additional qualification. It will be shown that finite termination is assured for any labeling rule which satisfies certain conditions on a "band" which separates the starting point from infinity in the nonnegative orthant. The existence of a labeling rule with the specified properties implies the existence of an NLCP solution. The specified rule required for the existence of a solution will be exhibited under a variety of conditions on the function f, including most of those previously known to guarantee that f has a solution.

In principle the existence theorems of this paper are constructive. That is, an algorithm is provided for finding an approximate solution given the starting point and the labeling rule. From a practical viewpoint, however, the prescription for guaranteed convergence restricts the set of allowed starting points and the choice of initial grid size. This restriction may result in a mode of operation which is computationally inefficient. Those faced with solving the NLCP problem may prefer to use a more efficient algorithm and hope for finite convergence.

Kojima [13] has recently given a theory for the NLCP problem which unifies many of the known existence theorems and Lüthi [18] has provided an algorithm which finds an approximate solution under the basic condition of Kojima. The existence theorem presented here is more general than that of Kojima, but the algorithm for finding the solution is less satisfactory in that arbitrary restarts are not in general permissible.

The organization of the paper is as follows: § 2 contains the fundamental properties of complementary pivoting on pseudomanifolds and the terminology and notation necessary for our applications; § 3 contains the development and proof of the fundamental existence theorem; in § 4 various new and known existence theorems are shown to follow from the theorem of § 3; and in § 5, remarks concerning implications of the paper are added.

2. Complementary pivoting.

2.1. Notation. It will be assumed that the reader is familiar with the general technique of complementary pivoting and simplicial approximation as described in Scarf [23]. A detailed exposition of both the theory and its applications is also presented in [8].

We employ the following notation. Points in R^{n+1} will be denoted with capital letters; points in R^n and scalars are denoted with small letters. Thus, we write $X = \begin{bmatrix} x \\ z \end{bmatrix}$, $X \in R^{n+1}$, $z \in R^1$. We construct the product space $R^n_+ \times [0, 1]$, which will be triangulated in such a way that all vertices are in $R^n_+ \times \{0\}$ and $R^n_+ \times \{1\}$. The simplices of interest will be:

1. those of dimension n ($n+1$ vertices) which lie either in

$$R^n_+ \times \{0\} \quad \text{or} \quad R^n_+ \times \{1\};$$

2. those of dimension $n+1$ ($n+2$ vertices) which lie in

$$R^n_+ \times [0, 1].$$

The boundary of the triangulation is by definition composed of those n-simplices which lie completely in one of the sets $R_+^n \times \{1\}$, $R_+^n \times \{0\}$, or $\{x : x_i = 0\} \times [0, 1]$, $i = 1, \cdots, n$. Each vertex of the triangulation will be labeled with an integer in $\{1, 2, \cdots, n+1\}$. Each simplex generated by the algorithm will have among its vertex labels all the labels $\{1, \cdots, n+1\}$. Such a simplex will be called an $(n+2)$-almost completely labeled (acl) simplex. The labeling will have the following properties:

A_1. There is a unique $(n+2)$-acl boundary simplex in $R_+^n \times \{0\}$.

A_2. There are no $(n+2)$-acl boundary simplices in $\{x : x_i = 0\} \times [0, 1]$ for any $i = 1, \cdots, n$.

A_3. Any $(n+2)$-acl n-simplex in $R_+^n \times \{1\}$ contains an approximate NLCP solution.

The algorithm will be initiated at the simplex described in A_1. Complementary pivoting will produce a sequence of $(n+2)$-acl simplices. Such a sequence is either unbounded or, by A_2, it terminates in $R_+^n \times \{1\}$ with an $(n+2)$-acl n-simplex. Such a simplex, according to A_3, furnishes an approximate solution to the problem of interest. Conditions for finite termination will be given.

We shall employ the triangulation of $R^n \times [0, 1]$ described in [8]. This triangulation is a modification of Kuhn's triangulation of a unit cube [14], [15]. The grid size of the triangulation will be denoted by δ and hence, if X, Y are points in an $(n+1)$-dimensional simplex of the triangulation, $\|x - y\| \leqq \sqrt{n}\delta$.

2.2. The labeling functions. A labeling function is any rule which assigns an integer between 1 and $n+2$ to each $X \in R_+^n \times \{0\} \cup R_+^n \times \{1\}$. In general, a labeling L will employ different rules for the vertices in $R_+^n \times \{0\}$ and $R_+^n \times \{1\}$. Consequently, it is convenient to let L_0 denote the rule in $R_+^n \times \{0\}$ and L_1 the rule in $R_+^n \times \{1\}$. The entire labeling will be denoted by $L = (L_0, L_1)$.

The rule we use to label X in $R_+^n \times [0, 1]$ depends on a specified point $w \in R_+^n$ and on the particular point X; that is, the rule for specifying the label will vary over the space $R_+^n \times [0, 1]$. Let w be a given point in R_+^n and let x denote a vector in R_+^n. Let $S_1^0(x, w), \cdots, S_{m_0}^0(x, w)$ and $S_1^1(x, w), \cdots, S_{m_1}^1(x, w)$ be two partitions of $N = \{1, 2, \cdots, n\}$ and set $h^0(x) = x - w$ and $h^1(x) = f(x)$. For each $x \in R_+^n$ these partitions and functions induce two orderings on N, $<_x^k$, $k = 0, 1$, as follows. Let $i \in S_l^k(x, w)$ and $j \in S_m^k(x, w)$; then $i <_x^k j$ if one of the following holds:

(i) $l < m$,

(ii) $l = m$ and $h_i^k(x) < h_j^k(x)$,

(iii) $l = m$, $h_i^k(x) = h_j^k(x)$ and $i < j$.

The labels of $\begin{bmatrix} x \\ 0 \end{bmatrix}$ and $\begin{bmatrix} x \\ 1 \end{bmatrix}$, denoted by $L_0(x)$ and $L_1(x)$, are now defined by

$$L_0(x) = \begin{cases} n+1 & \text{if } x - w > 0, \\ \min_{(0,x)} \{l : x_l - w_l \leqq 0\} & \text{otherwise,} \end{cases}$$

$$L_1(x) = \begin{cases} n+1 & \text{if } x > 0, f(x) > 0, \\ \min_{(1,x)} \{l : x_l = 0\} & \text{if } x \not> 0, \\ \min_{(1,x)} \{l : f_l(x) \leqq 0\} & \text{if } x > 0, f(x) \not> 0, \end{cases}$$

where $\min_{(k,x)}$ refers to the smallest integer with respect to the order $<_x^k$.

It will be seen in § 4 that different choices of the partitions of N will lead to different existence theorems for the NLCP problem.

Given L_0, the simplex required by A_1 is easily constructed as in [5]. This initial simplex is near w, and hence w can be regarded as the starting point of the algorithm. The definitions of L_0 and L_1 taken together establish immediately that the label $n+1$ cannot appear on a coordinate plane $\{x : x_i = 0\}$. Thus, property A_2 is satisfied for these labelings. Now consider property A_3. If C is a given compact subset of R^n_+ and $X = \begin{bmatrix} y \\ 1 \end{bmatrix}$ is any vector in an $(n+2)$-acl simplex in $C \times \{1\}$, then it follows from the definition of the labeling L_1 and the uniform continuity of f on C that y is an approximate solution to the NLCP problem. A proof of this assertion is included in Theorem 3.1. Note that properties A_1, A_2, and A_3 are enjoyed by any labeling rules L_0 (respectively L_1) that assign the label $n+1$ only when $x > w$ (respectively $x > 0$ and $f(x) > 0$) and the label $l \in \{1, \cdots, n\}$ only when $x_l > w_l$ (respectively $x_l = 0$ or $f_l(x) \leqq 0$). The sets $S_j^0(x, w)$ and $S_j^1(x, w)$ given above provide for the selection of a unique label when more than one component of $x - w$ or $f(x)$ is nonpositive.

3. Finite termination and the fundamental existence theorem. Given w in R^n_+, we employ the concept of a *separating set* and a *band*. The idea of a separating set is purely topological. The concept of a band is labeling-dependent.

DEFINITION 1. Suppose w, A, and B are such that A is bounded and open in R^n, $w \in A \cap R^n_+$, and $B = \partial A \cap R^n_+$. Then B is said to *separate* w *from infinity*. Whenever we refer to the pair (w, B), it is understood that B separates w from infinity.

DEFINITION 2. Suppose (w, B) and L are such that for each x in B there is a label $l \in \{1, \cdots, n+1\}$, l depending upon x, such that this label cannot occur in some neighborhood of x. Then the triple (w, B, L) is said to be a *band*.

Our methods for proving termination and existence are based upon the next two results. Henceforth, the symbol A will always denote the bounded open set such that $B = \partial A \cap R^n_+$, and $\bar{A}$ will denote the closure of A.

THEOREM 3.1. *Suppose (w, B, L) is a band. Then if the grid size δ is sufficiently small, the algorithm with the labeling L terminates in $R^n \times \{1\}$ and $\bar{A} \cap R^n_+$ contains an NLCP solution.*

Proof. B is compact. Cover B with a finite number of open balls of radius $\varepsilon_i/2$ about $x^i \in B$, such that for each ball there is an index j_i which cannot appear as the label of any point Y if y is in a ball of radius ε_i about x^i. Choose δ so that the first simplex is entirely in the set A and so that $\delta < \min_i \varepsilon_i/(2\sqrt{n})$. If the algorithm does not terminate in $A \times \{1\}$, there must be a simplex $S \in R^n \times [0, 1]$ which intersects $B \times [0, 1]$. Consequently, there will be an $\begin{bmatrix} x \\ x_{n+1} \end{bmatrix}$ in S such that $x \in B$, and hence x is contained in the $\varepsilon_i/2$ ball about some x^i. By the choice of δ, the projection of S into R^n is entirely in the ball of radius ε_i about x^i and consequently the label j_i cannot occur on any vertex of S. Thus, S cannot be $(n+2)$-acl. Therefore, the path of $(n+1)$-dimensional simplices generated by the algorithm initiated at w cannot cross B; it must remain in a bounded region. Since it is a basic property of the complementary pivoting algorithm that no simplex can occur

twice, it follows that the algorithm must terminate with an $(n+2)$-acl simplex in $A \times \{1\}$.

To prove that $A \cap R_+^n$ contains an NLCP solution, let η_k be a sequence of positive scalars with $\eta_k \to 0$. Then choose δ_k to be a sequence of positive scalars tending to 0 such that if $\begin{bmatrix} x^k \\ 1 \end{bmatrix}$ is a point of the $(n+2)$-acl terminal simplex obtained by initiating the algorithm at w with labeling L and mesh size δ_k, then x^k satisfies $f_i(x^k) \geqq -\eta_k$ for all i and $f_i(x^k) \leqq \eta_k$ if $x_k > \delta_k$. That this can be done is a consequence of the definition of the labeling L_1 and the uniform continuity of f on $\bar{A}$. The point x^k is contained in $A \cap R_+^n$ and since $\eta_k \to 0$, any limit point of the sequence x^k is an NLCP solution. □

The application of Theorem 3.1 to a particular NLCP problem relies upon being able to show that at each point $\hat{x}$ on a set B there is a neighborhood of $\hat{x}$, $N_\varepsilon(\hat{x})$, and an index $j \in \{1, \cdots, n+1\}$ such that $L_0(x) \neq j$ and $L_1(x) \neq j$ for all $x \in N_\varepsilon(\hat{x})$. In Theorem 3.2, we shall delineate the circumstances under which an index does not occur in a neighborhood of a point $\hat{x}$. Then in § 4 we shall provide conditions on the function f which will guarantee the existence of a separating set B and a labeling L such that one of the hypotheses of Theorem 3.2 will be satisfied at each point in the set. The triple (w, B, L) will then be a band, and Theorem 3.1 will assure existence of a solution to the NLCP problem.

THEOREM 3.2. *Let $\hat{x} \in R_+^n$, $w \in R_+^n$, and partitions $S_j^0(x, w)$, $j = 1, \cdots, m_0$, and $S_j^1(x, w)$, $j = 1, \cdots, m_1$, be given. Then there is a neighborhood $N_\varepsilon(\hat{x})$ of $\hat{x}$ and an index $j \in \{1, \cdots, n+1\}$ such that $L_0(x) \neq j$ and $L_1(x) \neq j$ for all $x \in N_\varepsilon(\hat{x})$ if any one of the following conditions is satisfied*:

Γ_1. $(\hat{x} - w) \ngeqq 0$ *and* $f(\hat{x}) \ngeqq 0$.

Γ_2. *There exists an index j such that $(\hat{x}_j - w_j) f_j(\hat{x}) > 0$.*

Γ_3. *There exist indices j, k such that $f_k(\hat{x}) < 0$, $\hat{x}_j - w_j > 0$, and $k <_x^1 j$ for all x in an open set containing $\hat{x}$.*

Γ_4. *There exist indices j, k such that $\hat{x}_j - w_k < 0$, $\hat{x}_j > 0$, $f_j(\hat{x}) > 0$, and $k <_x^0 j$ for all x in an open set containing $\hat{x}$.*

Proof. If Γ_1 holds, then there are indices i and k such that $f_i(x) < 0$ and $(x_k - w_k) < 0$ for x in a neighborhood of $\hat{x}$. It is then apparent from the definitions of $L_0(x)$ and $L_1(x)$ that the label $n+1$ cannot occur in this neighborhood.

If Γ_2 holds, then either $(\hat{x}_j - w_j) < 0$ and $f_j(\hat{x}) < 0$ or $(\hat{x}_j - w_j) > 0$ and $f_j(\hat{x}) > 0$. In the former case, Γ_1 holds. In the latter case, we also have $\hat{x}_j > w_j \geqq 0$, and it is easily seen that there is a neighborhood of $\hat{x}$ in which the label j cannot occur.

If Γ_3 holds then $\hat{x}_j - w_j > 0$ implies $L_0(x) \neq j$ in a neighborhood of $\hat{x}$, and $\hat{x}_j > w_j \geqq 0$, $f_k(\hat{x}) < 0$, and $k <_x^1 j$ in a neighborhood of $\hat{x}$ implies $L_1(x) \neq j$ in a neighborhood of $\hat{x}$.

The argument for Γ_4 is made in a manner analogous to that for Γ_3. □

It should be observed that in a sense the conditions given in Theorem 3.2 are necessary as well as sufficient for the a priori exclusion of a label in a neighborhood of $\hat{x}$. Any further conditions would be dependent on the specific form of the function f. For example, if there were a j such that $(\hat{x}_j - w_j) = 0$ and $f_j(\hat{x}) = 0$, then the label $n+1$ would be excluded at $\hat{x}$. But it is possible that in any neighborhood of $\hat{x}$ there would be an x such that $x - w > 0$ or $f(x) > 0$ so that the label $n+1$ could occur.

If the sets $S_j^1(x, w)$, $j = 1, \cdots, m$, are independent of x or are of special form, then the condition in Γ_3 that "$k <_x^1 j$ for all x in an open set containing $\hat{x}$" can be replaced by "$k <_x^1 j$". A similar simplification hold for Γ_4. In all the situations which we encounter in § 4, these simpler versions of Γ_3 and Γ_4 will be applicable.

Theorems 3.1 and 3.2 together with the definition of a band yield the fundamental existence theorem of this paper.

THEOREM 3.3. *Let A be an open bounded set in R^n and $w \in A \cap R_+^n$. Suppose there exist partitions $S_j^0(x, w), j = 1, \cdots, m_0$, and $S_j^1(x, w), j = 1, \cdots, m_1$, such that at each $\hat{x} \in B = \partial A \cap R_+^n$ at least one of the conditions Γ_1–Γ_4 of Theorem 3.2 hold. Then the set $\bar{A} \cap R_+^n$ contains a solution to the NLCP problem.*

The algorithm we have given may be used to actually compute, in the limit, the solution described in Theorem 3.3. provided that the vector w and a satisfactory δ are known. That is, if the algorithm is successfully applied, starting at w, with a sequence of sufficiently small grid sizes which tend to zero, then a point may be selected in each terminal simplex to obtain a sequence x^k which converges to an NLCP solution. The proof of each existence result given in § 4 specifies the particular labeling one would use to compute a solution in this manner. Generally, it would be advantageous to initiate the algorithm on iteration $k+1$ using a revised $w = x^k$ if possible. Reference [6] describes computational experience using this approach with some convex programming problems.

4. Existence theorems for the NLCP problem. In this section we state conditions on f which allow us to display pairs (w, B) and labeling sets $S_j^0(x, w)$ and $S_j^1(x, w)$ for which the hypotheses of Theorem 3.3 are satisfied. The theorems proved below include as special cases many of the known theoretic existence results.

THEOREM 4.1. *Suppose there is a set $B = \partial A \cap R_+^n$ separating the origin from infinity such that for each $x \in B$ the following system is inconsistent*:

$$
\begin{aligned}
& f_i(x) + t = 0, \qquad x_i \geqq 0, \\
& f_i(x) + t \geqq 0, \qquad x_i = 0, \\
& t \geqq 0.
\end{aligned}
\tag{4.1}
$$

Then NLCP has a solution x^* with $x^* \in A$.

Proof. We set $w = 0$, $S_1^0(x, w) = \{1, \cdots, n\}$, and $S_1^1(x, w) = \{1, \cdots, n\}$. Then for $\hat{x} \in B$ we set $I = \{i : \hat{x}_i > 0\}$ and distinguish three cases.

Case (i). There exists a $j \in I$ such that $f_j(\hat{x}) > 0$. It is immediate that Γ_2 holds at $\hat{x}$.

Case (ii). There exists a $j \in I$ such that $0 \geqq f_j(\hat{x}) > \min_k (f_k(\hat{x}))$. From the definition of $S_1^1(\hat{x}, w)$, it follows that Γ_3 holds at $\hat{x}$.

Case (iii). $0 \geqq f_j(\hat{x}) = \min_k (f_k(\hat{x}))$ for every $j \in I$. Setting $t = -\min_k (f_k(\hat{x}))$, we see that the system of the hypothesis is consistent which is a contradiction. Thus, this case cannot occur.

Now since Γ_2 or Γ_3 hold, Theorem 3.3 implies the existence of an NLCP solution $x^* \in \bar{A}$. The inconsistency of (4.1) precludes $x^* \in B$, so $x^* \in A$. □

As a consequence of Theorem 4.1, we can deduce the following existence result.

COROLLARY 4.2 (Karamardian [11]). *Let $G(x)=f(x)-f(0)$ be positively homogeneous of degree $d>0$ and suppose the system*

$$\begin{aligned} &G_i(x)+t=0, && x_i>0,\\ &G_i(x)+t\geqq 0, && x_i=0\\ &t\geqq 0, \end{aligned} \tag{4.2}$$

is inconsistent for all $x\geqq 0$, $x\neq 0$. Then NLCP has a solution.

Proof. The function $F(x)=G(x)+(1-\sum_{i=1}^n x_i)f(0)$ satisfies the assumption of Theorem 4.1 with $B=[x\geqq 0: \sum_{i=1}^m x_i=1\}$. Hence the NLCP $F(x)\geqq 0$, $x\geqq 0$, $\langle F(x), x\rangle=0$ has a solution x' with $\sum_{i=1}^n x_i'<1$. Now let $\bar{x}=x'/(1-\sum_{i=1}^n x_i')^{1/d}$. The homogeneity of G implies directly that $f_i(\bar{x})=F_i(x')/(1-\sum_{i=1}^n x_i')$ so that $\bar{x}$ is a solution of the original NLCP. □

The following theorem establishes the existence of a solution to the NLCP problem in terms of the existence of solutions to a linear system on a separating set, B. See Kojima [13], Lüthi [18], and Fisher, Gould and Tolle [5].

THEOREM 4.3. *Suppose a set $B=\partial A\cap R_+^n$ separates $w\geqq 0$ from infinity. Suppose that for each $x\in B$ there is a $y\geqq 0$ such that the system*

$$\begin{aligned} &(x-w-y)^T f(x)>0,\\ &(x-w-y)^T e\geqq 0 \quad \textit{when } (x-w)\geqq 0 \end{aligned} \tag{E_1}$$

has a solution. Then the set $\bar{A}$ contains a solution to the NLCP problem.

Proof. Set $S_1^0(x, w)=S_1^1(x, w)=\{1, \cdots, n\}$. Let $x\in B$. We distinguish several cases. If $f(x)\geqq 0$, then $(x-w)^T f(x)>y^T f(x)\geqq 0$ and Γ_2 holds at x. If $f(x)\ngeqq 0$ and $(x-w)\ngeqq 0$, then Γ_1 holds at x. If $f(x)\ngeqq 0$, $(x-w)\geqq 0$, and there exist integers l and k such that $(x_l-w_l)>0$, $f_k(x)<0$, and $f_l(x)>f_k(x)$, then it follows from the definition of $S_1^1(x, w)$ that Γ_3 holds at x. In the remaining possibility we have $M=\min_k (f_k(x))>0$ and $x_l-w_l=0$ for every l such that $f_l(x)>M$. Setting $I=\{l: f_l(x)=M\}$, we have from the second inequality of E_1, $\sum_{j\in I}(x_j-w_j-y_j)\geqq \sum_{j\notin I} y_j$. Thus,

$$\begin{aligned} \sum_{j=1}^n (x_j-w_j-y_j)f_j(x) &= \textstyle\sum_{j\in I}(x_j-w_j-y_j)f_j(x)-\sum_{j\notin I} y_j f_j(x)\\ &\leqq M\sum_{j\in I}(x_j-w_j-y_j)-M\Big(\sum_{j\notin I} y_j\Big)\\ &\leqq 0, \end{aligned}$$

which contradicts the first inequality of E_1. Thus, this last possibility cannot occur and we have that one of the conditions of Theorem 3.2 holds at each $x\in B$ for which E_1 holds. □

A modification of the inequalities E_1 yields the following result, stated without proof.

THEOREM 4.4. *For any $x \in B$ the system E_1 has a solution with $y \geqq 0$ if and only if the system E_2 given below has a solution with $y \geqq 0$.*

$$E_2 \qquad \begin{aligned} &(x-w-y)^T f(x) \geqq 0, \\ &(x-w)^T f(x) \neq 0, \qquad \text{when } f(x) \geqq 0, \\ &(x-w-y)^T e > 0, \qquad \text{when } (x-w) \geqq 0. \end{aligned}$$

We can obtain a slightly different version of Theorems 4.3 and 4.4.

COROLLARY 4.5. *Suppose B separates $w \geqq 0$ from infinity and there exists a positive vector d such that for every $x \in B$ there is a $y \geqq 0$ satisfying one of the two equivalent systems:*

$$E_1' \qquad \begin{aligned} &(x-w-y)^T f(x) > 0, \\ &(x-w-y)^T d \geqq 0, \qquad \text{when } (x-w) \geqq 0; \end{aligned}$$

$$E_2' \qquad \begin{aligned} &(x-w-y)^T f(x) \geqq 0, \\ &(x-w)^T f(x) \neq 0, \qquad \text{when } f(x) \geqq 0, \\ &(x-w-y)^T d > 0, \qquad \text{when } (x-w) \geqq 0. \end{aligned}$$

Then the NLCP problem has a solution.

Proof. Let C be the $n \times n$ diagonal matrix with diagonal elements $1/d_1$, $1/d_2, \cdots, 1/d_n$. Set $z = C^{-1}x$, $\hat{w} = C^{-1}w$, $\hat{y} = C^{-1}y$, $\hat{B} = \{z : z = C^{-1}x$ for some $x \in B\}$, and $g(z) = Cf(Cz)$. Then it is evident that $\hat{B}$, $\hat{w}$, and $\hat{y}$ satisfy E_1 (or E_2) for the function $g(z)$. Hence, the NLCP problem for $g(z)$ has a solution z^*. It follows that $x^* = Cz^*$ is a solution of the NLCP problem for $f(x)$. □

Note that knowledge of d would be necessary in order to implement the algorithm in the case of Corollary 4.5.

The next three corollaries are existence theorems found in the literature Kojima [13] has also shown them to be corollaries of Theorem 4.3.

COROLLARY 4.6 (Karamardian [11]). *If there is a nonempty compact set C in R_+^n such that for each $x \in R_+^n - C$ there is a $y \in C$ such that $(x-y)^T f(x) > 0$. then NLCP has a solution.*

Proof. Let $w = 0$ and $r > 0$ be any scalar such that $B = \{x \in R_+^n : e^T x = r\}$ separates each $y \in C$ from infinity. Then the system E_1 in Theorem 4.3 is satisfied for each $x \in B$. □

COROLLARY 4.7 (Karamardian [11]). *Let $H = \{x : |x| \leqq r, r > 0\}$ or $H = \{x : e^T x \leqq r, r > 0\}$. Then NLCP has a solution if $x^T f(x) \geqq 0$ for all $x \in \partial H \cap R_+^n$.*

Proof. If $x^T f(x) = 0$ when $f(x) \geqq 0$ on $\partial H \cap R_+^n$ then x is a solution. Otherwise, the system E_2 in Theorem 4.4 is satisfied by choosing $w = 0$, $y = 0$, and $B = \partial H \cap R_+^n$. □

COROLLARY 4.8 (Eaves [2]). *NLCP has a solution if there exists a positive n-vector d, a positive scalar r, and a set B separating $C = \{y \in R_+^n : y^T d \leqq r)$ from infinity such that for each $x \in B$ there is a $y \in C$ for which $(x-y)^T f(x) \geqq 0$.*

Proof. Set $w = 0$. Since B separates C from infinity, $x^T d > r$ for each $x \in B$. Thus for each $y \in C$, $(x-y)^T d > 0$. Now if $x^T f(x) = 0$ when $f(x) \geqq 0$, x is a solution to NLCP. Otherwise the system E_2' of Corollary 4.5 is satisfied for each $x \in B$. □

In the previous theorems the labeling sets $S_j^0(x, w)$ and $S_j^1(x, w)$ have been independent of x and w. In this next theorem $S_j^1(x, w)$ will depend on x and w. Let $m_1(x)$ denote the number of different values assumed by the components of the n-vector $x - w$. Then

$$S_1^1(x, w) = \left\{ i : x_i - w_i = \min_k (x_k - w_k) \right\},$$

$$S_j^1(x, w) = \left\{ i : x_i - w_i = \min_{k \notin \cup_{l=1}^{j-1} S_l^1(x,w)} (x_k - w_k) \right\}, \qquad j = 2, \cdots, m_1(x),$$

$$S_1^0(x, w) = \{1, \cdots, n\}.$$

THEOREM 4.9. *Suppose $B = \partial A \cap R_+^n$ separates $w \geqq 0$ from infinity and for every $x \in B$ there is a $y \geqq 0$ such that the following system is satisfied:*

$$\mathrm{E}_3 \qquad \begin{aligned} &\max_i [(x_i - w_i - y_i) f_i(x)] > 0, \\ &\max_i (y_i) \leqq \max_i (x_i - w_i) \quad \textit{if} \quad (x - w) \geqq 0. \end{aligned}$$

Then NLCP has a solution in $\bar{A}$.

Proof. Let $S_1^0(x, w)$ and the $S_j^1(x, w)$ be as described below. Let $\hat{x} \in B$ and set $I = \{i : f_i(\hat{x}) < 0\}$. We consider five possible cases.

(i) $f(\hat{x}) \geqq 0$. By E_3 there is an index i such that $(\hat{x}_i - w_i) f_i(\hat{x}) > y_i f_i(\hat{x}_i) \geqq 0$ so that Γ_2 holds at $\hat{x}$.

(ii) $f(\hat{x}) \ngeqq 0$ and $(\hat{x} - w) \ngeqq 0$. Then Γ_1 holds at $\hat{x}$.

(iii) $f(\hat{x}) \ngeqq 0$, $(\hat{x} - w) \geqq 0$, and there exists an i such that $(\hat{x}_i - w_i) f_i(\hat{x}) > 0$. Then Γ_2 holds at $\hat{x}$.

(iv) $f(\hat{x}) \ngeqq 0$, $(\hat{x} - w) \geqq 0$, and there exists a j such that $\hat{x}_j - w_j > \hat{x}_i - w_i$ for some $i \in I$. Then for x in some open neighborhood of $\hat{x}$, $i <_x^1 j$. Since $f_i(\hat{x}) < 0$ and $\hat{x}_j - w_j > 0$, it follows that Γ_3 holds.

(v) $f(x) \ngeqq 0$, $(\hat{x} - w) \geqq 0$, $\hat{x}_i - w_i = \max_k (\hat{x}_k - w_k)$ for all $i \in I$, and $(\hat{x}_j - w_j) f_i(\hat{x}) \leqq 0$ for all j. We demonstrate that the existence of a solution to E_3 precludes this case. Suppose $i \in I$; then $(\hat{x}_i - w_i) \geqq y_i$ and $f_i(\hat{x}) < 0$ imply $(\hat{x}_i - w_i - y_i) f_i(\hat{x}) \leqq 0$. If $i \notin I$, then $f_i(x) \geqq 0$ implies $(x_i - w_i - y_i) f_i(x) \leqq (x_i - w_i) f_i(x) < 0$. Thus for all i, E_3 cannot hold. □

COROLLARY 4.10 (Moré [21], [22]). *NLCP has a solution if there exists a positive scalar r such that for every $x \in R_+^n$ with $\max_i x_i = r$ there is a $y \geqq 0$ such that*

$$\mathrm{E}_4 \qquad \begin{aligned} &y_i \leqq r \quad \textit{for each } i, \\ &\max_i [(x_i - y_i) f_i(x)] > 0. \end{aligned}$$

Proof. The corollary follows immediately from Theorem 4.9 by setting $w = 0$ and $B = \{x \in R_+^n : \max_i x_i = r\}$. □

Consider the nonlinear programming problem

$$\begin{aligned} &\min g_0(z), \\ &g(z) \leqq 0, \\ &z \geqq 0 \end{aligned} \tag{4.3}$$

where g_0 maps R_+^k to R, g maps R_+^k to R^m, and g_0 and g are convex and continuously differentiable on an open set containing R_+^k. If u denotes an m-component vector of dual variables, the Kuhn–Tucker conditions for this problem are given by the NLCP with $x = (z, u)$ and

$$f(x) = (\nabla_z g_0(z) + u^T \nabla_z g(z), -g(z)). \tag{4.4}$$

The numerous constraint qualification results of nonlinear programming are concerned with existence conditions for this NLCP. The following theorem develops one of these results from Theorem 3.3.

THEOREM 4.11. *The NLCP defined by* (4.4) *has a solution if there exists a point* $z^0 \geqq 0$ *for which* $g(z^0) < 0$ *and if the set of optimal solutions for* (4.3) *is nonempty and bounded.*

Proof. We will show that the conditions of Theorem 3.3 hold with $S_1^1(x, w) = \{k+1, \cdots, k+m\}$, $S_1^0(x, w)$, $S_j^0(x, w)$, $S_j^1(x, w)$, $j \geqq 2$ arbitrary, and $w = (z^0, 0)$. Let

$$F^1 = \{x \geqq 0 : (x - w)^T f(x) \leqq 0\},$$

$$F^2 = \{x \geqq 0 : g(x) \leqq 0\}$$

and

$$F^3 = \{x \geqq 0 : x_i - w_i \leqq 0,\ i \notin S_1^1(x, w)\}.$$

We first establish the boundedness of $F^1 \cap (F^2 \cup F^3)$.

It will then be useful to have a lower bound on $(x - w)^T f(x)$. By definition,

$$(x - w)^T f(x) = (z - z^0)^T \nabla_z g_0(z) + \sum_{i=1}^{m} u_i (z - z^0)^T \nabla_z g_i(z) - u^T g(z).$$

By convexity,

$$g_i(z^0) \geqq g_i(z) - (z - z^0)^T \nabla_z g_i(z), \qquad i = 0, 1, \cdots, m.$$

Hence

$$(x - w)^T f(x) \geqq g_0(z) - g_0(z^0) - u^T g(z^0).$$

Since $-u^T g(z^0) \geqq 0$, if $g_0(z) > g_0(z^0)$ then $(x - w)^T f(x) > 0$. Because the optimal set is bounded, the set of feasible z for which $g_0(z) \leqq g_0(z^0)$ is also bounded. If $x = (z, u) \in F^2$, then z is feasible. If $x \in F^3$, then $|z| \leqq \gamma |z^0|$. Hence, there exists an $r > 0$ such that $F^1 \cap (F^2 \cup F^3) \subseteq \{x \in R_+^n : |z| \leqq r\}$. Because $g_0(z) - g_0(z^0)$ is bounded on $\{x \in R_+^n : |z| \leqq r\}$ and $g(z^0) < 0$, if $|u|$ is sufficiently large for any $x \in \{x : |z| \leqq r\}$ we have $(x - w)^T f(x) > 0$. Hence, $F^1 \cap (F^2 \cup F^3)$ is bounded.

Now let B be any set separating $F^1 \cap (F^2 \cup F^3)$ from infinity and satisfying $B \cap F^1) (F^2 \cup F^3) = \varnothing$. For $x \in B$, either $x \notin F^1$ in which case $(x_j - w_j) f_j(x) > 0$ for

some j and Γ_2 holds or $x \notin F^2 \cup F^3$ in which case $f_i(x) < 0$ for some $i \in S_1^1(x, w)$ and $x_j - w_j > 0$ for some $j \notin S_1^1(x, w)$; so Γ_3 holds. □

5. Additional remarks. We note that the existence results of Karamardian, Eaves, and Moré given in Corollaries 4.6, 4.7, 4.8, and 4.10 are obtained by setting $w = 0$ in more general theorems. The possibility of having $w \neq 0$ is an important aspect of our results. It can be essential in proving the existence of a solution since properties of $f(x)$ defined relative to the origin need not have a special significance in characterizing a solution which may not be near the origin. For example, consider the two-dimensional problem with $f_1(x) = (x_1 - 2)^2 - 1$ and $f_2(x) = x_1 - 2$. The example has a unique solution at $x_1 = 3$, $x_2 = 0$. Any set which separates the origin from infinity must include a point $\bar{x}$ with $\bar{x}_1 = 0$ and $\bar{x}_2 > 0$. At $\bar{x}$ we must have $f_1(\bar{x}) > 0$ and $f_2(\bar{x}) < 0$. This prevents the hypotheses of Corollaries 4.6, 4.7, 4.8, and 4.10 from being satisfied. Letting $G_1(x) = f_1(x) - f_1(0) = x_1(x_1 - 4)$ and $G_2(x) = f_2(x) - f_2(0) = x_1$, we have $G_1(\bar{x}) = G_2(\bar{x}) = 0$. Hence, $t = 0$ solves system (4.2) and Corollary 4.2 also fails. On the other hand, if $w = (2.5, 0)$ and $A = \{x : -1 \leqq x_1 \leqq 4,\ -1 \leqq x_2 \leqq 1\}$, then $B = \partial A \cap R_+^n$ is a band with any specification of the sets $S_j^0(x, w)$ and $S_j^1(x, w)$, and the hypotheses of Theorem 3.3, Theorem 4.9, and Corollary 4.5 for arbitrary $d > 0$ are all satisfied.

In general, the algorithm used to prove Theorem 3.3 must be restarted from the same w after each iteration. If the conditions for existence hold for some vector $\hat{w}$ in the terminal simplex of an iteration, then the algortihm can be reinitiated from this point. Such a procedure would have obvious computational advantages. However, to make an a priori assumption that this situation occurs would greatly weaken the generality of the existence theorems given in this paper. The following observation suggests that once a neighborhood of the true solution is reached, a special labeling will allow the algorithm to be restarted from the terminal simplex of the preceding iteration.

Suppose that x^* is a nondegenerate solution to NLCP and that the matrix $H = (h_{ij})$ is nonsingular, where

$$h_{ij} = \begin{cases} \left.\dfrac{\partial f_i(x)}{\partial x_j}\right|_{x = x^*} & x_i^* > 0, \\ \delta_{ij}, & x_i^* = 0. \end{cases}$$

Then, it can be shown that there exist sets $S_j^0(x, x^*)$, $S_j^1(x, x^*)$ and a separating set B such that (x^*, B, L) is a band for the function f.[2] Thus, the existence of a band is both necessary and sufficient for the existence of a solution satisfying the above condition. This suggests that the basic existence theorem of § 3 is about as general as one could expect for f continuously differentiable.

The algorithm given here is a nonlinear analogue of the algorithm developed by Lemke [16] for solving the LCP problem. As in Lemke's algorithm, an "almost-complementary" path is followed along the contours at which $(n - 1)$ of the variables (possibly including artificial variables) are zero, and switching from one contour to another occurs near a "vertex" where n of the variables are zero.

[2] For a proof of a similar result for the problem of finding zeros of a function, see [5] or [9].

The path will not reach a neighborhood of a solution if an "infinite" almost-complementary contour is followed. The major existence theorem of this paper, Theorem 3.3, is achieved by postulating the existence of a "band" or "barrier" on the almost-complementary contours which prohibits the algorithm from following these infinite paths.

Finally, it should be noted that in many cases at least one of the conditions Γ_1–Γ_4 will automatically be satisfied at all but a finite number of points on a potential separating set B. For instance, if $S_j^0(x, w) = \{j\}$ and $S_j^1(x, w) = \{j\}$ for $j = 1, \cdots, n$ and all x, then for any $\hat{x} \in B$ at which there exist two indices k, l such that

$$\hat{x}_j(\hat{x}_j - w_j)f_j(\hat{x}) \neq 0, \qquad j = k, l, \tag{5.1}$$

at least one of the conditions Γ_1–Γ_4 hold. The points at which (5.1) does not hold are on the intersection of at least $n-1$ zero contours and such points will often comprise only a finite number of points on B. For further results in this vein, the reader is referred to [9].

Acknowledgments. This paper has benefited from numerous detailed suggestions from the referees and from conversations with F. J. Gould.

REFERENCES

[1] R. W. COTTLE, *Nonlinear programs with positively bounded Jacobians*, SIAM J. Appl. Math., 14 (1966), pp. 147–158.

[2] B. C. EAVES, *On the basic theorem of complementarity*, Math. Programming, 1 (1971-d), pp. 68–75.

[3] B. C. EAVES AND R. SAIGAL, *Homotopies for computation of fixed points on unbounded regions*, Ibid., 3 (1972), pp. 225–237.

[4] M. L. FISHER AND F. J. GOULD, *A simplicial algorithm for the nonlinear complementarity problem*, Ibid., 6 (1974), pp. 281–300.

[5] M. L. FISHER, F. J. GOULD AND J. W. TOLLE, *A simplicial approximation algorithm for solving systems of nonlinear equations*, Proc. Conf. on Math. Programming and Its Applications, National Institute of Higher Mathematics, City University, Rome, Italy, April 1974, Academic Press, New York, 1976.

[6] ———, *A simplicial algorithm for the mixed nonlinear complementarity problem with applications to convex programming*, Center for Math. Studies in Business and Economics Rep. 7424, Graduate School of Business, Univ. of Chicago, June 1974.

[7] C. B. GARCIA, *The complementarity problem and its applications*, Ph.D. dissertation, Rensselaer Polytechnic Institute, Troy, N.Y., June 1973.

[8] F. J. GOULD AND J. W. TOLLE, *A unified approach to complementarity in optimization*, J. Discrete Math., 7 (1974), pp. 225–271.

[9] ———, *An existence theorem for solutions to* $f(x) = 0$, Math. Programming, to appear.

[10] S. KARAMARDIAN, *The nonlinear complementarity problem with applications. Part I and Part II*, J. Optimization Theory Appl., 4 (1969), pp. 87–88 and pp. 167–181.

[11] ———, *The complementarity problem*, Math. Programming, 2 (1972), pp. 107–129.

[12] M. KOJIMA, *Computational methods for solving the nonlinear complementarity problem*, Keio Engrg. Rep., 27 (1974), No. 1, pp. 1–41.

[13] ———, *A unification of the existence theorems of the nonlinear complementarity problem*, Math. Programming, 9 (1975), pp. 257–277.

[14] H. W. KUHN, *Simplicial approximation of fixed points*, Proc. Nat. Acad. Sci. U.S.A., 61 (1968), pp. 1238–1242.

[15] ———, *Some combinatorial lemmas in topology*, IBM J. Res. Develop., 4 (1960), pp. 518–524.
[16] C. E. LEMKE, *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11 (1965), pp. 681–689.
[17] G. LUNA, *A remark on the nonlinear complementarity problem*, Proc. Amer. Math. Soc., 48 (1975), pp. 132–134.
[18] H. J. LÜTHI, *A simplicial approximation of a solution for the nonlinear complementarity problem*, Math. Programming, 9 (1975), pp. 278–293.
[19] O. L. MANGASARIAN, *Equivalence of the complementarity problem to a system of nonlinear equations*, Tech. Rep. 227, Computer Sci. Dep. Univ. of Wisconsin, Madison, November 1974.
[20] O. H. MERRILL, *Applications and extensions of an algorithm that computes fixed points of certain nonempty upper semicontinuous point to set mappings*, Tech. Rep. 71–7, Dep. of Industrial Engineering, Univ. of Michigan, Ann Arbor, September 1971.
[21] J. J. MORÉ, *Classes of functions and feasibility conditions in nonlinear complementarity problems*, Math. Programming, 6 (1974), pp. 327–338.
[22] ———, *Coercivity conditions in nonlinear complementarity problems*, SIAM Rev., 16 (1974), pp. 1–16.
[23] H. SCARF, *The approximation of fixed points of a continuous mapping*, SIAM J. Appl. Math., 15 (1967), pp. 1328–1343.
[24] L. A. WOLSEY, *Convergence, simplicial paths, and acceleration methods for simplicial approximation algorithms for finding a zero of a system of nonlinear equations*, Center for Operations Research and Econometrics, Heverlee, Belgium, Discussion Paper 7427, December 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

SOME ANALYTIC AND MEASURE THEORETIC PROPERTIES OF SET-VALUED MAPPINGS*

MARY BRADLEY† AND RICHARD DATKO‡

Abstract. A theory of differentiation for set-valued mappings in a separable reflexive Banach space X is presented. The investigation is centered around the differentiability of the support functionals and thus only computation of limits of real functions is required. Our results include a Radon–Nikodym theorem for set-valued measures taking closed bounded convex values.

1. Introduction. In this paper a theory of differentiation for set-valued mappings taking values in a separable reflexive Banach space is presented. The basic idea is to consider mappings taking values in closed bounded convex sets and to approach differentiation through consideration of the differentiability of the support functionals of the sets. This approach is made possible by the fact that the integrals of the support functionals of set-valued mappings and the support functionals of the closures of the integrals are equal (see [12]). Hence any property of integration which is common to every support functional is common to the set-valued mapping if its values are taken in closed bounded convex sets.

Other notions of differentiability for set-valued mappings have been discussed by Artstein [1], Banks and Jacobs [4], Bridgland [5], Hermes [21] and Hukuhara [23]. Banks and Jacobs use Radström's embedding result [28] and reduce the investigation to differentiation in a normed linear space. Artstein and Hermes both obtain results for set-valued functions in R^n which are representable as indefinite integrals. Hukuhara considers mappings with values in the compact convex subsets of R^n and defines differentiation in terms of the convex set difference and the Hausdorff metric.

In § 4 we consider a Radon–Nikodym theorem for weak set-valued measures with values in a separable reflexive Banach space. This theorem is similar to the work in [9], [10], [11], [26] and particularly [17]. However, although the space considered in [17] is more general than ours, our result is stronger due to the fact that we consider integrably bounded measures as opposed to uniformly bounded measures. Furthermore we believe our proof is more direct in that it utilizes only the standard Radon–Nikodym theorem for scalar measures and a variant of the Hahn–Banach theorem.

In § 5 we give some applications of our results.

2. Preliminaries. Let X denote a separable reflexive Banach space over the real numbers R. The topological dual of X will be denoted by X', and S' will denote the surface of the unit ball in X'. If $x' \in X'$ and $K \subset X$, $x'(K)$ is defined as

$$x'(K) = \sup \{x'(k): k \in K\}.$$

The norm in X and X' will be denoted by $|\cdot|$. If $x' \in X'$ then

$$|x'| = \sup \{|x'(x)|: x \in X \text{ and } |x| = 1\}.$$

* Received by the editors December 18, 1975, and in revised form August 10, 1976.
† Department of Mathematics, George Mason University, Fairfax, Virginia 22030.
‡ Department of Mathematics, Georgetown University, Washington, D.C. 20057.

$\overline{\text{co}}\mathcal{K}(X)$ will denote the space of nonempty closed bounded and convex subsets of X. If $A \subset X$ then co$A$ will denote the convex hull of A and $\overline{\text{co}}A$ the closed convex hull of A.

Let $V=\{x_i'\}$ be any countably dense subset of S'. The distance function d, defined as

$$d(A, B)=\sum_{i=1}^{\infty} \frac{1}{2^i}|x_i'(A)-x_i'(B)|$$

for every pair $A, B \in \overline{\text{co}}\mathcal{K}(X)$, is a metric on $\overline{\text{co}}\mathcal{K}(X)$. If $X=R^n$ it is not difficult to show that d is equivalent to the more common Hausdorff metric.

T will denote a locally compact Polish space and μ a nonatomic positive regular Borel measure on T such that $\mu(T)<\infty$. When $T=[a, b]$ we will use Lebesgue measure. Integrals of Banach-valued functions which occur are to be considered as Bochner integrals. The symbol $L(T, R)$ will denote the equivalence classes of functions from T into R which are μ-integrable.

A set-valued mapping P is said to be measurable (see [6]) if for each closed subset $A \subset X$ the set $P^{-}A=\{t \in T: P(t) \cap A \neq \varnothing\}$ is measurable. P is said to be integrably bounded if there exists $g \in L(T, R)$ such that $\sup\{|x|: x \in P(t)\} \leqq g(t)$ a.e. on $T\}$. For any measurable set $A \subset T$, the set-valued integral is defined as

$$\int_A P(t)\, d\mu(t)=\Big\{\int_A \sigma(t)\, d\mu(t): \sigma: T \to X \text{ is measurable and } \sigma(t) \in P(t) \text{ a.e. on } A\Big\}.$$

If $P: T \to \overline{\text{co}}\mathcal{K}(X)$ is measurable and integrably bounded then $\int_A P(t)\, d\mu(t) \in \overline{\text{co}}\mathcal{K}(X)$ for any measurable set A (see [13]).

3. Differentiation. We will begin this section with the definition of differentiability given by Hukuhara [23].

DEFINITION 3.1. A set-valued mapping P of an interval $[a, b] \subset R$ into $\overline{\text{co}}\mathcal{K}(R^n)$ is said to be *Hukuhara differentiable* at $t_0 \in [a, b]$ if there exists $D_H P(t_0) \in \overline{\text{co}}\mathcal{K}(R^n)$ such that both of the limits

$$\lim_{h \to 0^+} \frac{P(t_0+h)-P(t_0)}{h}, \qquad \lim_{h \to 0^+} \frac{P(t_0)-P(t_0-h)}{h}$$

exist and are equal to $D_H P(t_0)$. Here the limit is taken in the Hausdorff sense and the difference $B-C$ of two sets $B, C \in \overline{\text{co}}\mathcal{K}(X)$ is the set $D \in \overline{\text{co}}\mathcal{K}(X)$, if it exists, such that $D+C=B$.

A comparison of Hukuhara's work with other theories of set-valued differentiation may be found in [4]. Through the use of the support functionals we now extend the above definition.

DEFINITION 3.2. A set-valued mapping $P: [a, b] \to \overline{\text{co}}\mathcal{K}(X)$ is said to be *differentiable* at $t_0 \in [a, b]$ if for each $x' \in S'$, $(d/dt)x'(P(t))$ exists at t_0 and equals $q(x', t_0)$ where $p(x')=q(x', t_0)$ is a continuous positively homogeneous sublinear functional on S'. In this case we define $DP(t_0)$ as

$$DP(t_0)=\bigcap_{x' \in S'} \{x: x'(x) \leqq q(x', t_0)\}.$$

LEMMA 3.3. *If* $P: [a, b] \to \overline{\text{co}}\mathcal{K}(X)$ *is differentiable at* $t_0 \in [a, b]$ *then* $DP(t_0)$ *is a nonempty closed bounded and convex set in* X.

Proof. See Lemma 3 in [12].

It is now straightforward to verify that properties holding for differentiable scalar-valued functions are also valid for differentiable set-valued mappings. We illustrate with Theorem 3.5. Proposition 3.6 has already been proved in different settings by Artstein [1], Banks and Jacobs [4], Bridgland [5] and Hermes [21]. It is included here to demonstrate the convenience of the metric d defined earlier.

LEMMA 3.4. *If* $C, D \in \overline{\text{co}}\mathcal{K}(X)$ *then* $D - C$ *exists if and only if* $p(x') = x'(D) - x'(C)$ *is a continuous positively homogeneous sublinear functional on* X'. *In that case,*

$$D - C = \bigcap_{x' \in S'} \{x : x'(x) \leqq x'(D) - x'(C)\}.$$

Proof. (i) If $D - C$ exists then clearly $p(x') = x'(D) - x'(C) = x'(D - C)$ is a continuous positively homogeneous sublinear functional on X'.

(ii) Conversely by Lemma 3 in [12] and its corollary,

$$K = \bigcap_{x' \in S'} \{x : x'(x) \leqq x'(D) - x'(C)\}$$

is in $\overline{\text{co}}\mathcal{K}(X)$ and $x'(K) = x'(D) - x'(C)$ for all $x' \in S'$. Since X is reflexive $x'(K) + x'(C) = x'(K + C)$. Thus we have $D = K + C$.

THEOREM 3.5. *If* $P: [a, b] \to \overline{\text{co}}\,\mathcal{K}(X)$ *is differentiable on* $[a, b]$ *and if* $DP(t)$ *is Borel measurable and integrably bounded on* $[a, b]$ *then for* $t \in [a, b]$

$$\int_a^t DP(s)\, ds = P(t) - P(a).$$

Proof. From the definition of $DP(t)$ we have for each $x' \in S'$, $x'(DP(t)) = (d/dt)x'(P(t))$ on $[a, b]$. Let $x' \in S'$ and $t \in [a, b]$. Since $\int_a^t (d/ds)x'(P(s))\, ds = x'(P(t)) - x'(P(a))$ it follows from the above lemma that $P(t) - P(a)$ exists. The theorem is now a consequence of Corollary 2 in [12, p. 234].

PROPOSITION 3.6. *If* $P: T = [a, b] \to \overline{\text{co}}\mathcal{K}(X)$ *is Borel measurable and integrably bounded by* $g \in L(T, R)$ *then*

$$d\left(\frac{\int_t^{t+h} P(s)\, ds}{h}, P(t)\right) \to 0 \quad as \quad h \to 0^+ \ a.e.\ in\ T.$$

Proof. Choose $V = \{x_i'\}_{i=1}^{\infty}$ dense in S'. Let $t \in T$ and $h > 0$. Then for any integer $N \geqq 1$ the following estimate can be obtained:

$$\sum_{i=N}^{\infty} \frac{1}{2^i} \left| x_i'\left(\frac{\int_t^{t+h} P(s)\, ds}{h}\right) - x_i'(P(t)) \right| \leqq \frac{1}{2^{N-2}} g(t).$$

The result now follows by referring to the real-valued case.

4. A Radon–Nikodym theorem. As in [17], a weak set-valued measure Ω on T is defined to be a set-valued mapping defined on the Borel subsets of T and taking values in $\overline{\text{co}}\mathcal{K}(X)$ such that for each $x' \in X'$, $x'(\Omega(\cdot))$ is a measure on T. If Ω

is integrably bounded, that is, $\|\Omega(E)\| = \sup\{|x|: x \in \Omega(E)\} \leqq \int_E g(t)\, d\mu(t)$ for all measurable subsets E of T and some $g \in L(T, R)$, then it is shown that Ω possesses a set-valued Radon–Nikodym derivative.

Artstein [2] and Debreu and Schmeidler [15] investigated Radon–Nikodym derivatives for set-valued measures in R^n. In [17] a related result is shown for set-valued measures in a separably locally convex topological vector space. Costé [9] using the Hausdorff set metric to define a strong set-valued measure gives a necessary and sufficient condition for the existence of Radon–Nikodym derivatives. The technique used in the proof of the following theorem is the same as in Lemma 3 of [12].

THEOREM 4.1. *If Ω is an integrably bounded weak set-valued measure defined on T with values in $\overline{\text{co}}\mathcal{K}(X)$, then there is an integrably bounded measurable set-valued mapping P taking values a.e. in $\overline{\text{co}}\mathcal{K}(X)$ such that $\Omega(E) = \int_E P(t)\, d\mu(t)$ for all measurable subsets E of T.*

Proof. For each $x' \in X'$, $x'(\Omega(\cdot)) \ll \mu$ and hence there exists $f(x', s) \in L(T, R)$ such that for every measurable set E, $x'(\Omega(E)) = \int_E f(x', t)\, d\mu(t)$. Also for each $x' \in X'$, $x'(\Omega(\cdot))$ has a Hahn decomposition; that is, there exist measurable sets A and B such that

$$\int_E f^+(x', t)\, d\mu(t) = x'(\Omega(E \cap A)) = |x'(\Omega(E \cap A))|$$
$$\leqq |x'| \int_{E \cap A} g(t)\, d\mu(t)$$

and

$$\int_E f^-(x', t)\, d\mu(t) = -x'(\Omega(E \cap B)) = |x'(\Omega(E \cap B))|$$
$$\leqq |x'| \int_{E \cap B} g(t)\, d\mu(t)$$

where E is any measurable set. Thus for each $x' \in X'$,

$$f(x', s) \leqq |f(x', s)| \leqq |x'| g(s) \quad \text{a.e. in } T.$$

Because $p: x' \to x'(\Omega(E))$ is a subadditive functional we have for $x', y' \in X'$, $f(x'+y', s) \leqq f(x', s) + f(y', s)$ for $s \in E$, where $\mu(E) = \mu(T)$. Since p is also positively homogeneous it follows that for $x' \in X'$ and $\alpha \in R^+$, $f(\alpha x', s) = \alpha f(x', s)$ for $s \in E_\alpha$ where $\mu(E_\alpha) = \mu(T)$.

Choose $V = \{x_i'\}_{i=1}^{\infty}$ dense in S' such that if $x_i' \in V$ then $-x_i' \in V$. Notice that by the above there exists a set $E \subset T$ where $\mu(E) = \mu(T)$ and such that:

(i) If α is a positive rational then $f(\alpha x_i', s) = \alpha f(x_i', s)$ for all $s \in E$ and $x_i' \in V$;

(ii) For any finite subset J of the natural numbers

$$f\left(\sum_{i \in J} \alpha_i x_i', s\right) \leqq \sum_{i \in J} \alpha_i f(x_i', s)$$

for all $s \in E$, where $x_i' \in V$ and α_i is a positive rational for each i in J.

Form the vector space X_0' over the field Q of rationals of all finite linear combinations of elements from V. It now follows that there exists a measurable set E_0 where $\mu(E_0) = \mu(T)$ and such that if $s \in E_0$ then $p(x') = f(x', s)$ is sublinear and positively homogeneous (Q^+) on X_0'. Moreover we may assume that $|f(x', s)| \leqq g(s)|x'|$ for all $s \in E_0$ and $x' \in X_0'$.

Let $s \in E_0$ and choose $x_0' \in V$. As in [12] define the subspace $M = \{\alpha x_0' : \alpha \in Q\}$ and a continuous linear functional x_0'' on M by the relation $x_0''(\alpha x_0') = \alpha f(x_0', s)$. Then $x_0''(\alpha x_0') \leqq f(\alpha x_0', s)$ for $\alpha \in Q$. By the Hahn–Banach theorem, x_0'' can be extended to a continuous linear functional x'' on X_0' such that $x''(x') \leqq f(x', s)$ for all $x' \in X_0'$ and $x''(x_0') = f(x_0', s)$. Hence $x''(x') \leqq g(s)|x'|$ for each $x' \in X_0'$, and we can again by the Hahn–Banach theorem extend x'' to all of X' over Q, giving $x''(x') \leqq g(s)|x'|$ for all $x' \in X'$. Now x'' has a unique extension to X' over R and since X is reflexive there exists $x_0 \in X$ such that $x''(x') = x'(x_0)$ for all $x' \in X'$. Thus $x_i'(x_0) \leqq f(x_i', s)$ for all $x_i' \in V$ and $x_0'(x_0) = f(x_0', s)$. Hence

$$P(s) = \bigcap_{x_i' \in V} \{x : x_i'(x) \leqq f(x_i', s)\} \neq \varnothing.$$

Also, $P(s) \in \overline{\text{co}}\mathcal{K}(X)$ for each $s \in E_0$ and P is measurable and integrably bounded. In addition, $f(x_i', s) = x_i'(P(s))$ for all $x_i' \in V$ and $s \in E_0$. Integrating over a measurable set E yields

$$\int_E f(x_i', s)\, d\mu(s) = \int_E x_i'(P(s))\, d\mu(s) = x_i'(\Omega(E))$$

and thus the proof is complete.

5. Applications. In this section some of the techniques developed concerning differentiation are applied to certain set-valued mappings arising in control theory.

Example 5.1. Consider the control of the undamped harmonic oscillator,

$$\ddot{x} + x = u, \qquad |u| \leqq 1. \tag{5.1}$$

Equivalently,

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -x + u, \end{aligned} \qquad |u| \leqq 1. \tag{5.2}$$

Then if $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, we are considering the control system

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad |u| \leqq 1. \tag{5.3}$$

Direct computation yields that the fundamental matrix solution of the homogeneous equation $\dot{x}(t) = Ax(t)$ equals

$$X(t) = e^{tA} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}$$

and thus the reachable set of (5.3) (see [8]) is

$$R(t) = \int_0^t X^{-1}(s)BU\, ds = \int_0^t \begin{bmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} U\, ds = \int_0^t \begin{bmatrix} -\sin s \\ \cos s \end{bmatrix} U\, ds.$$

Let $x' \in R^2$ and $x' = (\cos \beta, \sin \beta)$ where $0 \leqq \beta \leqq 2\pi$. Thus

$$x'\left\{\int_0^t \begin{bmatrix} -\sin s \\ \cos s \end{bmatrix} u(s)\, ds\right\} = \left\{x', \int_0^t \begin{bmatrix} -\sin s \\ \cos s \end{bmatrix} u(s)\, ds\right\}$$

$$= \int_0^t \sin(\beta - s) u(s)\, ds$$

for $u(s) \in U$. Choose $u^*(s) = \operatorname{sgn}(\sin(\beta - s))$. Then

$$x'\left(\int_0^t e^{-sA} BU\, ds\right) = \int_0^t \sin(\beta - s) u^*(s)\, ds = \int_0^t |\sin(\beta - s)|\, ds$$

and

$$\frac{d}{dt} x'\left(\int_0^t e^{-sA} BU\, ds\right) = |\sin(\beta - t)|.$$

Hence,

$$DR(t) = \bigcap_{\beta} \left\{x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : \cos \beta x_1 + \sin \beta x_2 \leqq |\sin(\beta - t)|\right\}.$$

When $t = 0$ the set-valued derivative of $R(t)$ is the line segment in R^2 connecting $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ -1 \end{bmatrix}$. As t increases the segment is rotated in a counterclockwise manner through an angle t.

The following is an infinite dimensional version of a result obtained by Chukwu [8] and by Hautus and Olsder [20] concerning identification problems for autonomous linear control systems.

Let $\mathcal{L}(X, X)$ denote the space of bounded linear operators on X.

THEOREM 5.2. *Suppose that for each $t \in [0, T]$, $F_i(t)$ and $F_i^{-1}(t)$ $(i = 1, 2)$ are one-to-one, in $\mathcal{L}(X, X)$ and satisfy $F_1(0) = F_2(0) = I$. Further assume that both $F_i(t)$ and $F_i^{-1}(t)$ $(i = 1, 2)$ are strongly continuous on $[0, T]$ and that $|F_1(t)|, |F_2(t)| \leqq M$ on $[0, T]$ for some $M > 0$. If V_i $(i = 1, 2)$ is a compact convex subset of X having at most countably many extreme points, $V_i^e (i = 1, 2)$, then $\int_0^t F_1(s) V_1\, ds = \int_0^t F_2(s) V_2\, ds$ on $[0, T]$ if and only if $V_1 = V_2$ and for each $v \in \operatorname{span} V_1^e$ $F_1(t)v = F_2(t)v$ on $[0, T]$.*

Proof. Since V_i $(i = 1, 2)$ is compact and F_i $(i = 1, 2)$ is strongly continuous it is not difficult to show that given $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\text{if} \quad |t - \bar{t}| < \delta \quad \text{then} \quad |F_i(t)v_i - F_i(\bar{t})v_i| < \varepsilon$$

for arbitrary $v_i \in V_i$ $(i = 1, 2)$. Thus the function $t \to F_i(t) V_i$ $(i = 1, 2)$ is continuous (Hausdorff metric) on $[0, T]$. Now assume $\int_0^t F_1(s) V_1\, ds = \int_0^t F_2(s) V_2\, ds$ on $[0, T]$. Taking set-valued derivatives of both sides we have that $F_1(t) V_1 = F_2(t) V_2$ a.e. in $[0, T]$. But the continuity implies that this relation holds everywhere. Substituting $t = 0$ gives the equality of V_1 and V_2. Because of the one-to-one assumption we have

$$F_2^{-1}(t) F_1(t) V_1^e = V_1^e \quad \text{on } [0, T],$$

and therefore given any point $v \in V_1^e$ there exists $v_1(t) \in V_1^e$ for $t \in [0, T]$ such that $v_1(t) = F_1^{-1}(t) F_2(t) v$ on $[0, T]$. Clearly $v_1([0, T]) = \{y_i\}_{i=1}^{\infty} = Y$ is a connected space

and $\{y_i\}$ is a closed set for each i. Applying the Baire category theorem we obtain the existence of a positive integer i_0 such that $\{y_{i_0}\}$ is a nonempty open set. Thus $Y=\{y_{i_0}\}=\{v\}$ and since v was arbitrary the result follows.

The following is an example of an infinite dimensional set-valued mapping arising in control theory.

Example 5.3. Consider the scalar partial differential equation

$$u_{tt}=u_{xx}+f(x,t), \qquad 0\leqq x\leqq 1, \quad t\geqq 0, \tag{5.4}$$

with initial values

$$\begin{aligned} u(x,0)&=\sum_{n=1}^{\infty} a_n \sin n\pi x=h(x), && 0\leqq x\leqq 1,\\ u_t(x,0)&=\sum_{n=1}^{\infty} b_n \sin n\pi x=g(x), && 0\leqq x\leqq 1, \end{aligned} \tag{5.5}$$

with

$$\sum_{n=1}^{\infty}(n\pi)^2a_n^2+\sum_{n=1}^{\infty}b_n^2<\infty$$

and boundary values

$$u(0,t)=u(1,t)=0, \qquad t\geqq 0. \tag{5.6}$$

Here $f(x,t)$ is assumed to be of the form

$$f(x,t)=\sum_{n=1}^{\infty} f_n(t)\sin n\pi x \tag{5.7}$$

where each f_n is integrable and $\sum_{n=1}^{\infty}|f_n(t)|^2<\infty$ for t a.e. in $[0,\infty)$. The solution of (5.4)–(5.7) can be written in the form

$$u(x,t)= \tag{5.8}$$
$$\sum_{n=1}^{\infty}\left\{a_n\cos n\pi t+\frac{b_n\sin n\pi t}{n\pi}+\frac{1}{n\pi}\int_0^t[\sin n\pi(t-s)]f_n(s)\,ds\right\}\sin n\pi x.$$

For each $t\in[0,\infty)$, $u(\cdot,t)$ and $u_t(\cdot,t)\in L_2[0,1]$.

Consider the complete set $\{\sin n\pi x\}$ in $L_2[0,1]$ and define the one-to-one onto mapping i of $L_2[0,1]$ onto l_2 as follows

$$i\,(\sin n\pi x)=e_n$$

where

$$e_1=(1,0,0,\cdots)$$
$$e_2=(0,1,0,\cdots)$$

Then from (5.8),

$$i(u(\cdot, t)) = \left\{ a_n \cos n\pi t + \frac{b_n \sin n\pi t}{n\pi} + \frac{1}{n\pi}\int_0^t \sin n\pi(t-s) f_n(s)\, ds \right\}$$
$$= \hat{u}(t, h, g, f).$$

For each $t \in [0, \infty)$ the nth coordinate of $\hat{u}(t, h, g, f)$ satisfies the ordinary differential equation

$$\frac{d^2 x_n}{dt^2} + (n\pi)^2 x_n = f_n$$

or letting

$$\hat{x}_n = \begin{bmatrix} x_n \\ \dfrac{dx_n}{dt} \end{bmatrix}$$

we have

$$\frac{d\hat{x}_n}{dt} = \begin{bmatrix} 0 & 1 \\ -(n\pi)^2 & 0 \end{bmatrix} \hat{x}_n + \begin{bmatrix} 0 \\ 1 \end{bmatrix} f_n. \tag{5.9}$$

Thus we can reduce the study of solutions of (5.4)–(5.7) to a study of an infinite number of solutions of the forced harmonic oscillator equation (5.9).

Consider the representation of $\hat{u}(t, h, g, f)$ and take the Hilbert cube I^∞ as the control set U. We shall look at all measurable mappings

$$f(s) = (f_1(s), \cdots, f_n(s), \cdots)$$

taking values a.e. in I^∞. Let $S(t-s) = \{(1/n\pi) \sin [n\pi(t-s)]\}$ and define the set-valued mapping $R(t)$ as

$$R(t) = \int_0^t S(t-s) I^\infty \, ds.$$

Taking the set-valued derivative of R we obtain

$$DR(t) = S(t) I^\infty = \left\{ \frac{1}{n\pi} \sin n\pi t \, a_n : |a_n| \leqq \frac{1}{n} \right\}.$$

Example 5.4. If we phrase the above example in a slightly different context and change our control set to a compact one with a countable number of extreme points we shall have an example of a system which satisfies Theorem 5.2. Thus consider the real Hilbert space H which consists of a countable direct sum of R^2. That is points $\hat{x}$ such that $\hat{x} = (\hat{x}_n, n = 1, 2, \cdots)$ where $\hat{x}_n = \begin{bmatrix} x_n^1 \\ x_n^2 \end{bmatrix}$ and

$$|\hat{x}| = \left[\sum_{j=1}^{\infty} j^2 (x_j^1)^2 + \sum_{j=1}^{\infty} (x_j^2)^2 \right]^{1/2} < \infty.$$

In place of (5.9) consider the equation

$$\frac{d\hat{x}_n}{dt} = \begin{bmatrix} 0 & 1 \\ -(n\pi)^2 & 0 \end{bmatrix} \hat{x}_n + \begin{bmatrix} b^n_{11} & b^n_{12} \cdots b^n_{1m} \\ & \\ b^n_{21} & b^n_{22} \cdots b^n_{2m} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \tag{5.10}$$

where we assume $|u_j| \leqq 1$, $j = 1, 2, \cdots, m$ and

$$\sum_{n=1}^{\infty} \left[\sum_{j=1}^{m} b^n_{1j} \right]^2 n^2 + \sum_{n=1}^{\infty} \left[\sum_{j=1}^{m} b^n_{2j} \right]^2 < \infty.$$

The homogeneous solutions of (5.10) (i.e. $(u_j(t)) = 0, j = 1, 2, \cdots, m$) give rise to a strongly continuous group $F(t)$ defined by the expression

$$F(t)\hat{x} = \left[\begin{pmatrix} \cos n\pi t & \dfrac{\sin n\pi t}{n\pi} \\ -n\pi \sin n\pi t & \cos n\pi t \end{pmatrix} \begin{pmatrix} x^1_n \\ x^2_n \end{pmatrix}, n = 1, 2, \cdots \right]$$

where $\hat{x}$ is in H. The control set U is

$$U = \{u \in R^m : |u_j| \leqq 1, j = 1, 2, \cdots, m\}.$$

Notice that if we consider the reachable set of the system defined above we are able to obtain certain uniqueness results.

As a final application we shall consider a variant of a control problem studied by Neustadt [27].

Example 5.5. Let $T(t)$ be a strongly continuous semi-group of class C_0 defined on a separable reflexive Banach space X. Let U be a compact set in R^m and $f: R^m \to X$ a continuous mapping. Define the set-valued mapping

$$F(t) = T(t)f(U).$$

Since $f(U)$ is compact and $T(t)$ is strongly continuous, $F(t)$ can be easily shown to be upper continuous and hence measurable (see [6]). Furthermore over finite intervals, $F(t)$ is integrably bounded. This is due to the fact that $T(t)$ is a strongly continuous C_0 semi-group and hence we can obtain estimates of the form

$$\sup \{|x| : x \in F(t)\} \leqq Me^{\omega t} \sup \{|y| : y \in f(U)\}$$

where ω can be assumed positive and $M \geqq 1$.

If $X = R^n$ and we look at the set-valued mapping

$$R(t) = \int_0^t T(t-s)f(U)\, ds = \int_0^t T(s)f(U)\, ds$$

we are in fact studying a finite dimensional control problem of the form

$$\dot{x}(t) = Ax(t) + f(u(t))$$

where A is an $n \times n$ matrix which satisfies $(d/dt)T(t) = AT(t)$ and u is a measurable mapping from $R^m \to R^n$ with values in U. By the Lyapunov theorem [25], $R(t)$

is for each t a compact convex set in R^n and

$$R(t) = \int_0^t T(s)f(U)\,ds = \int_0^t \overline{\text{co}}\, T(s)f(U)\,ds.$$

Thus since $f(U)$ is compact the following holds:

$$DR(t) = \overline{\text{co}}\, T(t)f(U) = \text{co}\, T(t)f(U) = T(t)\,\text{co}\, f(U).$$

In the case where X is infinite dimensional the most that we can conclude is that $\text{cl}\,(R(t)) = \int_0^t \overline{\text{co}}\,(T(s)f(U))\,ds$. Hence by our theory and the Krein–Milman theorem [16] we can conclude that

$$D(\text{cl}\,(R(t))) = \overline{\text{co}}\, T(t)f(U) = T(t)\,\overline{\text{co}}\, f(U) \quad \text{a.e.}$$

(Here $\text{cl}\,(A)$ denotes the closure of A in the normed topology.) Problems of the type mentioned above can arise from distributed parameter problems of which the following is a simple example.

In Example 5.3 we replace $f(x, t)$ in equation (5.4) by $f(x, u)$ where

$$f(x, u) = \sum_{n=1}^{\infty} a_n(u) \sin n\pi x$$

and $u \in U$, some compact subset of R.

REFERENCES

[1] Z. ARTSTEIN, *On the calculus of closed set-valued functions*, Indiana Univ. Math. J., 24 (1974), pp. 433–441.

[2] ———, *Set-valued measures*, Trans. Amer. Math. Soc., 165 (1972), pp. 103–125.

[3] R. J. AUMANN, *Integrals of set-valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1–12.

[4] H. T. BANKS AND M. Q. JACOBS, *A differential calculus for multifunctions*, Ibid., 29 (1970), pp. 246–272.

[5] T. F. BRIDGLAND, *Trajectory integrals of set-valued functions*, Pacific J. Math., 33 (1970), pp. 43–68.

[6] C. CASTAING, *Sur les multi-applications mesurables*, Doctoral thesis, L'Universite de Caen, Caen, France, 1967.

[7] ———, *Le théorème de Dunford–Pettis généralisé*, C. R. Acad. Sci. Paris Sér. A, 268 (1969), pp. 327–329.

[8] E. CHUKWU, *Symmetries of autonomous linear control systems*, this Journal, 12 (1974), pp. 436–448.

[9] A. COSTÉ, *La propriété de Radon–Nikodym en intégration multivoque*, C. R. Acad. Sci. Paris Sér. A, 280 (1975), pp. 1515–1518.

[10] ———, *Sur les multimesures à valeurs fermées bornées d'un espace de Banach*, Ibid., 280 (1975), pp. 567–570.

[11] A. COSTÉ AND R. PALLU DE LA BARRIÈRE, *Un théorème de Radon–Nikodym pour les multimesures à valeurs convexes fermées localement compactes sans droite*, Ibid., 280 (1975), pp. 255–258.

[12] R. DATKO, *Measurability properties of set-valued mappings in a Banach space*, this Journal, 8 (1970), pp. 226–238.

[13] ———, *On the integration of set-valued mappings in a Banach space*, Fund. Math., 78 (1973), pp. 205–208.

[14] G. DEBREU, *Integration of correspondences*, Proc. 5th Berkeley Symp. Math. Stat. and Prob., Vol. 2, 1967, pp. 351–372. University of California Press, Berkeley, 1967.

[15] G. DEBREU AND D. SCHMEIDLER, *The Radon–Nikodym derivative of a correspondence*, Proc. 6th Berkeley Symp. Math. Stat. and Prob., Vol. 2, 1972, pp. 41–56.
[16] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, John Wiley, New York, 1958.
[17] C. GODET-THOBIE, *Sélections de multimesures. Application à un théorème de Radon–Nikodym multivoque*, C. R. Acad. Sci. Paris Sér. A, 279 (1974), pp. 603–606.
[18] O. HÁJEK, *Identification of control systems by performance*, Math. Systems Theory, 5 (1971), pp. 349–352.
[19] P. R. HALMOS, *Measure Theory*, Van Nostrand, Princeton, NJ, 1950.
[20] M. L. J. HAUTUS AND G. J. OLSDER, *A uniqueness theorem for linear control systems with coinciding reachable sets*, this Journal, 11 (1973), pp. 412–416.
[21] H. HERMES, *Calculus of set-valued functions and control*, J. Math. Mech., 18 (1968), pp. 47–59.
[22] L. HÖRMANDER, *Sur la fonction d'appui des ensembles convexes dans ún espace localement convexe*, Ark. Mat., 3 (1954), pp. 181–186.
[23] M. HUKUHARA, *Intégration des applications mesurables dont la valeur est un compact convexe*, Funkcial. Ekvac. 10 (1967), pp. 205–223.
[24] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[25] A. LYAPUNOV, *Sur les fonctions-vecteurs complètement additives*, Bull. Acad. Sci. URSS Sér. Math., 4 (1940), pp. 465–478.
[26] H. METHLOUTHI AND R. PALLU DE LA BARRIÈRE, *Multi-mesures et champs d'espaces de Banach*, C. R. Acad. Sci. Paris Sér. A, 281 (1975), pp. 1027–1030.
[27] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
[28] H. RADSTRÖM, *An embedding theorem for spaces of convex sets*, Proc. Amer. Math. Soc., 3 (1952), pp. 165–169.
[29] H. L. ROYDEN, *Real Analysis*, 2nd ed., Macmillan, London, 1968.
[30] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

INFORMATIONALLY NONUNIQUE EQUILIBRIUM SOLUTIONS IN DIFFERENTIAL GAMES*

TAMER BAŞAR†

Abstract. This paper is concerned with two-person deterministic nonzero-sum differential games (NZSDG) characterized by quadratic objective functionals and with state dynamics described by linear differential equations. It is first shown that such games admit uncountably many (informationally nonunique) noncooperative (Nash) equilibrium solutions when at least one of the players has access to dynamic information. We provide a characterization of all Nash equilibrium solutions to the problem for a particular dynamic information pattern, and propose an optimal unique selection of an element of the Nash equilibrium set, which exhibits a robust behavior by being insensitive to additive random perturbations in the state dynamics. We model these random perturbations as a local martingale process and obtain the abovementioned optimal Nash strategy pair as the unique noncooperative equilibrium solution of a related stochastic NZSDG. With regard to the latter, it is shown that the unique Nash equilibrium strategy of the player with dynamic closed-loop information can be realized by affine control laws.

1. Introduction. A significant portion of the research activities concerning nonzero-sum differential games (NZSDG) has been concentrated on linear-quadratic problems (i.e. differential games characterized by quadratic cost functionals, and with state dynamics described by linear differential equations) since (i) these problems are more tractable and admit noncooperative equilibrium solutions that can be expressed in closed form, (ii) a thorough treatment of this class of problems and investigation of the properties of their permissible equilibrium solutions help to gain insight into the solutions of structurally more complicated NZSDG.

However, even though it constitutes, in some sense, the simplest class of NZSDG to deal with, the LQNZSDG theory is by no means complete today. In particular, there are no general enough results in the literature for problems characterized by dynamic information patterns. In obtaining the noncooperative (Nash) equilibrium solutions to these problems, it has been common practice to start with specific functional forms for the strategies of the players and then to derive the necessary conditions for existence of a Nash equilibrium solution within the class restricted by the a priori selected functional structure (this restriction has always resulted in the class of linear feedback control laws for the general closed-loop information structure). (See e.g. [7], [8], [11], [16], [17]). Attempts have been made in [11] and [8] to prove uniqueness of the linear feedback solution under the closed-loop information structure. However, Lukes assumes in [11] that the permissible strategies are also in pure-feedback form in addition to being linear, which is a rather severe restriction on the strategy spaces. Friedman claims in [8] that he has established uniqueness of linear Nash strategies under the general closed-loop information structure; however he has made an explicit assumption on the functional form of the optimal return function (see [8, p. 318]) which places a similar a priori restriction on the strategy spaces. To sum up: until

* Received by the editors February 28, 1975, and in revised form October 7, 1976.

† Applied Mathematics Division, Marmara Scientific and Industrial Research Institute, Gebze-Kocaeli, Turkey.

very recently, there were no existence and uniqueness results in the literature concerning LQNZSDG with dynamic information structures.

The first systematic and constructive approach to problems of this kind has been taken in [1] for the class of two-person LQNZSDG in which the state dynamics are described by a difference equation instead of a differential equation. It has been shown in [1] that when at least one of the players has access to dynamic information, then such differential games admit uncountably many Nash equilibrium solutions. Moreover, it has been shown in [2] that the solution can also be nonlinear. In order to resolve the dilemma arising from existence of nonunique Nash solutions, it has been proposed in [1] to further restrict the Nash solution concept for a given information pattern by including a zero-mean additive random perturbation term in the state dynamics and require the solution to be insensitive to this additive noise. Within the context of two-person LQNZSDG described by difference equations and for different information structures for either player, existence of unique Nash equilibrium strategies which exhibit this robust behavior has then been established in [1]. These results have further been extended to similar M-person differential games ($M>2$) in [3], for the case when a subset of the players have access to closed-loop information and the rest to open-loop information.

The results reported in [1], [2] and [3] hinted that an analogous structural behavior might be seen in LQNZSDG described by differential equations. A rigorous investigation of the validity of this conjecture, however, requires a mathematical approach different from the one taken in [1] and [3]. Hence, this paper constitutes the first systematic treatment of LQNZSDG described by differential equations and has a twofold objective: 1) To put into perspective the difficulties encountered in obtaining the entire class of noncooperative equilibrium solutions of an LQNZSDG for a given dynamic information pattern, and to display the fact that a direct extension of single criterion optimization techniques and dynamic programming does not suffice to characterize all such solutions; 2) to propose an *optimal unique* selection from the Nash strategy set composed of uncountably many elements, and to obtain the corresponding solution in closed form when one of the players has access to closed-loop information and the other one to open-loop information.

In the next section, we formulate the two-person LQNZSDG under a particular information structure which provides the first player with the classical closed-loop information and the second player with open-loop information. In § 3 we obtain a characterization of the entire class of Nash equilibrium solutions to the problem, elaborate on the *informational nonuniqueness* of the Nash equilibrium point, and then provide an illustrative example. In § 4 we relate the process of the optimal unique selection from the Nash equilibrium set to the derivation of the unique noncooperative solution of a particular stochastic differential game. A similar type of stochastic differential game has previously been considered by Friedman in [9] for the case when both players have access to closed-loop information. However, Friedman further restricts the strategies of every player to be functions of only the current value of the state vector. Here, we make no such assumptions and prove existence of a unique equilibrium solution under the information pattern of § 2.

The paper ends with a conclusion section and three appendices.

2. Formulation of deterministic LQNZSDG. The differential game under consideration is a two-person, fixed-duration ($[t_0, t_f]$) NZSDG, described by linear state dynamics

$$\dot{x} = F(t)x + G_1(t)u_1(t) + G_2(t)u_2(t); \qquad x(t_0) = x_0 \tag{1}$$

where $\dim(x) = n$, $\dim(u_i) = r_i$, the matrices F, G_1 and G_2 have appropriate dimensions and are continuous on $[t_0, t_f]$. $x(t_0)$ is the initial state vector and its value x_0 is known to both players. The functions $u_1(t)$ and $u_2(t)$ represent the control policies of players 1 and 2, respectively, and assume values in $\mathbb{R}^{r_1}$ and $\mathbb{R}^{r_2}$, respectively.

To delineate the information structure of the problem we let $C_n = C_n[t_0, t_f]$ denote the space of continuous functions on $[t_0, t_f]$ with values in $\mathbb{R}^n$. We further let $\mathcal{F}_t$ be the sigma-field in C_n generated by the cylinder sets $\{x \in C_n, x(s) \in B\}$ where B is a Borel set in $\mathbb{R}^n$ and $t_0 \leqq s \leqq t$. Then, the information gained by player 1 during the course of the game is completely determined by the information field $\mathcal{F}_t$ for all $t \geqq t_0$; i.e., player 1 has access to perfect *nonanticipative closed-loop* information concerning the state of the game. Player 2, on the other hand, gains no information during the course of the game (i.e., he plays open-loop).

Permissible strategy for player 1 will be a mapping $\gamma_1(\cdot, \cdot)$ of $[t_0, t_f] \times C_n$ into $\mathbb{R}^{r_1}$ with the following properties:

(i) $\gamma_1(t, \eta)$ is continuous in t for each $\eta \in C_n$.

(ii) It is uniformly Lipschitz in η; i.e., for some $k > 0$

$$|\gamma_1(t, \eta) - \gamma_1(t, \xi)| \leqq k\|\eta - \xi\|; \qquad t \in [t_0, t_f], \quad \eta, \xi \in C_n,$$

where $\|\cdot\|$ is the standard sup norm in C_n.

(iii) $u_1(t) = \gamma_1(t, x)$ is adapted to the information field $\mathcal{F}_t$; i.e., it is $\mathcal{F}_t$-measurable.

We denote the class of strategies described above by U_1, to be referred to as the *permissible strategy set for player* 1. Since player 2 has access to open-loop information, we let the permissible strategy set U_2 for player 2 be C_{r_2}.

Since the strategy set U_1 does not only contain Markovian (pure-feedback) controls (i.e., controls that depend only on the current value of the state vector), equation (1) is actually a functional differential equation rather than an ordinary differential equation, which should better be written as

$$dx/dt = F(t)x + G_1(t)\gamma_1[t, x(\cdot)] + G_2(t)\gamma_2(t); \qquad x(t_0) = x_0. \tag{2}$$

It is known that this equation admits a unique continuous solution on $[t_0, t_f]$ for every pair $\{\gamma_1 \in U_1, \gamma_2 \in U_2\}$ (see, for example, [6]).

For any pair of strategies $\{\gamma_i \in U_i, i = 1, 2\}$, the loss (or minus the payoff) incurred to player i is given by the quadratic cost function

$$J_i(u_1, u_2) = x^T(t_f)C_{i_f}x(t_f) + \int_{t_o}^{t_f} [x^T C_i(t)x + u_1^T D_{i1}(t)u_1 + u_2^T D_{i2}(t)u_2]\, dt \tag{3}$$

where $u_1 = \gamma_1[\cdot, \cdot]$, $u_2 = \gamma_2(\cdot)$; the weighting matrices C_{i_f}, $C_i(t)$ are nonnegative definite for all $t \in [t_0, t_f]$ and each entry of $C_i(\cdot)$ is continuous on $[t_0, t_f]$. The matrices $D_{i1}(t) > 0$, $D_{i2}(t) > 0$ are also defined and continuous on $[t_0, t_f]$, $i = 1, 2$.

The objective of player i is to pick the permissible strategy that will yield the minimum value of the cost function J_i against some rationally selected strategy of

player $j, j \neq i$. With no direct cooperation between the players allowed, this reasoning leads to what is known as the *noncooperative Nash equilibrium solution* [13].

DEFINITION 1. A pair $\{\gamma_1^* \in U_1, \gamma_2^* \in U_2\}$ is said to be *in* (*Nash*) *equilibrium* if the following inequalities hold for all $\gamma_1 \in U_1$, $\gamma_2 \in U_2$:

$$J_1(\gamma_1^*, \gamma_2^*) \leqq J_1(\gamma_1, \gamma_2^*), \tag{4a}$$

$$J_2(\gamma_1^*, \gamma_2^*) \leqq J_2(\gamma_1^*, \gamma_2). \tag{4b}$$

If there exist more than one set of Nash equilibrium solutions, then we can define a partial ordering within this solution set as follows:

DEFINITION 2. For a given information structure, a permissible Nash pair $\{\gamma_1, \gamma_2\}$ is said to result in a *better* performance than (or be *superior* to) another permissible Nash pair $\{\beta_1, \beta_2\}$ if

$$J_i(\gamma_1, \gamma_2) \leqq J_i(\beta_1, \beta_2), \qquad i = 1, 2 \tag{5}$$

with strict inequality for at least one i. We will call a permissible Nash pair $\{\gamma_1, \gamma_2\}$ *admissible* if there exists *no other* permissible Nash solution pair that is better than $\{\gamma_1, \gamma_2\}$.

It is this author's strong opinion that a given Nash solution pair can be considered as a reasonable equilibrium solution for noncooperative nonzero-sum situations if it is also admissible. Even an admissible strategy pair might sometimes be considered unreliable if the problem admits more than one admissible equilibrium solution. As a matter of fact, it turns out that this is the case for most of the nonzero-sum differential game problems that admit nonunique Nash equilibrium solutions.

It is not only the structure of the cost functions or the differential equation involved that is responsible for the nonuniqueness of the Nash equilibrium solution, but in addition (and more commonly), the dynamic nature of the information structure plays a role in the appearance of nonunique equilibrium solutions. In particular, an NZSDG that admits a unique noncooperative equilibrium solution in open-loop policies will admit nonunique solutions when the information structure is made dynamic. We will say that this kind of nonuniqueness arises purely because of the *informational structure* of the problem.

Existence of this phenomena in NZSDG has only recently been established in the literature (see Başar [1], [2]), and it has been shown within the context of LQNZSDG described by difference equations and with dynamic information that this is a rule in such games rather than an exception.

In the next section, we will describe a technique of obtaining "informationally nonunique" Nash solutions and find a characterization of the entire class of Nash equilibrium strategies for the LQNZSDG of this section. It will be apparent from the sequel that the same technique can also be applied to nonlinear and nonquadratic deterministic NZSDG described by differential state equations, and under any dynamic information structure.

3. Nonuniqueness of the Nash solution under dynamic information. It has previously been shown by Lukes et al. [12] that the Nash equilibrium solution of the LQNZSDG of § 2, with the open-loop information structure for both players,

is unique whenever it exists, and that the solution exists when the interval $[t_0, t_f]$ is taken to be sufficiently small. We will, however, show in what follows that the same differential game admits nonunique and both linear and nonlinear permissible Nash solutions if at least one of the players (say, player 1) has nonanticipative closed-loop information.

To this end, let us first fix player 2's strategy at a $\bar{\gamma}_2 \in U_2$ and note that for this fixed control policy of player 2, player 1 is faced with a "one-sided" optimal control problem, namely

$$\min J_1(\gamma_1, \bar{\gamma}_2), \qquad \gamma_1 \in U_1 \tag{6}$$

where J_1 is given by (3) and the state vector x is defined by (2), with $u_2 = \bar{\gamma}_2(\cdot)$. It is a well-known result in optimal control theory (linear regulator) that the global minimum of J_1 exists and a candidate that yields that minimum is the Markovian (pure-feedback) strategy

$$\gamma_1^*[t, x(\cdot)] = \gamma_1^*(t, x) = -D_{11}^{-1}(t)G_1^T(t)[P(t)x + k(t)] \tag{7}$$

where $P(t)$ is obtained as a solution to a matrix-Riccati equation and *does not depend on* $\bar{\gamma}_2$, and $k(t)$ is obtained as a solution to a linear time-varying differential equation and does depend on $\bar{\gamma}_2$ (see e.g. [4]). However, (7) is not the only element of U_1 that yields the global minimum. To see this, we first introduce for each σ, $t_0 < \sigma < t_f$, an affine mapping of $\mathbb{R}^n$ onto $C_n[\sigma, t_f]$ by

$$Z_\sigma(t) = \psi(t, \sigma)z(\sigma) - \int_\sigma^t \psi(t, \tau)[G_1 D_{11}^{-1} G_1^T k(\tau) - G_2 \bar{\gamma}_2(\tau)]\, d\tau \tag{8a}$$

where $z(\sigma) \in \mathbb{R}^n$, $Z_\sigma(\cdot) \in C_n(\sigma, t_f]$, and $\psi(t, \sigma)$ is the transition matrix associated with the differential equation

$$\dot{y} = (F - G_1 D_{11}^{-1} G_1^T P)(t)y. \tag{8b}$$

If $\bar{x}(t)$, $t_0 \leqq t \leqq t_f$, denotes the unique trajectory resulting from application of the strategies γ_1^* (given by (7)) and $\bar{\gamma}_2$, then we note that with $z(\sigma)$ picked to be equal to $\bar{x}(\sigma)$, $Z_\sigma(t)$ agrees with the restriction of $\bar{x}(\cdot)$ to $[\sigma, t_f]$. Now, we let $\varphi_\sigma^t[x(\cdot), Z_\sigma(\cdot)]$ denote any element of U_1, with the additional properties:

(i) $\varphi_\sigma^t[\cdot, \cdot]$: $C_n[\sigma, t] \times C_n[\sigma, t] \to \mathbb{R}^{r_1}$ for each fixed t, σ, $t_0 \leqq \sigma \leqq t \leqq \leqq t_f$;

(ii) $\varphi_\sigma^t[\cdot, \cdot]$ satisfies the boundary condition

$$\varphi_\sigma^t[x(\cdot), Z_\sigma(\cdot)] = -D_{11}^{-1}(t)G_1^T(t)[P(t)x(t) + k(t)] \tag{9}$$

whenever $Z_\sigma(s)$ is replaced by $x(s)$, for all s, $\sigma \leqq s \leqq t$. A typical candidate would be

$$\varphi_\sigma^t[x(\cdot), Z_\sigma(\cdot)] = \int_\sigma^t [x(s) - Z_\sigma(s)]\, ds - D_{11}^{-1}(t)G_1^T(t)[P(t)x(t) + k(t)]. \tag{10}$$

It is now not difficult to see that any such φ_σ^t, with the additional side condition $z(\sigma) \equiv x(\sigma)$, achieves the global minimum of J_1, and therefore is an optimal solution to the problem for every fixed $\bar{\gamma}_2 \in U_2$. Hence, we have a nontrivial subset of U_1, to be denoted by $\mathcal{U}_1(\bar{\gamma}_2)$, which has uncountably many atoms, each of them

yielding the global minimum of J_1. $\mathcal{U}_1(\bar{\gamma}_2)$ can in fact rigorously be defined as a subset of U_1 as follows:

$$(11)\quad \mathcal{U}_1(\bar{\gamma}_2)=\{\gamma_1\in U_1:\ \gamma_1(t, Z_{t_0}(t))=-D_{11}^{-1}(t)G_1^T(t)[P(t)Z_{t_0}(t)+k(t)],\ z(t_0)=x_0\}$$

where $Z_{t_0}(t)$ is defined by (8a) and the dependence of $\bar{\gamma}_2$ is through $k(t)$ as described before. Hence, we have

OBSERVATION 1. If $\mathcal{U}_1(\bar{\gamma}_2)$ is as defined by (11), then for every fixed $\bar{\gamma}_2\in U_2$, any element of $\mathcal{U}_1(\bar{\gamma}_2)$ yields the global minimum of $J_1(\cdot, \bar{\gamma}_2)$. Furthermore, it follows from (11) that for every fixed $\bar{\gamma}_2\in U_2$, different γ_1 in $\mathcal{U}_1(\bar{\gamma}_2)$ stand for different *representations* (in terms of the state vector) of the same control value which is $\gamma_1(\cdot, Z_{t_0}(\cdot))$.

PROPOSITION 1. *For the differential game of* § 2, *every Nash strategy for player* 1 *is in* $\mathcal{U}_1(\gamma_2)$ *for at least one* $\gamma_2\in U_2$.

Proof. Assume that there exists a Nash pair $\{\bar{\gamma}_1, \bar{\gamma}_2\}$ with $\bar{\gamma}_1\notin\mathcal{U}_1(\bar{\gamma}_2)$. However, since J_1 is a strictly convex functional over U_1, there exists a unique minimizing open-loop policy for player 1, which also determines $\mathcal{U}_1(\bar{\gamma}_2)$ [see (11)]. Hence, if $\bar{\gamma}_1\notin\mathcal{U}_1(\bar{\gamma}_2)$, then $J_1(\bar{\gamma}_1, \bar{\gamma}_2)>\min_{U_1} J_1(\cdot, \bar{\gamma}_2)$, which contradicts the Nash optimality of $\{\bar{\gamma}_1, \bar{\gamma}_2\}$. Q.E.D.

In order to obtain a permissible Nash strategy for player 2, we start with an element $\bar{\gamma}_1\in\mathcal{U}_1(\gamma_2^0)$ and solve the optimization problem

$$(12a)\qquad \min_{U_2} J_2(\bar{\gamma}_1, u_2).$$

Assuming that a solution exists, it will in general also depend on γ_2^0; i.e. any candidate u_2^* will be of the form

$$(12b)\qquad u_2^*=\gamma_2(t, \gamma_2^0).$$

We now require consistency in the solution and solve for a ${}^*\gamma_2^0\in U_2$ that satisfies the fixed-point equation

$$(12c)\qquad \gamma_2^0=\gamma_2(t, \gamma_2^0).$$

The pair $\{{}^*\bar{\gamma}_1\in\mathcal{U}_1({}^*\gamma_2^0), {}^*\gamma_2^0\}$ will then constitute a Nash solution for the differential game that we have considered.

If the procedure described above is executed for every $\gamma_2\in U_2$ and every $\gamma_1\in\mathcal{U}_1(\gamma_2)$, then this provides us with a nontrivial subset of the product space $U_1\times U_2$ to be denoted by $\mathcal{U}_N$ which has uncountably many elements and every element is a pair of strategies that are Nash optimal against each other. $\mathcal{U}_N$ will be called the "Nash strategy set" of this differential game corresponding to the given information structure. We have only a partial ordering in $\mathcal{U}_N$ determined by the notion of betterness introduced in Definition 2; and a subset of $\mathcal{U}_N$ can be formed consisting of all the admissible Nash pairs. Several illustrative examples worked out by the author indicate that this subset is, in general, *not* a singleton, and its elements yield different Nash costs. It now follows from the definition of $\mathcal{U}_1(\gamma_2)$ that given any pairs (γ_1^1, γ_2^1), (γ_1^2, γ_2^2) in $\mathcal{U}_N$ a sufficient condition for $J_1(\gamma_1^1, \gamma_2^1)=J_1(\gamma_1^2, \gamma_2^2)$ and $J_2(\gamma_1^1, \gamma_2^1)=J_2(\gamma_1^2, \gamma_2^2)$ is that $\gamma_2^1(\cdot)\equiv\gamma_2^2(\cdot)$. It is a property of the

two-person linear-quadratic team problem[1] that this sufficiency condition is satisfied for all elements of $\mathcal{U}_N$, i.e.,

PROPOSITION 2. *For the LQNZSDG of* § 2, *and under the parametric restrictions* $C_{1_f} = C_{2_f}$, $C_1 \equiv C_2$, $D_{11} \equiv D_{21}$, *the corresponding Nash strategy set* $\mathcal{U}_N$ *has the property that the second components of every element of* $\mathcal{U}_N$ *are identical.*

Proof. Under the parametric restrictions given above, there exists a Nash solution to the LQNZSDG if and only if there exists a person-by-person team optimal solution for the problem described by the objective function

$$(13)\quad J(u_1, u_2) = x^T(t_f)C_{1_f}x(t_f) + \int_{t_0}^{t_f} (x^T C_1(t)x + u_1^T D_{11}(t)u_1 + u_2^T D_{22}(t)u_2)\, dt = x_0\}$$

and the state dynamics (2). Since J_1 is convex in the pair (u_1, u_2), every person-by-person team optimal solution is also globally optimal and furthermore there exists a unique open-loop solution for both players. Because of the nature of the permissible strategy set of player 1, he can pick different representations of the same open-loop control value; however, for player 2, he is permitted to use only open-loop policies and is therefore committed to a unique representation. Hence, player 2 has a unique Nash strategy. Q.E.D.

Remark. It should be noted that, for the team problem, even though player 1 can employ different representations and player 2 is faced with a different objective functional to be minimized in each case, these objective functionals all have the same global minimum yielded by the same control vector. The reason for this is that the objectives of both players are essentially the same. However, this does not hold true in NZSDG which cannot be converted to equivalent team problems (with a possible exception of 2-person NZSDG that can be converted into equivalent zero-sum differential games), since for two different representations player 2 is in general faced with two entirely different optimization problems, the global minima of which do not necessarily coincide. This reasoning brings us to the following conclusion.

CONCLUSION 1. Nonzero-sum differential games which admit a unique Nash equilibrium solution in open-loop policies, will, in general, admit uncountably many Nash equilibrium solutions ("informationally nonunique" solutions) when the information structure of at least one of the players becomes dynamic. Deterministic LQNZSDG have this property of admitting informationally nonunique solutions.

In order to illustrate this conclusion, let us now consider the following scalar LQNZSDG in which the strategy set of player 2 is taken to be the class of all piecewise-constant controls, which makes the computational aspects of the problem rather straightforward. This example will therefore allow us to focus our full concentration on the technique of obtaining informationally nonunique solutions.

Example 1. Within the framework of the formulation given in § 2, let the state dynamics be described by

$$(14a)\qquad \dot{x} = u + v, \quad x(0) = x_0, \quad t \in [0, 1], \quad u \triangleq u_1, \quad v \triangleq u_2,$$

[1] A team problem is a dynamic optimization problem with a single objective functional to be jointly minimized (or maximized) by several controllers with possibly different information.

and the cost functionals by

$$J_1 = x^2(1) + \int_0^1 u^2\,dt, \tag{14b}$$

$$J_2 = x^2(1) + \int_0^1 v^2\,dt + \beta \int_0^1 u^2\,dt, \qquad \beta \geqq 0. \tag{14c}$$

The control spaces are $\mathbb{R}$, and the information structure of the problem is as given before. The permissible strategy set for player 2 is taken to be a proper subset of U_2, which consists of all constant maps, whereas the permissible strategy set for player 1 is as described before in §2.

Now, for every fixed $\gamma_2(\cdot) = \bar{v} \in \mathbb{R}$, minimization of J_1 over U_1 yields the solution

$$u = \gamma_1(t, x) = -(Sx + k), \tag{15a}$$

$$S(t) = 1/(2-t), \tag{15b}$$

$$k(t) = [(1-t)/(2-t)]\bar{v} \tag{15c}$$

which is unique in value but not in representation. It follows from Observation 1 that any strategy of the form (16) is permissible Nash against $\gamma_2(\cdot) \equiv \bar{v}$ for any measurable $p(\cdot)$ and $q(\cdot)$:

$$\gamma_1(t, x) = -(Sx + k) + \int_0^t [x(\sigma) - \bar{x}(\sigma)]p(\sigma)\,d\sigma + [x(t) - \bar{x}(t)]q(t) \tag{16a}$$

$$\dot{\bar{x}} = -S\bar{x} - k + \bar{v}; \qquad \bar{x}(0) = x_0, \tag{16b}$$

or in more compact form with the use of (15b) and (15c):

$$\gamma_1(t, x) = \left(q - \frac{1}{2-t}\right)x - \left(\frac{qt}{2} + \frac{1-t}{2-t}\right)\bar{v} - q\left(1 - \frac{t}{2}\right)x_0 + y(t), \tag{16c}$$

$$\dot{y} = xp - (1 - t/2)px_0 - (tp/2)\bar{v}; \qquad y(0) = 0. \tag{16d}$$

Adopting the functional form (16c) as player 1's strategy and using the technique described following Proposition 1, we obtain a unique Nash strategy for player 2 for every fixed q and p. This Nash strategy is given by (refer to Appendix A for details of the derivation)

$$\bar{v}^*(p, q) = -\frac{\beta + (1-\beta)A(p, q)}{2 + \beta + (1-\beta)A(p, q)}x_0, \tag{17a}$$

where

$$A(p, q) = \left[\int_0^1 \varphi(1, \sigma)\,d\sigma\right]_{11}, \tag{17b}$$

and $\varphi(\cdot, \cdot)$ is the 2×2 state transition matrix satisfying

$$\dot{\varphi} = \begin{bmatrix} q - 1/(2-t) & 1 \\ p & 0 \end{bmatrix}\varphi, \qquad \varphi(\sigma, \sigma) = I. \tag{17c}$$

It is important to note that the NZSDG (14a)–(14c) can be converted into an equivalent team problem if and only if $\beta = 1$ and that $\bar{v}^*$ becomes independent of different representations of (15a) (i.e. independent of p and q in this case) if and only if $\beta = 1$ (see (17a)), which is in accordance with Proposition 2.

In order to illustrate the informational nonuniqueness of the Nash solution more explicitly, we now assume $p(t) \equiv 0$ and $q(t) = \alpha + 1/(2-t)$ where α is a scalar parameter. Then we have

$$u^*(\alpha) = \alpha x - (1+\alpha t)v^*/2 - (1+\alpha(2-t))x_0/2, \tag{18a}$$

$$v^*(\alpha) = -(1-2/[2+\beta+(1-\beta)(e^{\alpha}-1)/\alpha])x_0, \tag{18b}$$

which is a Nash equilibrium pair for each value of α. If (18a) and (18b) are substituted into (14b) and (14c), the corresponding Nash costs will be

$$J_1^*(\alpha) = 2x_0^2/[2+\beta+(1-\beta)(e^{\alpha}-1)/\alpha]^2, \tag{18c}$$

$$J_2^*(\alpha) = \{(1+\beta+[\beta+(1-\beta)(e^{\alpha}-1)/\alpha]^2)/[2+\beta+(1-\beta)(e^{\alpha}-1)/\alpha]^2\}x_0^2. \tag{18d}$$

Since α was arbitrary, it follows from the above that the Nash costs are definitely nonunique whenever $\beta \neq 1$, and that they are not strictly ordered as a function of α. Especially, in the limiting case when α is sufficiently large, we observe a phenomenon similar to the one noted in [2, p. 428] within the context of LQNZSDG described by difference equations. For sufficiently large values of α, J_1^* approaches its lowest possible value, zero, whereas J_2^* approaches an unfavorable cost of x_0^2. When α approaches zero, however, $J_1^* \to (2/9)x_0^2$ and $J_2^* \to [(2+\beta)/9]x_0^2$.

After first discovering this informational nonuniqueness of the Nash solution within the context of LQNZSDG described by difference equations [1], [2], we have proposed in [1] a possible resolution of this dilemma by requiring the Nash equilibrium solution to possess some kind of a stability property. This requirement was imposed on the problem by inclusion of additive zero-mean random perturbations in the state dynamics, and we sought a robust noncooperative equilibrium solution that is insensitive to these perturbations.

In the next section, we impose similar restrictions on the continuous-time model by including an additive zero-mean independent increment process (a local martingale) in the state dynamics. We formulate this new version of the LQNZSDG of § 2 under the Ito interpretation for the stochastic differential equation and obtain explicit expressions for the unique element of $\mathcal{U}_N$ that possesses the abovementioned robust behavior.

4. Derivation of robust noncooperative equilibrium solutions. We now assume that the state of the game evolves according to the Ito stochastic differential equation

$$dx_t = F(t)x_t\,dt + G_1(t)u_1^t\,dt + G_2(t)u_2\,dt + dw_t, \qquad t \geqq t_0, \quad x_{t_0} = x_0 \tag{19a}$$

where we have adopted a different notation (from that in (1)) for the state vector and the first player's control function in order to emphasize that they are no longer deterministic functions. x_t, $t \geqq t_0$, is an n-dimensional vector stochastic process

with continuous sample paths, and x_0 is a known deterministic vector. The additive perturbation term w_t, $t \geqq t_0$, is any separable process with zero-mean independent increments, the covariance matrix of each increment being positive definite. The results obtained and the conclusions drawn in this section will, however, be valid for a larger class of processes; namely, w_t is allowed to be an n-dimensional local martingale with respect to the increasing family of sigma-fields $\mathcal{B}_t$, $t \geqq t_0$, where $\mathcal{B}_t$ is the sigma-field generated by w_s and x_s, $t_0 \leqq s \leqq t$. (See Wong [18, p. 165] for a definition of local martingale.) Furthermore, $w_{t_0} = 0$, $E[w_t w_t^T] = K(t) < \infty$, for each $t \geqq t_0$, where $K(t)$ is an $n \times n$ matrix function which is independent of the processes u_1^t, $u_2(t)$, $t_0 \leqq t \leqq t_f$, and $E[dw_t\, dw_t^T] = \Lambda(t)$ which is a positive definite matrix for all $t \geqq t_0$. An example of a stochastic process that satisfies these requirements is the zero-mean independent increment Gaussian process in which case $K(t) = \int_{t_0}^{t} \Lambda(s)\, ds$.

The coefficient matrices $F(\cdot)$, $G_1(\cdot)$ and $G_2(\cdot)$ are, as defined before, continuous matrix functions of appropriate dimensions. Again, as before, player 1 has access to perfect nonanticipative closed-loop information concerning the state of the game, and player 2 has access to open-loop information. Hence, we can take the strategy sets of players 1 and 2 to be U_1 and U_2, respectively, as defined in § 2. It is known that corresponding to any pair of strategies $\{\gamma_1 \in U_1, \gamma_2 \in U_2\}$, the stochastic functional differential equation

(19b)

$$dx_t = F(t)x_t\, dt + G_1(t)\gamma_1[t, x_s; s \leqq t]\, dt + G_2(t)\gamma_2(t)\, dt + dw_t, \qquad x_{t_0} = x_0, \quad t \geqq t_0,$$

admits a unique solution that is a sample path continuous second order stochastic process [6], [19], [20]. Furthermore, the control process $u_1^t = \gamma_1[t, x_s; s \leqq t]$ is a second order process with continuous sample paths and adapted to $\mathcal{B}_t$, which is the sigma-field generated by x_s and w_s, $t_0 \leqq s \leqq t$.

The objective functions of the players are expected values of the expressions given by (3), which we denote by $\bar{J}_i$.

Hence, the prime objective of the rest of this paper is verification of existence of a unique noncooperative equilibrium solution $\{\gamma_1^* \in U_1, \gamma_2^* \in U_2\}$ such that

(20a) $$\bar{J}_1(\gamma_1^*, \gamma_2^*) \leqq \bar{J}_1(\gamma_1, \gamma_2^*),$$

(20b) $$\bar{J}_2(\gamma_1^*, \gamma_2^*) \leqq \bar{J}_2(\gamma_1^*, \gamma_2)$$

for all $\gamma_1 \in U_1$, $\gamma_2 \in U_2$.

Now, for every fixed $\gamma_2 \in U_2$ player 2 is faced with the following stochastic control problem:

(21a) $$\min_{U_1} \bar{J}_1(u_1^t, \gamma_2)$$

with

(21b) $$\bar{J}_1 = E\left\{x_{t_f}^T C_{1_f} x_{t_f} + \int_{t_0}^{t_f} (x_t^T C_1 x_t + (u_1^t)^T D_{11} u_1^t + \gamma_2^T D_{12} \gamma_2)\, dt\right\},$$

(21c) $$dx_t = (Fx_t + G_1 u_1^t + G_2\gamma_2(t))\, dt + dw_t.$$

The unique solution to this (generalized linear regulator) problem is given below in Lemma 1. A proof of this lemma is provided in Appendix B.

LEMMA 1. *For every fixed* $\gamma_2 \in U_2$ *there exists a* unique *element of* U_1 *that solves the stochastic control problem* (21a)–(21c). *This unique solution is given by*

$$\gamma_1^*(t, x_t) = -D_{11}^{-1} G_1^T [P(t)x_t + k(t)], \tag{22a}$$

$$\dot{P} + F^T P + PF - PG_1 D_{11}^{-1} G_1^T P + C_1 = 0; \tag{22b}$$

$$P(t_f) = C_{1_f}, \qquad t_0 \leqq t \leqq t_f,$$

$$\dot{k} + (F^T - PG_1 D_{11}^{-1} G_1^T)k + PG_2 \gamma_2(t) = 0; \tag{22c}$$

$$k(t_f) = 0.$$

Remark. We note that (22a) is functionally similar to the solution (7) obtained for the deterministic version of this problem, and hence γ_1^* is an element of $\mathcal{U}_1(\gamma_2)$. However, as it has been shown in Appendix B, we do not have the nonuniqueness problem arising from different representations in this case, since $\bar{J}_1(\gamma_1^*, \gamma_2) < \bar{J}_1(\gamma_1, \gamma_2)$ for every other element γ_1 of $\mathcal{U}_1(\gamma_2)$, because of the assumption that the local martingale w_t has a positive definite incremental covariance.

It now follows from Lemma 1 that in characterizing all solutions to (20), we can restrict ourselves (without any loss of generality) to a proper subset of U_1 consisting of all measurable affine mappings γ_1 of the form

$$\gamma_1(t, x_t) = -D_{11}^{-1}(t) G_1^T(t) [P(t)x_t + l(t)], \tag{23}$$

where P is given by (22b) and l is any element of C_n.

Hence, every Nash strategy for player 1 will be of the form (23) for some l in C_n. Now replacing u_1^t by the expression given by (23) in both (19) and (3) with $i = 2$, we observe that every Nash policy for player 2 will be an optimizing solution to the following stochastic control problem for some $l \in C_n$:

$$\min_{U_2} L(u_2) \tag{24a}$$

with

(24b)

$$L(u_2) \triangleq E\left\{ x_{t_f}^T C_{2_f} x_{t_f} + \int_{t_0}^{t_f} [x_t^T \tilde{C}_2 x_t + 2x_t^T \tilde{l} + u_2^T D_{22} u_2 + l^T G_1 D_{11}^{-1} D_{21} D_{11}^{-1} G_1^T l]\, dt \right\},$$

$$\tilde{l} \triangleq PG_1 D_{11}^{-1} D_{21} D_{11}^{-1} G_1^T l, \tag{24c}$$

$$\tilde{C}_2 \triangleq C_2 + PG_1 D_{11}^{-1} D_{21} D_{11}^{-1} G_1^T P, \tag{24d}$$

and subject to

$$dx_t = (\tilde{F}x_t + G_2 u_2 + s)\, dt + dw_t, \qquad x_{t_0} = x_0, \tag{24e}$$

$$\tilde{F}(t) \triangleq F - G_1 D_{11}^{-1} G_1^T P, \tag{24f}$$

$$s(t) \triangleq -G_1 D_{11}^{-1} G_1^T l, \tag{24g}$$

where $P(t)$ is given by (22b), and l is any a priori picked element of C_n that is functionally independent of u_2. (In the above description we have intentionally suppressed the time dependence in order to avoid unnecessary repetition).

This stochastic control problem admits a unique globally optimal solution which is given below in Lemma 2. A proof of Lemma 2 is also provided in Appendix B.

LEMMA 2. *Corresponding to every a priori fixed $l \in C_n$ there exists a* unique *element γ_2^* of U_2 that solves the stochastic control problem* (24a)–(24g). *This optimal strategy can explicitly be written as a function of $l(t)$ as follows* (*where time dependence is again suppressed*):

$$u_2^* = \gamma_2^*(t) = -D_{22}^{-1} G_2^T [S(t)y(t) + b(t)], \tag{25a}$$

$$\dot{S} + \tilde{F}^T S + S\tilde{F} - SG_2 D_{22}^{-1} G_2^T S + \tilde{C}_2 = 0; \tag{25b}$$

$$S(t_f) = C_{2_f}, \qquad t_0 \leqq t \leqq t_f,$$

$$\dot{b} + \tilde{F}^T b - SG_2 D_{22}^{-1} G_2^T b + Ss + \tilde{l} = 0; \tag{25c}$$

$$b(t_f) = 0,$$

$$\dot{y} + (G_2 D_{22}^{-1} G_2^T S - \tilde{F}) y + G_2 D_{22}^{-1} G_2^T b - s = 0; \tag{25d}$$

$$y(t_0) = x_0.$$

Lemmas 1 and 2 provide us with a characterization of all possible Nash strategies of players 1 and 2, respectively, in terms of a continuous function $l(\cdot)$ yet to be determined. Now, for (23a) and (25a) to be mutually consistent as a permissible Nash strategy pair, they have to be optimal against each other, which implies that $k(t)$ should be equivalent to $l(t)$ when $\gamma_2(t)$ is replaced in (22c) by its optimal value from (25a). It is now not difficult to see that the NZSDG of this section will admit a noncooperative equilibrium solution if and only if the abovementioned compatibility condition is satisfied. This brings us to the following important result which directly follows from Lemmas 1 and 2 and the discussion given above:

THEOREM 1. *Every noncooperative Nash equilibrium solution to the LQNZSDG of this section is given by*

$$*u_1^t = \gamma_1^*(t, x_t) = -D_{11}^{-1} G_1^T (Px_t + \bar{k}), \tag{26a}$$

$$u_2^*(t) = \gamma_2^*(t) = -D_{22}^{-1} G_2^T (S\bar{y}(t) + \bar{b}) \tag{26b}$$

where $P(t)$ and $S(t)$ are as defined by (22*b*) *and* (25*b*), *respectively, and $\bar{k}$, $\bar{y}$, $\bar{b}$ satisfy the coupled differential equations*

$$\dot{\bar{k}} + \tilde{F}^T \bar{k} - PG_2 D_{22}^{-1} G_2^T (S\bar{y} + \bar{b}) = 0; \qquad \bar{k}(t_f) = 0, \tag{27a}$$

$$\dot{\bar{b}} + (\tilde{F}^T - SG_2 D_{22}^{-1} G_2^T)\bar{b} + (PG_1 D_{11}^{-1} D_{21} - SG_1) D_{11}^{-1} G_1^T \bar{k} = 0; \qquad \bar{b}(t_f) = 0, \tag{27b}$$

$$\dot{\bar{y}} - (\tilde{F} - G_2 D_{22}^{-1} G_2^T S)\bar{y} + G_2 D_{22}^{-1} G_2^T \bar{b} + G_1 D_{11}^{-1} G_1^T \bar{k} = 0; \qquad \bar{y}(t_0) = x_0. \tag{27c}$$

Furthermore, a necessary and sufficient condition for existence of a Nash equilibrium point is existence of a solution to the two-point boundary value problem (27a)–(27c).

Remark. A proof of the last part of the statement of Theorem 1 follows from the fact that since $C_1(t) \geqq 0$, $\tilde{C}_2(t) \geqq 0$, $D_{22}(t) > 0$, $D_{11}(t) > 0$, both of the Riccati differential equations (22b) and (25b) admit unique bounded nonnegative definite matrix solutions (see, e.g., Reid [14, p. 121]).

Via Theorem 1, we have now converted the original problem of investigating existence of a unique Nash equilibrium point to investigation of existence of a unique solution to the two-point boundary value problem (27a)–(27c). This, in turn, is equivalent to existence of a unique

$$z \triangleq -D_{22}^{-1} G_2^T (S\bar{y} + \bar{b}), \tag{28}$$

which also constitutes the optimal unique Nash strategy for player 2, whenever it exists. Now, adopting a Lebesgue interpretation for the integral appearing in the cost functions (3), we seek existence of a unique element $z \in L^{r_2}$ such that (27a)–(27c) are satisfied. (Here L^{r_2} denotes the Banach space of all r_2-dimensional real-valued Lebesgue square-integrable functions on $[t_0, t_f]$; i.e., if $z \in L^{r_2}$, then $\int_{t_0}^{t_f} z^T z \, dt < \infty$.) As it will be clear from the sequel, any such element (if it exists) will also be in C_{r_2} because of the nature of the differential equation involved (or the nature of operator $\mathscr{L}$ defined below by (32b)). Furthermore, it will be a unique element of C_{r_2} since any two elements of C_{r_2} are equivalent under the sup norm if and only if they are equivalent under the norm of L^{r_2}.

Denoting the state-transition matrices associated with the differential equations (27a), (27b) and (27c) by Φ_k, Φ_b and Φ_y, respectively, we write these equations in the equivalent form

$$\bar{k} = \mathscr{L}_1 \bar{z}, \tag{29a}$$

$$\bar{b} = \mathscr{L}_2 \bar{k}, \tag{29b}$$

$$\bar{y} = \Phi_y(t, t_0)x_0 + \mathscr{L}_3 \bar{k} + \mathscr{L}_4 \bar{b}, \tag{29c}$$

where z is defined by

$$z = -D_{22}^{-1} G_2^T \bar{z}, \tag{30}$$

and $\mathscr{L}_i$, $i = 1, \cdots, 4$, are linear operators mapping L^n into L^n, and are defined by

$$\mathscr{L}_1 \bar{z} = \int_{t_f}^{t} \Phi_k(t, \sigma)[PG_2 D_{22}^{-1} G_2^T \bar{z}](\sigma) \, d\sigma, \tag{31a}$$

$$\mathscr{L}_2 \bar{k} = \int_{t_f}^{t} \Phi_b(t, \sigma)[SG_1 D_{11}^{-1} G_1^T \bar{k} - PG_1 D_{11}^{-1} D_{21} D_{11}^{-1} G_1^T \bar{k}](\sigma) \, d\sigma, \tag{31b}$$

$$\mathscr{L}_3 \bar{k} = -\int_{t_0}^{t} \Phi_y(t, \sigma)[G_1 D_{11}^{-1} G_1^T \bar{k}](\sigma) \, d\sigma, \tag{31c}$$

$$\mathscr{L}_4 \bar{b} = -\int_{t_0}^{t} \Phi_y(t, \sigma)[G_2 D_{22}^{-1} G_2^T \bar{b}](\sigma) \, d\sigma. \tag{31d}$$

Compatibility condition now requires solvability of the operator equation

$$\bar{z} = S(t)\Phi_y(t, t_0)x_0 + \mathscr{L}\bar{z} \tag{32a}$$

$$\mathscr{L} \triangleq (S\mathscr{L}_3 + S\mathscr{L}_4\mathscr{L}_2 + \mathscr{L}_2)\mathscr{L}_1. \tag{32b}$$

We now find a bound on the norm $\|\cdot\|_0$ of $\mathscr{L}$, where $\|\cdot\|_0$ is the standard norm on the Banach space $\mathscr{B}(L^n)$ of continuous linear transformations of L^n onto itself, i.e.,

$$\|\mathscr{L}\|_0 \triangleq \sup_{\bar{z}\in L^n} \|\mathscr{L}\bar{z}\|_{t_f}, \qquad \|\bar{z}\|_{t_f} \leqq 1, \tag{33a}$$

where

$$\|\bar{z}\|_{t_f}^2 \triangleq \int_{t_0}^{t_f} \bar{z}^T\bar{z}\, dt. \tag{33b}$$

Preliminary remarks and notation. For every $t' \in [t_0, t_f]$, let $P(t', t)$ and $S(t', t)$ denote the unique matrix solutions to the Riccati differential equations (22b) and (25b) respectively, for $t_0 \leqq t \leqq t'$ and with the original boundary conditions replaced by $P(t') = C_{1_f}$, $S(t') = C_{2_f}$. We know that a unique nonnegative definite solution exists for every $t' \in [t_0, t_f]$ and furthermore that the solution is a continuous function of t, $t_o \leqq t \leqq t'$ (see, e.g., Reid [14]).

Since Φ_k, Φ_b and Φ_y are continuous functions of entries of S and P, they will also be continuous functions of t, $t \leqq t'$, for every fixed $t' \in (t_0, t_f]$. To denote the explicit dependence on t', we adjoin the variable t' to their argument. That is, we write

$$\Phi_k = \Phi_k(t', t, \sigma), \tag{34a}$$

$$\Phi_b = \Phi_b(t', t, \sigma), \tag{34b}$$

$$\Phi_y = \Phi_y(t', t, \sigma), \qquad t, \sigma \leqq t'. \tag{34c}$$

We now define bounded scalar numbers α_i, $i = 1, \cdots, 4$, by

$$\alpha_1 = \max_{i,j,\sigma,t,t'} |\Phi_k(t', t, \sigma)[P(t', \sigma)G_2D_{22}^{-1}G_2^T](\sigma)|_{ij}, \tag{35a}$$

$$\alpha_2 = \max_{i,j,\sigma,t,t'} |\Phi_b(t', t, \sigma)[S(t', \sigma)G_1D_{11}^{-1}G_1^T - P(t', \sigma)G_1D_{11}^{-1}D_{21}D_{11}^{-1}G_1^T](\sigma)|_{ij}, \tag{35b}$$

$$\alpha_3 = \max_{i,j,\sigma,t,t'} |\Phi_y(t', t, \sigma)G_1(\sigma)D_{11}^{-1}(\sigma)G_1^T(\sigma)|_{ij}, \tag{35c}$$

$$\alpha_4 = \max_{i,j,\sigma,t,t'} |\Phi_y(t', t, \sigma)G_2(\sigma)D_{22}^{-1}(\sigma)G_2^T(\sigma)|_{ij}, \tag{35d}$$

where $|\cdot|_{ij}$ denotes, in each case, the absolute value of the ijth element of the $n \times n$ matrix $(\cdot)$, and the maxima are taken over $i, j = 1, \cdots, n$; $t_0 \leqq \sigma$, $t \leqq t' \leqq t_f$. The maximum exists, in each case, since the matrices involved have continuous and bounded elements for every $i, j = 1, \cdots, n$, and every $t' \in [t_0, t_f]$, and the interval $[t_0, t_f]$ is compact. We further let $\bar{\lambda}_s^2$ denote the maximum value of trace of the nonnegative definite matrix $S(t', t)S(t', t)$; i.e.,

$$\bar{\lambda}_s^2 = \max_{t', t} [\operatorname{tr}(S(t', t)S(t', t))] \tag{35e}$$

where $t_0 \leqslant t \leqslant t' \leqq t_f$.

LEMMA 3. *If $\mathscr{L}$ is defined by* (32b) *and α_i, $i=1,\cdots,4$, $\bar{\lambda}_s$ by* (35), *then we have the bound*

$$\|\mathscr{L}\|_0 \leqq \left[\alpha_2+\bar{\lambda}_s\alpha_3+\bar{\lambda}_s(t_f-t_0)\frac{n}{\sqrt{2}}\alpha_2\alpha_4\right]\frac{n^2\alpha_1}{2}(t_f-t_0)^2. \tag{36}$$

Proof. Since each $\mathscr{L}_i$, $i=1,\cdots,4$, is in $\mathscr{B}(L^n)$ and since $\mathscr{B}(L^n)$ is an algebra, we have (see Simmons [15, p. 222]):

$$\begin{aligned}\|\mathscr{L}\|_o &= \|(S\mathscr{L}_3+S\mathscr{L}_4\mathscr{L}_2+\mathscr{L}_2)\mathscr{L}_1\|_0 \\ &\leqq \|S\mathscr{L}_3+S\mathscr{L}_4\mathscr{L}_2+\mathscr{L}_2\|_0\|\mathscr{L}_1\|_0 \\ &\leqq \{\|S\mathscr{L}_3\|_0+\|S\mathscr{L}_4\|_0\|\mathscr{L}_2\|_0+\|\mathscr{L}_2\|_0\}\|\mathscr{L}_1\|_0\end{aligned} \tag{37a}$$

where the last relation follows from the Minkowski inequality applied to L^n. We now note that for any $v\in L^n$,

$$\begin{aligned}\|S\mathscr{L}_3v\|_{t_f}^2 &= \int_{t_0}^{t_f}(\mathscr{L}_3v)^TS(t_f,t)S(t_f,t)\mathscr{L}_3v\,dt \\ &\leqq \max_{t_0\leqq t\leqq t_f}\lambda(S^2(t_f,t))\|\mathscr{L}_3v\|_{t_f}^2 \\ &\leqq \max_{t',t}\lambda(S^2(t',t))\|\mathscr{L}_3v\|_{t_f}^2; \qquad t'\geqq t, \\ &= \bar{\lambda}_s^2\|\mathscr{L}_3v\|_{t_f}^2,\end{aligned}$$

which implies the inequality

$$\|S\mathscr{L}_3\|_0 \leqq \bar{\lambda}_s\|\mathscr{L}_3\|_0. \tag{37b}$$

Similarly,

$$\|S\mathscr{L}_4\|_0 \leqq \bar{\lambda}_s\|\mathscr{L}_4\|_0. \tag{37c}$$

We now refer the reader to Lemma 4, Appendix C, for a proof of the bound

$$\|\mathscr{L}_i\|_0 \leqq n(t_f-t_0)\alpha_i/\sqrt{2}; \qquad i=1,\cdots,4. \tag{37d}$$

By using the relations (37b)–(37d) in (37a), we obtain the desired result (36). Q.E.D.

Lemma 3 can now be used to prove existence of a unique solution to (32a) for sufficiently small time intervals $[t_0, t_f]$. We first make the following crucial observation:

OBSERVATION 2. If the original differential game is instead defined on a shorter time interval $[t_0, t_0+\delta]$, $0<\delta\leqq t_f-t_0$, everything else remaining the same, then the statement of Lemma 3 will still be valid with t_f-t_0 replaced by δ, since $\bar{\lambda}_s$ and α_i are independent of the length δ of the time interval as long as $\delta\leqq t_f-t_0$. Therefore we have

$$\|\mathscr{L}\|_0 \leqq \left[\alpha_2+\bar{\lambda}_s\alpha_3+\bar{\lambda}_s\delta\frac{n}{\sqrt{2}}\alpha_2\alpha_4\right]\frac{n^2}{2}\alpha_1\delta^2, \tag{38}$$

for a differential game defined on a time interval of length $\delta \leqq t_f - t_0$. This implies that $\|\mathscr{L}\|_0$ can be made arbitrarily small by a sufficiently small choice of $\delta > 0$.

THEOREM 2. *The nonzero-sum differential game of* § 4 *admits a* unique *Nash equilibrium solution given by* (26a)–(26b), *if the time interval on which the game is defined is taken to be sufficiently small.*

Proof. We have previously shown that there exists a unique Nash equilibrium solution if there exists a unique $\bar{z} \in L^n$ that solves (32a).

Since $\|\mathscr{L}\|_0$ can be made less than one by a proper choice of δ (this follows from (38)), $\mathscr{L}$ is a contraction mapping for sufficiently small δ. This consequently implies existence of a unique solution to (32a) by Banach's classical fixed point theorem (Simmons [15, p. 338]), which further implies existence of a unique $z \in L^{r_2}$ through (30). Q.E.D.

Remark. Theorems 1 and 2 indicate that every robust Nash equilibrium solution of the original deterministic differential game of § 2 will be of the structural form given by (26a)–(26b), under the given information pattern. Furthermore, the solution will exist and be unique if the length of the time interval on which the game is defined is taken to be sufficiently small. As it has been noted in Theorem 1, the existence condition, in general, is solvability of a two-point boundary value problem.

The same deterministic LQNZSDG and with the same information pattern was considered previously by Foley et al. [7], who asserted a linear strategy for player 1 that is a function of only the current state vector. Their condition of existence for such a Nash strategy was nonexistence of conjugate-points in the solution of some coupled matrix-Riccati differential equations. The solution given in [7] does not correspond to the one given in Theorem 1 and neither do the existence conditions. It should be noted that the Nash equilibrium solution proposed in [7] is not Nash optimal under the additional restriction that the solution should be insensitive (robust) to external disturbances in the state dynamics, which are modeled by a local-martingale process.

5. Concluding remarks and possible extensions. In the first part of this paper (§§ 2 and 3), we have looked into the role dynamic information plays in the characterization of the Nash equilibrium set in nonzero-sum differential games (NZSDG), and have shown within the context of DG described by linear differential state dynamics and quadratic cost functionals that dynamic information pattern gives rise to existence of uncountable number of admissible (informationally nonunique) Nash solutions for DG which otherwise admit unique solutions.

In order to overcome the dilemma arising from existence of informationally nonunique Nash solutions, we proposed in the second part of the paper (§ 4) further to restrict the definition of a Nash solution in deterministic NZSDG by requiring the strategy of each player to be insensitive to external zero-mean disturbances in the state dynamics. We have modeled these disturbances by an additive local-martingale process and have shown, again within the context of LQNZSDG, that if player 1 is provided with the classical nonanticipative closed-loop information and player 2 with open-loop information, then every permissible Nash strategy of the former will be an affine transformation on the current state

vector. Furthermore, we have proven *existence* of a *unique* Nash equilibrium solution if the time interval is taken to be sufficiently small. The significance of this result, in our opinion, is that it is the first existence and uniqueness proof in the literature on NZSDG characterized by differential state dynamics and nontrivial dynamic information patterns.

If the closed-loop information pattern for player 1 is replaced by an ε-delay information pattern, it seems that it is possible (if not immediate) to extend and generalize the approach presented in § 4 in order to prove existence and uniqueness of Nash equilibrium solution for that problem also. If, however, the information available to player 2 is also dynamic, the approach taken in § 4 will no longer be applicable since it strongly makes use of the assumption that one of the players' information structure is static. Hence, we can only conjecture at this point that under a general ε-delay information pattern for both players, the stochastic differential game of § 4 will only admit Nash strategies that are affine in the available information (i.e., a counterpart of Theorem 1 will be valid). Verification of this conjecture still remains a challenge at this stage.

Appendix A. In this appendix, we supply the reader with the missing steps in the derivation of informationally nonunique Nash solutions for Example 1 (§ 3).

If player 1's strategy is fixed a priori in the functional form (16c)–(16d), player 2's best response will be an optimizing solution to

$$\min_{v\in\mathbb{R}} J_2 \tag{A.1}$$

where

$$J_2 = z^T(1)Q(1)z(1) + v^2 + \beta\int_0^1 [\alpha^T z + k_1\bar{v} + k_2 x_0]^2\, dt, \tag{A.2}$$

$$\dot{z} = Az + B\bar{v} + Cx_0 + Dv, \qquad z(0) = z_0, \tag{A.3a}$$

$$A \triangleq \begin{bmatrix} q - 1/(2-t) & 1 \\ p & 0 \end{bmatrix}, \tag{A.3b}$$

$$B^T \triangleq [qt/2 - (1-t)/(2-t),\ -tp/2], \tag{A.3c}$$

$$C^T \triangleq [-q(1-t/2),\ -p(1-t/2)], \tag{A.3d}$$

$$D^T \triangleq [1, 0], \tag{A.3e}$$

$$Q(1) \triangleq \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \alpha^T \triangleq [q - 1/(2-t),\ 1], \tag{A.3f}$$

$$k_1 \triangleq -qt/2 - (1-t)/(2-t), \tag{A.3g}$$

$$k_2 \triangleq -q(1-t/2), \tag{A.3h}$$

$$z_0^T \triangleq [x_0, 0], \tag{A.3i}$$

Now, denoting the state transition matrix associated with the differential equation (A.3a) by $\varphi(\cdot,\cdot)$ (this is also given in § 3 by (17c)), we write out the solution to (A.3a):

$$\begin{aligned} z(t) &= \varphi(t,0)z_0 + \int_0^t \varphi(t,\sigma)B(\sigma)\,d\sigma\bar{v} + \int_0^t \varphi(t,\sigma)C(\sigma)\,d\sigma x_0 \\ &\quad + \int_0^t \varphi(t,\sigma)D\,d\sigma v \\ &\triangleq \tilde{A}z_0 + \tilde{B}\bar{v} + \tilde{C}x_0 + \tilde{D}v. \end{aligned} \tag{A.4}$$

Substitution of this expression in (A.2) and minimization with respect to $v \in \mathbb{R}$ yields the unique solution

$$\begin{aligned} v = {} & -[\tilde{D}^T(1)Q(1)\tilde{D}(1) + \beta\int_0^1 (\tilde{D}^T\alpha\alpha^T\tilde{D})\,dt + 1]^{-1} \\ & \cdot \Big\{ D^T(1)Q(1)[\tilde{A}z_0 + \tilde{B}\bar{v} + \tilde{C}x_0]|_{t=1} \\ & + \beta\int_0^1 \alpha^T\tilde{D}[\alpha^T\tilde{A}z_0 + \alpha^T\tilde{B}\bar{v} + \alpha^T\tilde{C}x_0 + k_1\bar{v} + k_2x_0]\,dt \Big\}. \end{aligned} \tag{A.5}$$

Now, for consistency we require $v = \bar{v}$ in (A.5), and make use of the fact that we are operating at the equilibrium point. The latter observation provides us with a simplified version of (A.4), i.e.

$$\bar{z} = [\bar{x}, \bar{y}]^T, \quad \bar{x} = (1 - t/2)x_0 + (t/2)\bar{v}, \quad \bar{y} = 0$$

where $\bar{x}$ denotes the value of the state vector at equilibrium, and $\bar{y}$ denotes the value of the solution to (16d) at equilibrium. Using this simplification in (A.5) (with $v = \bar{v}$), we obtain the solution given by (17a), after some rather straightforward but extensive manipulations.

Appendix B. We now provide a proof for Lemmas 1 and 2 of § 4. Lemma 1 is a generalization of Kwakernaak's result [10] and hence our proof will parallel his. We first give a restatement of Lemma 1 for the sake of completeness.

LEMMA 1. *For every fixed $\gamma_2 \in U_2$ there exists a unique element of U_1 that solves the stochastic control problem* (21a)–(21c). *This unique solution is given by*

$$\gamma_1^*(t, x_t) = -D_{11}^{-1}G_1^T[P(t)x_t + k(t)], \tag{B.1a}$$

$$\dot{P} + F^TP + PF - PG_1D_{11}^{-1}G_1^TP + C_1 = 0; \tag{B.1b}$$

$$P(t_f) = C_{1f}, \qquad t_0 \leqq t \leqq t_f,$$

$$\dot{k} + (F^T - PG_1D_{11}^{-1}G_1^T)k + PG_2\gamma_2(t) = 0; \tag{B.1c}$$

$$k(t_f) = 0.$$

Proof. The objective function to be minimized is given by

$$\bar{J}_1(u_1^t) = E\Big\{ x_{t_f}^TC_{1_f}x_{t_f} + \int_{t_0}^{t_f} [x_t^TC_1x_t + (u_1^t)^TD_{11}u_1^t + \gamma_2^TD_{12}\gamma_2]\,dt \Big\} \tag{B.2a}$$

and an integral representation of the state dynamics (19) is

$$\text{(B.2b)}\quad x_t = x_0 + \int_{t_0}^{t} F(s)x_s\,ds + \int_{t_0}^{t} G_1(s)u_1^s\,ds + \int_{t_0}^{t} G_2(s)\gamma_2(s)\,ds + w_t, \qquad t \geqq t_0.$$

Since $\gamma_2 \in U_2$ it follows from [10] that $\{x_t, \mathcal{F}_t, t \geqq t_0\}$ is a semi-martingale and if $\psi(y, t)$ is any function having first and second partial derivatives with respect to $y \in \mathbb{R}^n$ and a first partial derivative with respect to t, then $\psi(x_t, t)$ can be expressed in the form (B.3) by applying the Doléans–Dade–Meyer "differentiation rule":

$$\begin{aligned}\psi(x_t, t) = \psi(x_{t_0}, t_0) &+ \int_{t_0}^{t} \psi_t(x_{s^-}, s)\,ds \\ &+ \int_{t_0}^{t} \psi_x^T(x_{s^-}, s)\,dx_s + \tfrac{1}{2}\operatorname{tr}\left[\int_{t_0}^{t} \psi_{xx}(x_{s^-}, s)\,d\Phi_s\right] \\ &+ \sum_{t_0 \leqq s \leqq t} [\psi(x_s, s) - \psi(x_{s^-}, s) - \psi_x^T(x_{s^-}, s)(x_{s^-} - x_s)]\end{aligned} \tag{B.3}$$

where Φ_t, $t \geqq t_0$ is a matrix-valued stochastic process, the ijth element of which is $\langle x_i^c, x_j^c\rangle_t$, x_i^c being the ith component of the continuous local martingale part of the process x_i. (See [10] or [5] for further elaboration on this representation.) Now taking $\psi(x_t, t) = x_t^T P(t)x_t + 2x_t^T k(t)$, where $P(\cdot)$ and $k(\cdot)$ are uniquely defined by (B.1b) and (B.1c), and applying the above differentiation rule at $t = t_f$, we have

$$\begin{aligned}x_{t_f}^T C_{1f} x_{t_f} = x_0^T P(t_0)x_0 + 2x_0^T k(t_0) &+ \int_{t_0}^{t_f} [x_{s^-}^T \dot{P}(s)x_{s^-} + 2x_{s^-}^T \dot{k}(s)]\,ds \\ &+ 2\int_{t_0}^{t_f} [P(s)x_{s^-} + k(s)]^T\,dx_s + \operatorname{tr}\int_{t_0}^{t_f} P(s)\,d\Phi_s \\ &+ \sum_{t_0 \leqq s \leqq t_f} \{x_s^T P(s)x_s - x_{s^-}^T P(s)x_{s^-} + 2x_s^T k(s) - 2x_{s^-}^T k(s) \\ &\qquad - 2[P(s)x_{s^-} + k(s)]^T(x_s - x_{s^-})\}.\end{aligned} \tag{B.4}$$

Substitution of $\dot{P}, \dot{k}$ and dx_s (from (B.1b), (B.1c) and (B.2b), respectively) into (B.4) yields the following expression for $\bar{J}_1(u_1^t)$:

$$\begin{aligned}\bar{J}_1(u_1^t) = E\Big\{ \int_{t_0}^{t_f} &\{x_s^T C_1 x_s + (u_1^s)^T D_{11} u_1^s + \gamma_2^T D_{12}\gamma_2 \\ &+ x_{s^-}^T[-C_1 + PG_1D_{11}^{-1}G_1^T P - F^T P - PF]x_{s^-} \\ &+ 2x_{s^-}^T[(PG_1D_{11}^{-1}G_1^T - F)k - PG_2\gamma_2] \\ &+ 2[Px_{s^-} + k]^T[Fx_s + G_1u_1^s + G_2\gamma_2]\}\,ds \\ &+ 2\int_{t_0}^{t_f} [Px_{s^-} + k]^T\,dw_s\end{aligned} \tag{B.5}$$

$$+x_0^T P(t_0)x_0+2x_0^T k(t_0)+\operatorname{tr}\int_{t_0}^{t_f} P\,d\Phi_s$$

$$+\sum_{t_0\leqq s\leqq t_f}[x_s^T Px_s-x_{s^-}^T Px_{s^-}+2x_s^T k(s)-2x_{s^-}^T k$$
$$-2[Px_{s^-}+k]^T(x_s-x_{s^-})]\Big\}$$

where we have suppressed the explicit time dependence when there is no ambiguity from the context.

Observations. 1) It follows from (B.2b) that $d\Phi_s = d\Phi_s^w$, where the ijth element of Φ_s^w is $\langle w_i^c, w_j^c\rangle_s$.

2) It follows from (B.2b) that $x_s - x_{s^-} = w_s - w_{s^-}$.

3) x_{s^-} and $(w_s - w_{s^-})$ are stochastically independent.

4) $(w_s - w_{s^-})$ is a zero-mean process.

5) It has been shown by Kwakernaak [10] that

$$E\Big\{\operatorname{tr}\int_{t_0}^{t_f} P\,d\Phi_s^w+\operatorname{tr}\sum_{t_0\leqq s\leqq t_f} P(w_s-w_{s^-})(w_s-w_{s^-})^T\Big\}=\int_{t_0}^{t_f} P(s)\,dK(s).$$

6) $x(t_0)=x_0$ is a deterministic vector.

Now, using the preceding observations and facts in (B.5), we obtain (after rearranging some terms):

$$\begin{aligned}\bar{J}_1(u_1^t)=E\Big\{&\int_{t_0}^{t_f}[u_1^t+D_{11}^{-1}G_1^T(Px_t+k)]^T D_{11}[u_1^t+D_{11}^{-1}G_1^T(Px_t+k)]\,dt\Big\}\\&+\int_{t_0}^{t_f}\gamma_2^T D_{12}\gamma_2\,dt+x_0^T P(t_0)x_0+2x_0^T k(t_0)\\&+2\int_{t_0}^{t_f}k^T(t)G_2(t)\gamma_2(t)\,dt\\&-\int_{t_0}^{t_f}k^T G_1 D_{11}^{-1}G_1^T k\,dt+\int_{t_0}^{t_f}P(t)\,dK(t).\end{aligned}\tag{B.6}$$

If we now make use of the assumption that the matrix function $K(t)$ is independent of the process u_1^t, $t_0\leqq t\leqq t_f$, it follows from (B.6) that minimization of $\bar{J}_1(u_1^t)$ over U_1 is equivalent to minimization of

$$\tilde{J}_1(u_1^t)=E\Big\{\int_{t_0}^{t_f}[u_1^t+D_{11}^{-1}G_1^T(Px_t+k)]^T D_{11}[u_1^t+D_{11}^{-1}G_1^T(Px_t+k)]\,dt\Big\}\geqq 0$$

over U_1. Since γ_1^* given by (B.1a) is in U_1, and since $\tilde{J}_1(\gamma_1^*)=0$, we can easily conclude that (B.1a) is a globally optimizing solution for $\bar{J}_1(u_1^t)$. In order to complete the proof of Lemma 1 we still have to show that (B.1a) is the only element of U_1 that achieves this global minimum.

If we restrict the permissible strategies to a subset of U_1 such that u_1^t is not functionally dependent on x_s, $t_0\leqq s\leqq t$, then it follows from strict convexity of $\tilde{J}_1$ that (B.1a) is the only such solution. Therefore, if there are other elements of U_1 which yield the same global minimum as γ_1^*, then they have to reflect different representations of ${}^*u_1^t$. However it follows from (B.2b) that x_t cannot be expressed in terms of x_s for any $s<t$, since $(w_t - w_{t^-})$ has a positive definite covariance for

all $t_0 < t \leqq t_f$, i.e. every element of U_1 has a unique representation. This completes the proof of the lemma. We note in passing that if this positive definiteness restriction were not imposed on the process w_t, then γ_1^* would not be the only element of U_1 that yields $\tilde{J}_1 = 0$. Q.E.D.

LEMMA 2. *Corresponding to every a priori fixed $l \in C_n$ there exists a* unique *element γ_2^* of U_2 that solves the stochastic control problem* (24a)–(24g). *This optimal strategy can explicitly be written as a function of $l(t)$ as follows*:

$$u_2^* = \gamma_2^*(t) = -D_{22}^{-1} G_2^T [S(t) y(t) + b(t)], \tag{B.7}$$

where S, b and y are uniquely defined by (25b)–(25d).

Proof. Working along the same lines of the proof of Lemma 1, we pick $\psi(x_t, t) = x_t^T S(t) x_t + 2 x_t^T b(t)$ and arrive at the following expression for $L(u_2)$ (after similar reasoning and manipulations):

$$\begin{aligned} L(u_2) = E\Big\{ & \int_{t_0}^{t_f} [u_2(t) + D_{22}^{-1} G_2^T (S x_t + b)]^T D_{22} [u_2(t) + D_{22}^{-1} G_2^T (S x_t + b)]\, dt \Big\} \\ & + \int_{t_0}^{t_f} l^T G_1 D_{11}^{-1} D_{21} D_{11}^{-1} G_1^T l \, dt + x_0^T S(t_0) x_0 + 2 x_0^T b(t_0) \\ & + 2 \int_{t_0}^{t_f} b^T s \, dt - \int_{t_0}^{t_f} b^T G_2 D_{22}^{-1} G_2^T b \, dt + \int_{t_0}^{t_f} S(t)\, dK(t). \end{aligned} \tag{B.8}$$

Since only the first term determines the optimal solution, minimization of $L(u_2)$ is equivalent to minimization of the nonnegative expression

$$\tilde{L}(u_2) = E\Big\{ \int_{t_0}^{t_f} [u_2(t) + D_{22}^{-1} G_2^T (S x_t + b)]^T D_{22} [u_2(t) + D_{22}^{-1} G_2^T (S x_t + b)]\, dt \Big\}. \tag{B.9}$$

We now decompose the solution to (24e) into two parts—a purely deterministic part to be denoted by $z(t)$, and a purely stochastic part that is functionally independent of u_2, which we denote by v_t. $z(t)$ satisfies the DE

$$\dot{z} = \tilde{F} z + G_2 u_2 + s, \qquad z(t_0) = x_0, \tag{B.10a}$$

and v_t satisfies the stochastic DE

$$dv_t = \tilde{F} v_t \, dt + dw_t, \qquad v_{t_0} = 0. \tag{B.10b}$$

If we now replace x_t in (B.9) by $z(t) + v_t$ and note that (i) v_t defines a zero-mean semi-martingale, (ii) $u_2(t)$ is functionally and statistically independent of v_t, we can rewrite (B.9) in the equivalent form

$$\begin{aligned} \tilde{L}(u_2) = E\Big\{ & \int_{t_0}^{t_f} [u_2 + D_{22}^{-1} G_2^T (S z + b)]^T D_{22} [u_2 + D_{22}^{-1} G_2^T (S z + b)]\, dt \Big\} \\ & + E\Big\{ \int_{t_0}^{t_f} v_t^T S G_2 D_{22}^{-1} G_2^T S v_t \, dt \Big\} \end{aligned}$$

and since the first term inside the expectation is completely deterministic,

$$\tilde{L}(u_2) = \int_{t_0}^{t_f} [u_2 + D_{22}^{-1} G_2^T (Sz + b)]^T D_{22} [u_2 + D_{22}^{-1} G_2^T (Sz + b)]\, dt + \operatorname{tr} \int_{t_0}^{t_f} SG_2 D_{22}^{-1} G_2^T S E[v_t v_t^T]\, dt. \tag{B.11}$$

Minimization of $\tilde{L}(u_2)$ over U_2 is now equivalent to minimization of the first term of (B.11) which is a strictly convex functional. It therefore follows that

$$u_2 = -D_{22}^{-1} G_2^T (Sz + b) \tag{B.12}$$

yields the unique global minimum of $\tilde{L}(u_2)$. Substitution of this solution into (B.10a) yields the DE (25d) and replacing z in (B.12) by the unique solution of this DE provides us with the *unique* element of U_2 given by (B.7), (It is important to remind the reader that the notion of uniqueness used here is with respect to equivalence classes formed in U_2 under the Lebesgue measure, i.e. two elements of U_2 are said to be equivalent if they differ at most on a set of Lebesgue measure zero, but since $U_2 = C_{r_2}$ this is equivalent to the notion of equivalence with respect to the sup norm.) This completes the proof of Lemma 2. Q.E.D.

Appendix C. In this appendix we provide a proof for Lemma 4 which has been used in the proof of Lemma 3 in § 4.

LEMMA 4. *If* $\mathscr{L}_i \in \mathscr{B}(L^n)$, $i = 1, \cdots, 4$, *are defined by* (31a)–(31d) *and* α_i *by* (35a)–(35d), *we have the bound*

$$\|\mathscr{L}_i\|_0 \leqq n(t_f - t_0)\alpha_i/\sqrt{2}; \qquad i = 1, \cdots, 4. \tag{C.1}$$

Proof. We first note that because of the structurally similar forms of (31a)–(31d), and taking into account the modifications (34a)–(34c), the proof will be completed if we can show that

$$\|\mathscr{S}\|_0 \leqq n(t_f - t_0)\alpha/\sqrt{2} \tag{C.2a}$$

where $\mathscr{S} \in \mathscr{B}(L^n)$ is defined by

$$\mathscr{S}v = \int_{t_0}^{t} A(t_f, t, \sigma) v(\sigma)\, d\sigma, \qquad t \leqq t_f \tag{C.2b}$$

for some $A(\cdot, \cdot, \cdot)$ which has square Lebesgue integrable elements, and α is given by

$$\alpha = \max_{i,j,\sigma,t,t'} |A(t', t, \sigma)|_{ij}; \qquad t_0 \leqq t, \sigma \leqq t' \leqq t_f, \quad i, j = 1, \cdots, n. \tag{C.2c}$$

Since $\|\mathscr{S}\|_0 \triangleq \sup \|\mathscr{S}v\|_{t_f}$; $\|v\|_{t_f} \leqq 1$, we start with

$$\int_{t_0}^{t} \int_{t_0}^{t} v^T(\sigma) K_t(\sigma, s) v(s)\, ds\, d\sigma \tag{C.3a}$$

where

$$K_t(\sigma, s) \triangleq A^T(t_f, t, \sigma) A(t_f, t, s). \tag{C.3b}$$

Denoting the ijth component of K_t by K_t^{ij} and the ith component of v by v_i, we can write (C.3a) as follows:

$$
\begin{aligned}
&\sum_{i,j}\int_{t_0}^{t} v_i(\sigma)\,d\sigma\int_{t_0}^{t} K_t^{ij}(\sigma,s)v_j(s)\,ds \\
&\quad\leqq\sum_{i,j}\int_{t_0}^{t} v_i(\sigma)\,d\sigma\left[\int_{t_0}^{t}|K_t^{ij}(\sigma,s)|^2\,ds\right]^{1/2}\left[\int_{t_0}^{t} v_j^2(s)\,ds\right]^{1/2} \\
&\quad\leqq\sum_{i,j}\left[\int_{t_0}^{t}\int_{t_0}^{t}|K_t^{ij}(\sigma,s)|^2\,ds\,d\sigma\right]^{1/2}\cdot\left[\int_{t_0}^{t} v_i^2(\sigma)\,d\sigma\right]^{1/2}\cdot\left[\int_{t_0}^{t} v_j^2(\sigma)\,d\sigma\right]^{1/2} \\
&\quad\triangleq\sum_{i,j}\|K_t^{ij}\|_t\cdot\|v_i\|_t\cdot\|v_j\|_t
\end{aligned}
\tag{C.4}
$$

where we have made repeated use of Buniakowski's inequality since $L^n(t_0, t)$ is a Hilbert space. Expression (C.4) can be bounded from above by

$$
\begin{aligned}
\sum_{i,j}\|K_t^{ij}\|_t\cdot(\|v_i\|_t^2+\|v_j\|_t^2)/2&=\sum_i\|v_i\|_t^2\sum_j(\|K_t^{ij}\|_t+\|K_t^{ji}\|_t)/2 \\
&=\sum_i\|v_i\|_t^2\cdot\sum_j\|K_t^{ij}\|_t,
\end{aligned}
\tag{C.5}
$$

where in arriving at the last equality we have made use of the symmetry property of K_t as defined by (C.3b).

It now follows from (C.5) that

$$
\|\mathcal{S}v\|_{t_f}^2\leqq\sum_{i,j}\int_{t_0}^{t_f}\|v_i\|_t^2\cdot\|K_t^{ij}\|_t\,dt
$$

Since $\|v_i\|_t$ is a monotonically nondecreasing function of t for any $v\in L^n$, we can bound the last expression from above by

$$
\sum_{i,j}\|v_i\|_{t_f}^2\int_{t_0}^{t_f}\|K_t^{ij}\|_t\,dt,
$$

which can further be bounded from above by

$$
\sum_i\bar{h}\|v_i\|_{t_f}^2=\bar{h}\|v\|_{t_f}^2
$$

where

$$
\bar{h}=\max_i\sum_j\int_{t_0}^{t_f}\|K_t^{ij}\|_t\,dt. \tag{C.6a}
$$

This implies that

$$
\|\mathcal{S}v\|_{t_f}^2\leqq\bar{h}\|v\|_{t_f}^2
$$

and hence

$$
\|\mathcal{S}\|_0\leqq\sqrt{\bar{h}}. \tag{C.6b}
$$

To find an explicit upper bound on $\bar{h}$, we first note that

$$\begin{aligned}
\|K_t^{ij}\|_t &= \left[\int_{t_0}^{t}\int_{t_0}^{t} |K_t^{ij}(\sigma, s)|^2\, ds\, d\sigma\right]^{1/2} \\
&= \left[\int_{t_0}^{t}\int_{t_0}^{t} |A^T(t_f, t, \sigma)A(t_f, t, s)|^2\, ds\, d\sigma\right]^{1/2} \\
&\leqq \max_{t,\sigma,s} |A^T(t_f, t, \sigma)A(t_f, t, s)|_{ij} \cdot (t-t_0); \qquad \sigma, s \leqq t, \\
&\leqq n \cdot \max_{i,j}\max_{t,\sigma} |A(t_f, t, \sigma)|_{ij}^2 \cdot (t-t_0) \\
&\leqq n \cdot \{\max_{i,j,t',t,\sigma} |A(t', t, \sigma)|_{ij}\}^2 \cdot (t-t_0), \qquad t_0 \leqq t, \sigma \leqq t' \leqq t_f,
\end{aligned}$$

and substitution of this bound in (C.6a) yields

$$\begin{aligned}
\bar{h} &\leqq n^2 \cdot \{\max_{i,j,t',t,\sigma} |A(t', t, \sigma)|_{ij}\}^2 \cdot (t_f - t_0)^2/2 \\
&\triangleq n^2\alpha^2(t_f - t_0)^2/2
\end{aligned} \tag{C.7}$$

which is the desired bound.

This completes verification of the bounds (C.1) for all $i = 1, \cdots, 4$, since if (C.2b) is instead defined with t_0 replaced by t_f the procedure described above will still yield the bound given by (C.7). Q.E.D.

REFERENCES

[1] T. BAŞAR, *On the uniqueness of the Nash solution in linear-quadratic differential games*, Internat. J. Game Theory, 5 (1976), no. 2/3, pp. 65–90.

[2] ———, *A counterexample in linear-quadratic games: Existence of nonlinear Nash solutions*, J. Optimization Theory Appl., 14 (1974), no. 4, pp. 425–430.

[3] ———, *Nash strategies for M-person differential games with mixed information structures*, Automatica, 11 (1975), no. 4, pp. 547–551.

[4] A. E. BRYSON AND Y. C. HO, *Applied Optimal Control*, Blaisdell, Waltham, MA, 1969.

[5] C. DOLÉANS-DADE AND P. MEYER, *Intégrales stochastiques par rapport aux martingales locales*, Seminaire de Probabilités IV, Lecture Notes in Mathematics, vol. 124, Springer-Verlag, Berlin, 1970, pp. 77–107.

[6] W. H. FLEMING AND M. NISIO, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), no. 5, pp. 777–794.

[7] M. H. FOLEY AND W. E. SCHMITENDORF, *On a class of non-zero-sum linear-quadratic differential games*, J. Optimization Theory Appl., 7 (1971), no. 5, pp. 357–377.

[8] A. FRIEDMAN, *Differential Games*, John Wiley, New York, 1971, Chap. 8.

[9] ———, *Stochastic differential games*, J. Differential Equations, 11 (1972), no. 1, pp. 79–108.

[10] H. KWAKERNAAK, *An extension of the stochastic linear regulator problem*, IEEE Trans. Automatic Control, AC-19 (1974), no. 2, pp. 121–123.

[11] D. L. LUKES, *Equilibrium feedback control in linear games with quadratic costs*, this Journal, 9 (1971), pp. 234–252.

[12] D. L. LUKES AND D. L. RUSSELL, *A global theory for linear-quadratic differential games*, J. Math. Anal. Appl., 33 (1971), pp. 96–123.

[13] J. F. NASH, *Noncooperative games*, Ann. of Math., 54 (1951), pp. 286–295.

[14] W. T. REID, *Riccati Differential Equations*, Academic Press, New York, 1972.

[15] G. F. SIMMONS, *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York, 1963.
[16] A. W. STARR AND Y. C. HO, *Nonzero-sum differential games*, J. Optimization Theory Appl. 3 (1969), no. 3, pp. 184–206.
[17] ———, *Further properties of nonzero-sum differential games*, Ibid., 3 (1969), no. 4, pp. 207–219.
[18] E. WONG, *Stochastic Processes in Information and Dynamical Systems*, McGraw-Hill, New York, 1971.
[19] W. H. FLEMING, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.
[20] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. II, A. T. Bharuca-Reid, ed., Academic Press, New York, 1970, pp. 131–212.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

ON A NONLINEAR EVASION PROBLEM

BARBARA KAŚKOSZ

Abstract. We prove a theorem which asserts the possibility of approximating a "relaxed controls" strategy by ordinary strategies. Next we consider a nonlinear game of evasion and give a sufficient condition of evasion.

1. Introduction. The differential game of evasion is a game given by a system of equations:

$$\dot{z} = P(z, u, v), \qquad z \in R^n, \quad u \in U, \quad v \in V, \tag{1.1}$$

two compact control sets U, V and a linear subspace M of R^n of codim $M \geqq 2$. The parameter u is controlled by the pursuer who wants the trajectory $z(t)$ to hit the subspace M; the parameter v controls the evader whose aim is opposite. We assume the following conditions:

(a) $P(z, u, v)$ is a continuous function on $R^n \times U \times V$,

(b) there exist positive constants A and B such that

$$|z \cdot P(z, u, v)| \leqq A|z|^2 + B \quad \text{for all } (u, v) \in U \times V,$$

(c) for each $R > 0$ there exists a constant C_R such that if $|z| \leqq R$, $|\bar{z}| \leqq R$ then

$$|P(z, u, v) - P(\bar{z}, u, v)| \leqq C_R|z - \bar{z}| \quad \text{for all } (u, v) \in U \times V.$$

The conditions (a)–(c) ensure that for any two measurable functions $u(t)$, $v(t)$ defined on $[0, +\infty)$, $u(t) \in U$, $v(t) \in V$, and any initial condition $z_0 \in R^n$ there exists on $[0, +\infty)$ the unique solution of the initial problem:

$$\begin{aligned} \dot{z}(t) &= P(z(t), u(t), v(t)), \\ z(0) &= z_0. \end{aligned} \tag{1.2}$$

The aim of the evader is to find for every initial state $z_0 \notin M$ and every pursuer's measurable control function $u(t)$, $u(t) \in U$, a measurable function $v(t)$, $v(t) \in V$, such that the corresponding trajectory of (1.2) satisfies $z(t) \notin M$ for all $t \in [0, +\infty)$. Similarly as in [1]–[5], to prove that evasion is possible we construct two closed sets W_1, W, containing M in their interiors, $W_1 \subset W$ and such that: if $z_0 \notin W$ the trajectory remains outside W_1 for all $t \geqq 0$; if $z_0 \in W$ then the trajectory leaves W after a certain short period of time and remains outside W_1 for the rest of the time. It is assumed, quite naturally, that the evader choosing at any moment t a value $v(t)$ of his control function does not know the future, but he can make use of information concerning the past and present. In particular, he does not know the future behavior of the opposer and hence his choice of control should be independent of the values $u(s)$ for $s > t$. To make the statement more precise we define a notion of strategy corresponding to this information pattern.

* Received by the editors June 29, 1976, and in revised form November 19, 1976.

† Instytut Matematyczny Polskiej Akademii Nauk, Warsaw, Poland.

Fix an arbitrary initial condition z_0. A mapping $v^u(z_0; t)$ which assigns to each pursuer's control function $u(t)$, $t \in [0, +\infty)$, an evader's control function $v(t) = v^u(z_0; t)$, $t \in [0, +\infty)$, is called a strategy if for any T and any two control functions $u_1(t)$, $u_2(t)$ the equality $u_1(t) = u_2(t)$ a.e. in $[0, T]$ implies that $v^{u_1}(z_0; t) = v^{u_2}(z_0; t)$ a.e. in $[0, T]$.

The problem as stated above has been considered by L. S. Pontryagin and E. F. Mishchenko in [1]–[2] for a linear game of the form $\dot{z} = Cz + u + v$, where C denotes a constant matrix, and then by R. V. Gamkrelidze and G. L. Kcharatishvili [4], E. F. Mishchenko [3], M. S. Nikolski [5] and others for the game of the form $\dot{z} = Cz + f(u, v)$. In each of papers [1]–[5] sufficient conditions of evasion are given, that is, conditions under which a strategy of evasion $v^u(z_0; t)$ can be constructed for each initial state $z_0 \notin M$.

In the present paper we consider a nonlinear evasion problem. First, in § 2 we prove an approximation theorem. The theorem answers a question stated in [5]. It asserts that for each strategy for the extended game of the form $\dot{z} = \sum_{i=1}^{k} \mu_i P(z, u, v_i)$, where the player v uses so-called relaxed controls, that is, collections $(\mu_1, \cdots, \mu_k, v_1, \cdots, v_k)$, $\sum_{i=1}^{k} \mu_i = 1$, $\mu_i \geqq 0$, $v_i \in V$, $i = 1, \cdots, k$, there exists a strategy for the game (1.1) such that for any $u(t)$ the corresponding trajectory of (1.1) approximates uniformly that of the extended game. Conditions of evasion contain usually an assumption that an intersection of certain sets, sort of a set of superiority, has a nonempty interior. The approximation theorem of § 2 allows us to consider a bigger set as a set of superiority. In Mishchenko's paper [3], for example, it is assumed that for some integer k and some two-dimensional subspace L orthogonal to M, each of the sets $\Pi C^r f(U, V)$, $r = 0, \cdots, k-2$, consists of a single point, where Π denotes the orthogonal projection onto L, and the set $\bigcap_{u \in U} \Pi C^{k-1} f(u, V)$ contains an interior point in L. The conditions of evasion we give in § 3, when applied to the linear case, lead to a weaker assumption, namely that the set $\bigcap_{u \in U} \operatorname{co} \Pi C^{k-1} f(u, V)$ contains an interior point in L and each of the sets $\Pi C^r f(U, V)$, $r = 0, \cdots, k-2$, consists of a single point.

The recent paper by P. B. Gusatnikov [6] concerns also a nonlinear game of evasion. Again our conditions of evasion are essentially more general than those in [6] because of the use of the approximation theorem. Our method of constructing the sets W, W_1 is different than that in [6].

2. The approximation theorem. Consider two games in R^n:

$$\dot{z} = P(z, u, v), \qquad u \in U, \quad v \in V, \tag{2.1}$$

$$\dot{z} = \sum_{i=1}^{n+1} \mu_i P(z, u, v_i), \quad u \in U, \quad \sum_{i=1}^{n+1} \mu_i = 1, \quad \mu_i \geqq 0, \quad v_i \in V \quad \text{for } i = 1, \cdots n+1, \tag{2.2}$$

where U, V are compact, $P(z, u, v)$ satisfies (a)–(c). In the game (2.2) the player v chooses instead of one control function $v(t)$ a collection $\tilde{v}(t) = (\mu_1(t), \cdots \mu_{n+1}(t), v_1(t), \cdots v_{n+1}(t))$ of measurable functions. This extends his chances. In fact, in the game (2.1) the player v chooses for each $u(t)$ a point from the set $P(z(t), u(t), V)$

while in the game (2.2) he is allowed to choose a point from the set $\operatorname{co} P(z(t), u(t), V)$.

Fix an arbitrary initial condition z_0 and a compact interval $[0, T]$. The following theorem holds:

THEOREM 2.1. *Let* $\tilde{v}^u(z_0; t) = \tilde{v}^u(t) = (\mu_1^u(t), \cdots \mu_{n+1}^u(t), v_1^u(t), \cdots, v_{n+1}^u(t))$ *be a strategy for the game* (2.2). *Then for every* $\varepsilon > 0$ *there exists a strategy* $v^u(z_0; t) = v^u(t)$ *for the game* (2.1) *such that for any control function* $u(t)$ *the trajectories* $z_1(t)$, $z_2(t)$ *of the games* (2.1), (2.2) *corresponding to* z_0, $u(t)$ *and* $v^u(t)$, $\tilde{v}^u(t)$, *respectively, satisfy*:

$$|z_1(t) - z_2(t)| < \varepsilon \quad \text{for } t \in [0, T]. \tag{2.3}$$

Before giving the proof of this theorem, we prove first the following

LEMMA 2.1. *Let* $p_1(t), \cdots, p_m(t)$ *be a collection of measurable, bounded functions defined on an interval* $[0, T]$ *taking values in* R^n, $\mu_1(t), \cdots, \mu_m(t)$ *a collection of measurable scalar functions on* $[0, T]$ *such that* $\sum_{i=1}^m \mu_i(t) = 1$, $\mu_i(t) \geqq 0$ *for* $i = 1, \cdots, m$, $t \in [0, T]$. *Then for any* $\varepsilon > 0$ *there exists a measurable function* $p(t)$, $p(t) \in \{p_1(t), \cdots p_m(t)\}$ *for* $t \in [0, T]$ *and such that*:

$$\sup_{t \in [0, T]} \left| \int_0^t \left(\sum_{i=1}^m \mu_i(\tau) p_i(\tau) - p(\tau) \right) d\tau \right| < \varepsilon. \tag{2.4}$$

Moreover, the value $p(t)$ *at any moment* t *does not depend on the values* $\mu_i(s)$, $p_i(s)$ *for* $s > t$.

Proof. Take a constant R such that $|p_i(t)| < R/2$ for $t \in [0, T]$, $i = 1, \cdots, m$ and an integer n such that

$$T \cdot \frac{R}{\sqrt{n}} < \varepsilon. \tag{2.5}$$

Divide the interval $[0, T]$ into n intervals of the length $\delta = T/n$: $I_1 \cup \cdots \cup I_n = [0, T]$, $I_j = [(j-1)\delta, j\delta]$, $j = 1, \cdots, n$. We will construct the function $p(t)$ step-by-step on each interval I_j. Put:

$$p(t) = p_1(t) \quad \text{for } t \in I_1.$$

In order to define $p(t)$ on I_j, $j = 2, \cdots, n$, take

$$s_{j-1} = \int_0^{(j-1)\delta} \left(\sum_{i=1}^m \mu_i(\tau) p_i(\tau) - p(\tau) \right) d\tau.$$

If $s_{j-1} = 0$ define $p(t) = p_1(t)$. If $s_{j-1} \neq 0$ take an orthogonal basis $\xi^{j-1} = (\xi_1^{j-1}, \cdots, \xi_n^{j-1})$ in R^n such that $\xi_1^{j-1} = s_{j-1}$. Put

$$p(t) = \max \operatorname{lex}_{\xi^{j-1}} \{p_1(t), \cdots, p_m(t)\} \quad \text{for } t \in I_j,$$

where "$\max \operatorname{lex}_{\xi^{j-1}}$" means the lexicographical maximum with respect to the basis ξ^{j-1}, that is, maximum with respect to the following order: if $x, y \in R^n$ and x^i, y^i are coordinates of x, y with respect to the basis ξ^{j-1} then x is not greater than y iff $x = y$ or there exists k, $1 \leqq k \leqq n$, such that $x^i = y^i$ for $i = 1, \cdots, k-1$ and $x^k < y^k$.

The function $p(t)$ is uniquely defined in this way; it is measurable and has the following property:

$$\langle p(t), s_{j-1}\rangle = \max_{i=1,\cdots,m} \langle p_i(t), s_{j-1}\rangle$$

and therefore

$$\left\langle \sum_{i=1}^{m} \mu_i(t)p_i(t) - p(t),\, s_{j-1}\right\rangle \leqq 0 \quad \text{for } t \in I_j. \tag{2.6}$$

We have the estimation

$$\int_{(j-1)\delta}^{j\delta} \left| \sum_{i=1}^{m} p_i(\tau)\mu_i(\tau) - p(\tau)\right| d\tau < R\delta \quad \text{for } j = 1, \cdots, n.$$

Because of (2.6) and the definition of s_{j-1} we have for $j = 2, \cdots, n$,

$$\left|\int_0^t \left(\sum_{i=1}^{m} p_i(\tau)\mu_i(\tau) - p(\tau)\right) d\tau\right|$$

$$= \left|s_{j-1} + \int_{(j-1)\delta}^{t} \left(\sum_{i=1}^{m} p_i(\tau)\mu_i(\tau) - p(\tau)\right) d\tau\right| \leqq \sqrt{|s_{j-1}|^2 + R^2\delta^2} \quad \text{if } t \in I_j.$$

Thus an easy induction argument gives the inequality

$$\left|\int_0^t \left(\sum_{i=1}^{m} p_i(\tau)\mu_i(\tau) - p(\tau)\right) d\tau\right| < R\delta\sqrt{j} \quad \text{for } t \in I_j, \quad j = 1, \cdots, n, \tag{2.7}$$

which together with (2.5) implies (2.4).

We proceed now to prove the theorem. Because of (b) all trajectories of (2.1) and (2.2) with the initial condition z_0 remain over the interval $[0, T]$ in a certain ball of radius r_0. Thus, because of (c), the function $P(z, u, v)$ satisfies the Lipschitz condition with a constant $C = C_{r_0}$ along any two trajectories starting from z_0. Using the integral form of the equations (2.1), (2.2) we obtain

$$|z_1(t) - z_2(t)|$$

$$= \left|\int_0^t \left(P(z_1(\tau), u(\tau), v^u(\tau)) - \sum_{i=1}^{n+1} \mu_i^u(\tau) P(z_2(\tau), u(\tau), v_i^u(\tau))\right) d\tau\right|;$$

therefore

$$\begin{aligned} &|z_1(t) - z_2(t)| \\ &\leqq \left|\int_0^t \left(P(z_2(\tau), u(\tau), v^u(\tau)) - \sum_{i=1}^{n+1} \mu_i^u(\tau) P(z_2(\tau), u(\tau), v_i^u(\tau))\right) d\tau\right| \\ &\quad + C\int_0^t |z_1(\tau) - z_2(\tau)|\, d\tau. \end{aligned} \tag{2.8}$$

Applying to (2.8) Gronwall's lemma, we deduce that it is enough to find a strategy $v^u(t)$ for the strategy $\tilde{v}^u$ such that for each function $u(t)$ and each

$t \in [0, T]$

$$\left| \int_0^t \Big(P(z_2(\tau), u(\tau), v^u(\tau)) - \sum_{i=1}^{n+1} \mu_i^u(\tau) P(z_2(\tau), u(\tau), v_i^u(\tau)) \Big) d\tau \right| < \tilde{\varepsilon}, \tag{2.9}$$

where $\tilde{\varepsilon} < \varepsilon \cdot e^{-CT}$. Take $p_i^u(t) = P(z_2(t), u(t), v_i^u(t))$ and apply Lemma (2.1). We obtain the function $p^u(t)$ and define $v^u(t)$ for $t \in [0, T]$ as follows:

$$v^u(t) = v_i^u(t) \quad \text{where } i \text{ is the smallest integer such that } p^u(t) = p_i^u(t).$$

The function $v^u(t)$ defined in this way is measurable for each $u(t)$ and satisfies (2.9). The mapping $v^u(t)$ is a strategy because of properties $p(t)$. Thus the proof of Theorem (2.1) is completed.

Remark. The function $P(z, u, v)$ above can be replaced by a function $P(t, z, u, v)$ which is measurable in t for fixed (z, u, v) and continuous in (z, u, v) for fixed t and which satisfies (b) uniformly for all t and (c) for all $t \in [0, T]$ with a constant $C_{R,T}$.

3. The evasion theorem. In this section we give a solution of the evasion problem when the game is described by the following equation:

$$\dot{z} = P_0(z) + f(z, u, v), \qquad z \in R^n, \quad u \in U \subset R^p, \quad v \in V \subset R^q, \tag{3.1}$$

two compact sets U, V and a subspace M of R^n such that codim $M \geqq 2$. Assume that the function $P(z, u, v) = P_0(z) + f(z, u, v)$ satisfies (a)–(c). The extended game takes the form:

$$\dot{z} = P_0(z) + \sum_{i=1}^{(n+1)} \mu_i f(z, u, v_i), \qquad u \in U, \quad \mu = (\mu_1 \cdots \mu_{n+1}) \in \Delta, \tag{3.2}$$

$$(v_1, \cdots, v_{n+1}) \in V \times \cdots \times V$$

where $\Delta = \{\mu \in R^{n+1} | \sum_{i=1}^{n+1} \mu_i = 1, \mu_i \geqq 0\}$. For every initial state $z_0 \in R^n$, a pair of measurable functions $u(t)$, $v(t)$ such that $u(t) \in U$, $v(t) \in V$ for $t \in [0, +\infty)$ defines a trajectory of the game (3.1), and a pair of measurable functions $u(t)$, $\tilde{v}(t)$ such that $u(t) \in U$, $\tilde{v}(t) \in \tilde{V} = \Delta \times V \times \cdots \times V \subset R^{(n+1)+(n+1)q}$ defines a trajectory of the game (3.2). Let P_i, P_j be two mappings from R^n into R^n. By $DP_i \cdot P_j$ we denote the mapping from R^n into R^n defined by $DP_i \cdot P_j(z) = DP_i(z) \cdot P_j(z)$, where $DP_i(z)$ is the differential of P_i at a point z. Let I denote the identity matrix. We shall use the following notation:

$$C_o(z) = I, \qquad C_1(z) = DP_0(z),$$

$$C_r(z) = D(D(\cdots \underbrace{(DP_0 \cdot P_0) \cdots) P_0)(z).}_{r \text{ times}}$$

We assume that the mapping P_0 is differentiable as many times as it is required in the conditions which follows. The conditions which we give next are basic for

constructing a strategy of evasion and we shall refer to them as conditions of evasion or conditions (E).

(E) For each $z_* \in M$ there exist a two-dimensional subspace $L = L(z_*)$ of R^n orthogonal to M, an open neighborhood $\mathcal{U}_{z_*} \ni z_*$ and an integer $p = p(z_*)$ such that the following two conditions (i), (ii), hold, where $\Pi = \Pi(z_*)$ denotes the orthogonal projection R^n onto $L(z_*)$:

(i) $\{\Pi C_r(z) f(z, u, v) | (u, v) \in U \times V\} = \{0\}$ for all $z \in \mathcal{U}_{z_*}$, $r = 0, \cdots p-2$,

(ii) $\bigcap_{u \in U} \operatorname{co}\{\Pi C_{p-1}(z_*) f(z_*, u, v) | v \in V\}$ contains an interior point with respect to L.

Under the conditions (E) an evasion strategy can be constructed for all $z_0 \notin M$. The following theorem holds:

THEOREM 3.1. *If for the game* (3.1) *the conditions* (E) *are satisfied, then there exist closed sets* W, W_1, $M \subset \operatorname{int} W_1 \subset \operatorname{int} W$, *a positive function* $T(\xi)$, $T(\xi) < 1$, $\xi \in (0, +\infty)$, *a positive function* $\gamma(\xi_1, \xi_2)$, $\xi_1, \xi_2 \in (0, +\infty)$ *and a strategy of evasion defined for all* $z_0 \notin M$ *such that any corresponding trajectory satisfies*:

if $z_0 \in W$ *then* $\rho(z(t), M) \geqq \gamma(\rho(z_0, M), |z_0|)$ *for* $t \in [0, T(|z_0|)]$ *and* $z(T(|z_0|)) \notin W$;

if for some t_1, $z(t_i) \notin W$ *then* $z(t) \notin W_1$ *for all* $t \geqq t_1$;

if $z(t_1) \in W$ *then for some* $t_2 \in [t_1, t + T(|z(t_1)|)]$ $z(t_2) \notin W$.

In the proof of the theorem we will construct the sets W, W_1 which are unions of cylinders around M and describe a strategy of evasion. Briefly speaking, according to the strategy, the evader is doing something, say choosing a constant control $\bar{v}$, $\bar{v} \in V$, as long as the trajectory remains outside the set W. When it hits W at a moment t_1 the evader begins an actual maneuvre of evading that takes some time $T = T(|z(t_1)|)$. The trajectory remains outside the set W_1 during the maneuvre and it appears outside W at the moment $t_1 + T$. The evader again puts his control function equal to $\bar{v}$ till the next moment t_2 such that $z(t_2) \in W$. If $z(0) = z_0 \in W$ then the evader starts with a maneuvre of evading which brings the trajectory outside W at time $T(|z_0|)$ and ensures the estimation given by the function γ over the interval $[0, T(|z_0|)]$.

Consider a particular case of the game (3.1) when the equation is linear:

$$\dot{z} = Cz + f(u, v) + a,$$

where a is a constant vector, C is a constant $n \times n$ matrix. The conditions (E) take in this case the following form: there exist a two-dimensional subspace L orthogonal to M and an integer p such that:

(i) each of the sets $\{\Pi C^{r-1} f(u, v) | (u, v) \in U \times V\}$ for $r = 1, \cdots, p-1$ consists of the origin,

(ii) the set $\bigcap_{u \in U} \operatorname{co}\{\Pi C^{p-1} f(u, v) | v \in V\}$ contains an interior point with respect to L.

Hence the conditions (E) contain in this case the conditions from [3].

We will prove Theorem 3.1 in several steps.

(A) Take $z_* \in M$ and a neighborhood $\mathcal{U}_{z_*}$ as in (E). There exist a neighborhood $\tilde{\mathcal{V}}_{z_*}$, $\tilde{\mathcal{V}}_{z_*} \subset \mathcal{U}_{z_*}$ and a number $\tilde{T}_{z_*} \in (0, 1)$ such that for any initial condition $z_0 \in \tilde{\mathcal{V}}_{z_*}$ every trajectory of (3, 1), (3.2) remains in $\mathcal{U}_{z_*}$ over the integral $[0, \tilde{T}_{z_*}]$. Integrating by parts the integral form of (3.1) we obtain the following expression

for $z(t)$:

$$z(t)=z_0+P_0(z_0)\cdot t+\cdots+C_{p-2}(z_0)\cdot P_0(z_0)\frac{t^{p-1}}{(p-1)!}+C_{p-1}(z_0)P_0(z_0)\frac{t^p}{p!}$$
$$+\int_0^t\Big(f(z(\tau),u(\tau),v(\tau))+C_1(z(\tau))f(z(\tau),u(\tau),v(\tau))(t-\tau)+\cdots$$
$$+C_{p-2}(z(\tau))f(z(\tau)u(\tau),v(\tau))\frac{(t-\tau)^{p-2}}{(p-2)!}\Big)\,d\tau$$
$$+\int_0^t C_{p-1}(z(\tau))f(z(\tau),u(\tau),v(\tau))\,\frac{(t-\tau)^{p-1}}{(p-1)!}$$
$$+\int_0^t C_p(z(\tau))(P_0(z(\tau))+f(z(\tau),u(\tau),v(\tau)))\frac{(t-\tau)^p}{p!}d\tau.$$

Because of (E) the projection $\Pi z(t)$ of the trajectory into $L=L(z_*)$ takes the form

$$\Pi z(t)=w_p(z_0;t)$$
$$+\int_0^t \Pi C_{p-1}(z(\tau))f(z(\tau),u(\tau),v(\tau))\,\frac{(t-\tau)^{p-1}}{(p-1)!}\,d\tau+R(t^{p+1}),\tag{3.3}$$

where

$$R(t^{p+1})=\int_0^t \Pi C_p(z(\tau))(P_0(z(\tau))+f(z(\tau),u(\tau),v(\tau)))\frac{(t-\tau)^p}{p!}\,d\tau$$

satisfies

$$|R(t^{p+1})|\leqq N_{z_*}t^{p+1}\quad\text{for } t\in[0,\tilde{T}_{z_*}]\tag{3.4}$$

for some constant N_{z_*}, $w_p(z_0;t)=\Pi z_0+\Pi P_0(z_0)t+\cdots+\Pi C_{p-1}(z_0)P_0(z_0)(t^p/p!)$ is a curve whose components are polynomials of degrees not greater than p.

(B) Take a square $Q\subset L$ and an integer p. There exists a positive constant θ such that for any curve $w_p(t)$ in L whose components are polynomials of degrees not greater than p there exists a point $w_0\in Q$ such that

$$|w_p(t)+w_0t^p|\geqq\theta t^p\quad\text{for all } t\in[0,+\infty).\tag{3.5}$$

This assertion is contained in a more general statement in [1]. An outline of the proof is as follows. Let $L=R^2$ and Q be a square whose sides are parallel to the axes. Divide $-Q$ by lines parallel to the axes into r^2 squares whose interiors are mutually disjoint. Consider the curve $(1/t^p)w_p(t)=((1/t^p)w_p^1(t),(1/t^p)w_p^2(t))$. If r is large enough, namely if $r>2p+1$, then there exists at least one among the small squares whose interior is not intersected by the curve. Suppose that the curve intersects all squares. Then there exists a line parallel to the first axis or a line parallel to the second axis, say a line parallel to the second axis, which is intersected by the curve at least $\frac{1}{2}(r-1)$ times, that is more than p times. It means that for some c the function $(1/t^p)w_p^1(t)-c$ has more than p zeros and so does the function $w_p^1(t)-t^pc$. Hence $w_p^1(t)-t^pc\equiv 0$ and the curve remains on one line all the time and so it cannot intersect all squares. We take the center of the square

whose interior is disjoint with the curve as $-w_0$. The inequality (3.5) holds with $\theta = (l/2r)$ where l denotes the length of a side of the square Q.

(C) We will describe a local maneuvre of evading in a neighborhood of a point $z_* \in M$. By (ii) there is a square Q_{z_*} such that

$$Q_{z_*} \subset \bigcap_{u \in U} \operatorname{co}\{\Pi C_{p-1}(z_*) f(z_*, u, v) | v \in V\}$$

$$= \bigcap_{u \in U} \{\Pi C_p(z_*) \sum_{i=1}^{n+1} \mu_i f(z_*, u, v_i) | (\mu_1, \cdots, \mu_{n+1}) \in \Delta, v_i \in V, i = 1, \cdots, n+1\}. \tag{3.6}$$

Take θ_{z_*} as in the section B for $Q = (1/p!)Q_{z_*}$. Choose neighborhoods $\mathcal{V}_{z_*}$, $\tilde{\mathcal{V}}_{z_*}$ and a number $\tilde{T}_{z_*} \in (0, \tilde{T}_{z_*})$ such that if $z_0 \in \mathcal{V}_{z_*}$ then any trajectory of (3.1) and (3.2) remains over the interval $[0, \tilde{T}_{z_*}]$ in the neighborhood $\tilde{\mathcal{V}}_{z_*}$ which has the following property:

$$|\Pi C_{p-1}(z) f(z, u, v) - \Pi C_{p-1}(z_*) f(z_*, u, v)| \leqq \frac{\theta_{z_*}}{2} \quad \text{for } z \in \tilde{\mathcal{V}}_{z_*} \subset \tilde{\mathcal{V}}_{z_*} \text{ and all } (u, v) \in U \times V. \tag{3.7}$$

Let $z_0 \in \mathcal{V}_{z_*}$. Denote by $w_{z_0}, w_{z_0} \in Q$ the point corresponding as in (B) to the curve $w_p(z_0; t)$. We have from (3.6) that if $w \in Q_{z_*}$ then for each $u \in U$ there exists $\tilde{v}(u) = (\mu_1(u), \cdots, \mu_{n+1}(u), v_1(u), \cdots, v_{n+1}(u)) \in \tilde{V}$ such that

$$\Pi C_{p-1}(z_*) \sum_{i=1}^{n+1} \mu_i(u) f(z_*, u, v_i) = w. \tag{3.8}$$

Fix a basis ξ in $R^{(n+1)+(n+1)q}$ and choose from the set of all solutions $\tilde{v}(u) \in \tilde{V}$ of (3.8) the lexicographical maximum $\tilde{v}_0(u)$ with respect to the basis ξ. We prove as in the well-known Fillipov's lemma that $\tilde{v}_0(u(t))$ is measurable for any measurable $u(t)$. Therefore for each $z_0 \in \mathcal{V}_{z_*}$ we have a strategy $\tilde{v}^u_{z_*}(z_0; t) = (\mu^u_1(t), \cdots, \mu^u_{n+1}(t), v^u_1(t), \cdots, v^u_{n+1}(t))$ in the extended game such that for any control function $u(t)$ the following holds:

$$\Pi C_{p-1}(z_*) \sum_{i=1}^{n+1} \mu^u_i(t) f(z_*, u(t), v^u_i(t)) = w_{z_0} p! \tag{3.9}$$

Since we have (3.3), (3.4) which are valid as well for trajectories of the extended game and since (3.5), (3.7), (3.9) hold, we conclude that each trajectory $\tilde{z}(t)$ corresponding to the strategy $\tilde{v}^u_{z_*}(z_0; t)$ satisfies

$$|\Pi \tilde{z}(t)| \geqq \frac{\theta_{z_*}}{2} t^p - N_{z_*} t^{p+1}.$$

Take $\tilde{K}_{z_*}$, T_{z_*} such that $\tilde{K}_{z_*} > 0$, $T_{z_*} \in (0, \tilde{T}_{z_*}]$ and

$$|\Pi \tilde{z}(t)| \geqq \tilde{K}_{z_*} t^p \quad \text{for } t \in [0, T_{z_*}]. \tag{3.10}$$

Take a positive constant C_{z_*} such that $C_{z_*} > (1/T_{z_*})$ and

$$|P_0(z) + f(z, u, v)| < \tfrac{1}{2} C_{z_*} \quad \text{for } z \in \mathcal{U}_{z_*}, \quad (u, v) \in U \times V. \tag{3.11}$$

Hence if $z_0 \in \mathcal{V}_{z_*}$, $t \in [0, T_{z_*}]$, then any trajectory of (3.1) satisfies

$$\rho(z(t), M) \geqq \rho(z_0, M) - t \cdot \frac{C_{z_*}}{2}. \tag{3.12}$$

We can assume that $\rho(z_0, M) \leqq 1$ for $z_0 \in \mathcal{V}_{z_*}$. Therefore we have from (3.12)

$$\rho(z(t), M) \geqq \frac{\rho(z_0, M)}{2} \quad \text{for } t \in \left[0, \frac{\rho(z_0, M)}{C_{z_*}}\right]. \tag{3.13}$$

Choose for each $z_0 \in \mathcal{V}_{z_*}$, $z_0 \notin M$, a positive number $\varepsilon(z_0)$ such that

$$\tilde{K}_{z_*} t^p \geqq \frac{\tilde{K}_{z_*}}{2} t^p + \varepsilon(z_0) \quad \text{for } t \geqq \frac{\rho(z_0, M)}{C_{z_*}}. \tag{3.14}$$

Take $\tilde{v}^u_{z_*}(z_0; t)$ and apply Theorem 2.1 for $\varepsilon = \varepsilon(z_0)$, $T = T_{z_*}$. We obtain a strategy $v^u_{z_*}(z_0; t)$ defined for all $z_0 \in \mathcal{V}_{z_*}$, $z_0 \notin M$ and $t \in [0, T_{z_*}]$ and such that any corresponding trajectory satisfies

$$\rho(z(t), M) \geqq \frac{\tilde{K}_{z_*}}{2} t^p \quad \text{for } t \in \left[\frac{\rho(z_0, M)}{C_{z_*}}, T_{z_*}\right]. \tag{3.15}$$

Moreover, it satisfies (3.13). Take K_{z_*} such that $K_{z_*} < C^p_{z_*}/2$, $K_{z_*} \leqq \tilde{K}_{z_*}/2$. We conclude that any trajectory corresponding to $v^u_{z_*}(z_0; t)$ satisfies

$$\rho(z(t), M) \geqq K_{z_*} t^{p(z_*)}, \qquad \rho(z(t), M) \geqq K_{z_*} \frac{\rho(z_0, M)^{p(z_*)}}{C^{p(z_*)}_{z_*}} \quad \text{for } t \in [0, T_{z_*}]. \tag{3.16}$$

(D) We proceed to construct the sets W, W_1. Denote by $K(0, r)$ the closed ball of radius r and center at the origin. Take a sequence $0 = r_0 < r_1 < \cdots < r_i < r_{i+1} \cdots$ such that for any trajectory $z(t)$ of (3.1) the following condition holds:

(3.17) if $z_0 \in K(0, r_i)$, $t \in [-1, 1]$ then $z(t) \in \operatorname{int} K(0, r_{i+1})$ for $i = 1, 2, \cdots$.

The existence of a such sequence follows from the growth condition (b).

Define $M_i = M \cap K(0, r_i)$ for $i = 1, 2, \cdots$. Each M_i is compact and $\{\mathcal{V}_{z_*}\}_{z_* \in M_i}$ is its open covering. Choose a finite covering $\mathcal{V}^{i,1}_{z_*}, \cdots, \mathcal{V}^{i,m_i}_{z_*}$ of M_i. Define

$$K_1 = \min\{K_{z^{1,1}_*}, \cdots K_{z^{1,m_1}_*}\}, \qquad K_i = \min\{K_{z^{i,1}_*}, \cdots, K_{z^{i,m_i}_*}, K_{i-1}\} \quad \text{for } i = 2, 3, \cdots.$$

Assume $T_0 = 1$, $p_0 = 1$, $C_0 = 1$ and define for $i = 1, 2, \cdots$,

$$T_i = \min\{T_{z^{i,1}_*}, \cdots, T^{i,m_i}_{z_*}, T_{i-1}\}, \qquad p_i = \max\{p_{z^{i,1}_*}, \cdots, p_{z^{i,m_i}_*}, p_{i-1}\},$$
$$C_i = \max\{C_{z^{i,1}_*}, \cdots, C_{z^{i,m_i}_*}, C_{i-1}\}.$$

We have therefore that each of the strategies $v^{u_{i,j}}_{z^{i,j}_*}(z_0; t)$, $j = 1, \cdots, m_i$, ensures the following estimation:

$$\rho(z(t), M) \geqq K_i t^{p_i}, \qquad \rho(z(t), M) \geqq K_i \frac{\rho(z_0, M)^{p_i}}{C^{p_i}_i} \quad \text{for } t \in [0, T_i]. \tag{3.18}$$

Denote by $W(\rho, M)$ the cylinder around M, $W(\rho, M) = \{z \in R^n | \rho(z, M) \leqq \rho\}$. Take a positive ρ_i for each $i = 1, 2, \cdots$, such that

$$W(\rho_i, M) \cap K(0, r_i) \subset \bigcup_{j=1}^{m_i} \mathcal{V}_{z_*^{i,j}}. \tag{3.19}$$

Take a sequence $\sigma_1, \sigma_2, \cdots, \sigma_i, \cdots$ of positive numbers such that

$$\begin{aligned} &\sigma_i \leqq \rho_i; \qquad \sigma_i < K_{i+1} T_{i+1}^{p_{i+1}} \quad \text{for } i = 1, 2, \cdots; \\ &\qquad\qquad\qquad\qquad\qquad\qquad \sigma_i \leqq \sigma_{i-1} \text{ for } i = 2, 3, \cdots. \end{aligned} \tag{3.20}$$

Define

$$W = \bigcup_{i=1}^{\infty} W(\sigma_i, M) \cap K(0, r_i).$$

We proceed to describe a strategy of evasion $v^u(z_0; t)$. Fix an element $\bar{v}$, $\bar{v} \in V$. Take an initial condition z_0, $z_0 \notin W$, and a control function $u(t)$ on $[0, +\infty)$. Put $v^u(z_0; t) \equiv \bar{v}$ as long as the corresponding trajectory $z(t)$ satisfies $z(t) \notin W$. Let t_1 be the first moment that $z(t_1) = z_1 \in W$ and let $|z_1| \in (r_{i-1}, r_i]$. Therefore $z_1 \in W(\sigma_i, M)$ and because of (3.19), (3.20), $z_1 \in \mathcal{V}_{z_*^{i,j}}$ for some j. Denote by j_0 the smallest of such integers. Put $v^u(z_0; t) = v_{z_*^{i,j_0}}^{\tilde{u}}(z_1; t - t_1)$ for $t \in [t_1, t_1 + T_i]$, where $\tilde{u}(t - t_1) = u(t)$. Since $\rho(z_1, M) \geqq \sigma_{i+1}$, we have from (3.18)

$$\rho(z(t), M) \geqq K_i(t - t_1); \qquad \rho(z(t), M) \geqq K_i \frac{\sigma_{i+1}^{p_i}}{C_i^{p_i}}, \qquad t \in [t_1, t_1 + T_i]. \tag{3.21}$$

Hence from (3.20), $\rho(z(t_1 + T_i), M) > K_i T_i^{p_i} > \sigma_{i-1} \geqq \sigma_i \geqq \sigma_{i+1}$ and since (3.17), $|z(t_1 + T_i)| \in (r_{i-2}, r_{i+1})$. Therefore $z(t_1 + T_i) \notin W$. Again we put $v^u(z_0; t) \equiv \bar{v}$ till the next moment t_2 such that $z(t_2) \in W$ when one of the strategies $v_{z_*}^u(z_2, t - t_2)$ is switched on again. Let $z_0 \in W$ now. Assume $|z_0| \in (r_{i-1}, r_i]$ and let j_0 be the smallest of the integers such that $z_0 \in \mathcal{V}_{z_*^{i,j}}$. Put $v^u(z_0; t) = v_{z_*^{i,j_0}}^u(z_0; t)$ for $t \in [0, T_i]$. We have

$$\rho(z(t), M) \geqq K_i t^{p_i}; \qquad \rho(z(t), M) \geqq K_i \frac{\rho(z_0, M)^{p_i}}{C_i^{p_i}} \quad \text{for } t \in [0, T_i]. \tag{3.22}$$

Hence $z(T) \notin W$ and we proceed as before. It may happen that $\sum_{i=1}^{\infty} T_i < +\infty$ but by (3.17), if $|z(t')| \leqq r_i$ and $|z(t'')| > r_{i+1}$ then $|t' - t''| > 1$ and therefore the game proceeds as described above over the whole interval $[0, +\infty)$; in other words, the procedure defines a strategy $v^u(z_0; t)$ for all $t \in [0, +\infty)$. Take

$$\eta_i = \frac{K_{i+1} \sigma_{i+1}^{p_{i+1}}}{2 C_{i+1}^{p_{i+1}}}$$

and define $W_1 = \bigcup_{i=1}^{\infty} W(\eta_i, M) \cap K(0, r_i)$. Notice that $\sigma_{i+1} > \eta_i \geqq \eta_{i+1}$. Take a trajectory corresponding to the strategy $v^u(z_0; t)$ and assume that $z(t_1) \notin W$. Let $t \geqq t_1$, $|z(t)| \in (r_{i-1}, r_i]$. Then, either $|z(t)| \geqq \sigma_{i+1} > \eta_i$ and hence $z(t) \notin W_1$, or $|z(t)| < \sigma_{i+1}$. The latter implies that $z(t) \in \operatorname{int} W$ and the trajectory is on the course

of action of a local maneuvre of evading which began at some earlier moment at a point $z_2 \in \partial W$, $|z_2| \in (r_{i-2}, r_{i+1})$. Therefore from (3.22)

$$\rho(z(t), M) \geqq K_{i+1} \frac{\sigma_{i+1}^{p_{i+1}}}{C_{i+1}^{p_{i+1}}} > \eta_i$$

and hence $z(t) \notin W_1$. Define: $T(\xi) = T_i$ for $\xi \in (r_{i-1}, r_i]$, $i = 1, 2, \cdots$,

$$\gamma(\xi_1, \xi_2) = \frac{K_i}{C_i^{p_i}} \xi_1^{p_i} \quad \text{for } \xi_1 \in (0, +\infty), \quad \xi_2 \in (r_{i-1}, r_i], \quad i = 1, 2, \cdots.$$

The functions $T(\xi)$, $\gamma(\xi_1, \xi_2)$ have properties required in the assertion of the Theorem 3.1. and this completes the proof.

Remarks. It is required in condition (i) that all functions $\Pi C_r(z) f(z, u, v)$, $r = 0, \cdots, p-2$ vanish in a neighborhood $\mathcal{U}_{z_*}$ of a point $z_* \in M$. Notice that this condition can be easily replaced by the assumption that the functions $\Pi C_r(z) f(z, u, v)$ vanish along the subspace M and grow up slowly enough withdrawing M. Namely, if we assume that for some constant F_{z_*} there is

$$|\Pi C_r(z) f(z, u, v)| \leqq F_{z_*} \rho(z, M)^{p-r} \quad \text{for } z \in \mathcal{U}_{z_*}, \quad r = 0, \cdots, p-2, \tag{3.23}$$

then for $z_0 \in \tilde{\mathcal{V}}_{z_*}$, $t \in [0, \tilde{T}_{z_*}]$ the expression

$$B(t) = \int_0^t \Big(\Pi f(z(\tau), u(\tau), v(\tau)) + \Pi C_1(z(\tau)) f(z(\tau), u(\tau), v(\tau))(t-\tau) + \cdots + \Pi C_{p-2}(z(\tau)) f(z(\tau), u(\tau), v(\tau)) \frac{(t-\tau)^{p-2}}{(p-2)!} \Big) d\tau$$

can be estimated together with $R(t^{p+1})$. Indeed, we have

$$|B(t)| \leqq F_{z_*} \sum_{i=1}^{p-1} t^i (\rho(z_0 M) + D_{z_*} t)^{p+1-i} \quad \text{for } t \in [0, \tilde{T}_{z_*}],$$

where $D_{z_*} = \sup \{|P_0(z) + f(z, u, v)| \, | \, z \in \mathcal{U}_{z_*}, (u, v) \in U \times V\}$. Therefore

$$|B(t)| \leqq \bar{N}_{z_*} t^{p+1} \quad \text{for } t \geqq \frac{\rho(z_0 M)}{C_{z_*}}.$$

It suffices since for $t \leqq \rho(z_0 M)/C_{z_*}$ we use the estimation $\rho(z(t), M) \geqq \rho(z_0 M)/2$ (see (3.13)). Notice, furthermore, that instead of considering a division of $P(z, u, v)$ into a sum $P(z, u, v) = P_0(z) + f(z, u, v)$ we can consider a division $\Pi(z_*) P(z, u, v) = P_1(z) + f_1(z, u, v)$ of the projection of $P(z, u, v)$ into the subspace $L(z_*)$ and assume that the functions $P_1(z), f_1(z, u, v)$ have properties corresponding to the properties of $\Pi P_0(z)$, $\Pi f(z, u, v)$ required in the conditions

(E). In this way the division depends on point $z_* \in M$. Namely the conditions (E) take the following form (compare [6]): for each $z_* \in M$ there exists a two-dimensional subspace $L(z_*)$ orthogonal to M, a neighborhood $\mathcal{U}_{z_*}$, and an integer p_{z_*} such that the projection $\Pi(z_*)P(z, u, v)$ of the right hand side into $L(z_*)$ can be written in the following form:

$$\Pi(z_*)P(z, u, v) = P_1(z) + f_1(z, u, v),$$
$$DP_i \cdot P(z, u, v) = P_{i+1}(z) + f_{i+1}(z, u, v), \qquad i = 1, 2, \cdots, p_{z_*} - 1$$

where the functions $f_i(z, u, v)$ for $i = 1, \cdots, p_{z_*} - 1$ vanish in $\mathcal{U}_{z_*}$ or are estimated as in (3.23) and the function $f_{p_{z_*}}(z, u, v)$ is such that the set $\bigcap_{u \in U} \operatorname{co} f_{p_{z_*}}(z_*, u, V)$ contains an interior point with respect to $L(z_*)$.

We have considered only the autonomous case. The nonautonomous case is not essentially different. The problem of possibility of evasion for each initial condition (t_0, z_0) for the game $\dot{z} = P(t, z, u, v)$, $t \in R$, $z \in R^n$, is equivalent to the following autonomous evasion problem in R^{n+1}: $\dot{\tilde{z}} = \tilde{P}(\tilde{z}, u, v)$, $\tilde{z} = (z^0, z) \in R^{n+1}$, $\tilde{P}(\tilde{z}, u, v) = (1, P(z^0, z, u, v))$, $\tilde{M} = R \times M$. Evasion conditions of such type as the conditions (E) and a division of the right hand side into a sum $P(z, u, v) = P_0(z) + f(z, u, v)$ appear in a natural way if, for example, one considers an evasion game between two objects x, y in R^m whose motions are described by equations of orders p and q, respectively:

$$x^{(p)} = F(x, x^{(1)}, \cdots, x^{(p-1)}, v), \qquad v \in V,$$

$$y^{(q)} = G(y, y^{(i)}, \cdots, y^{(q-1)}, u), \qquad u \in U.$$

Consider the corresponding game in $R^{(p+q)m}$; that is, take

$$z = (x, x^{(1)}, \cdots, x^{(p-1)}, y, \cdots, y^{(q-1)}) = (z_1, \cdots, z_{p+q}) \in R^{(p+q)m},$$
$$M = \{z \in R^{(p+q)m} | z_1 = z_{p+1}\},$$

and the corresponding system of equations $\dot{z} = P_0(z) + f(z, u, v)$, where

$$P_0(z) = (z_2, z_3, \cdots, z_p, 0, z_{p+2}, \cdots, z_{p+q}, 0),$$
$$f(z, u, v) = (0, \cdots, 0, F(z_1, \cdots, z_p, v), 0, \cdots, G(z_{p+1}, \cdots, z_{p+q}, u)).$$

In order to ensure the possibility of evasion for each initial state it is necessary to assume that the evader x has a sort of superiority. The following two conditions (e_1), (e_2) are in a sense natural conditions of superiority:

(e_1) $q < p$ and for each point $(x, x^{(1)}, \cdots, x^{(p-1)})$ there exists a two dimensional subspace $L = L(x, \cdots, x^{(p-1)})$ of R^n such that

$$\operatorname{int}_L \operatorname{co} \Pi F(x, x^{(1)}, \cdots, x^{(p-1)}, V) \neq \varnothing,$$

where by int_L we denote interior with respect to L, and Π is the orthogonal projection onto L.

(e_2) $q = p$ and for each pair of points $(x, x^{(1)}, \cdots, x^{(p-1)})$, $(y, y^{(1)}, \cdots, y^{(q-1)})$ such that $x = y$ there exists a two-dimensional subspace L of R^n such that for some $w_o \in L$,

$$w_0 + \Pi G(y, \cdots, y^{q-1}, U) \subset \operatorname{int}_L \operatorname{co} \Pi F(x, \cdots, x^{(q-1)}, V).$$

One can easily check by computing $C_r(z)f(z, u, v)$ that each of the conditions (e_1), (e_2) implies the condition (E).

Acknowledgments. I wish to express my gratitude to Professor Czesław Olech; without his assistance and encouragement this work could not have been done.

REFERENCES

[1] L. S. PONTRYAGIN, *A linear differential game of escape*, Trudy Mat. Inst. Steklov., 112 (1971), pp. 30–63.

[2] L. S. PONTRYAGIN AND E. F. MISHCHENKO, *On a problem of avoidance in linear differential games*, Differencial'nye Uravnenija, 7 (1971), pp. 436–445.

[3] E. F. MISHCHENKO, *On the problem of evading the encounter in differential games*, this Journal, 12 (1974), pp. 300–310.

[4] R. V. GAMKRELIDZE AND G. L. KCHARATISHVILI, *A differential game of evasion with nonlinear control*, this Journal, 12 (1974), pp. 332–349.

[5] M. S. NIKOLSKI, *On a quasi-linear problem of evading*, Dokl. Akad. Nauk SSSR, 221 (1975), pp. 539–543.

[6] P. B. GUSATNIKOV, *On a problem of l-evading*, Prikl. Mat. Meh., 40 (1976), pp. 25–47.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

STEEPEST DESCENT WITH RELAXED CONTROLS*

J. WARGA†

Abstract. We prove the convergence of a steepest descent iterative procedure for determining an "extremal" point of a function defined on a sequentially compact convex subset of a topological vector space. We then apply this procedure to the problem of determining an extremal of a relaxed optimal control problem defined by ordinary differential equations without endpoint or unilateral restrictions.

Let K be a sequentially compact convex subset of a (real) topological vector space $\mathscr{X}$ and $\phi: K \to \mathbb{R}$ a continuous function such that the directional derivative

$$D\phi(x; y-x) \triangleq \lim_{\alpha \to 0+} \frac{1}{\alpha}[\phi(x+\alpha[y-x]) - \phi(x)]$$

exists for all x, $y \in K$ and the family of functions

$$x \to D\phi(x; y-x): K \to \mathbb{R},$$

corresponding to all $y \in K$, is equicontinuous. [It is easy to see that if, moreover, the topology of $\mathscr{X}$ is metric and ϕ is Lipschitz continuous then the function

$$(x, y) \to D\phi(x; y-x): K \times K \to \mathbb{R}$$

is continuous.]

In its simplest form (which we shall apply to optimal control) our procedure is a very "natural" form of steepest descent: if the function $(x, y) \to D\phi(x; y-x)$ is continuous, we choose an arbitrary $x_0 \in K$ and determine $x_1, x_2, \cdots$ iteratively as follows: given x_i, we determine any y_i that minimizes $y \to D\phi(x_i; y-x_i)$ on K. We then determine (by a search on [0, 1]) a number $\theta_i \in [0, 1]$ that minimizes $\theta \to \phi(x_i + \theta(y_i - x_i))$, and choose as x_{i+1} any point such that

$$\phi(x_{i+1}) \leqq \phi(x_i + \theta_i(y_i - x_i)).$$

In a more general way, we apply

Procedure A. We choose some $c \in (0, 1)$ and $x_0 \in K$. Given x_i, we determine some y_i such that

$$D\phi(x_i; y_i - x_i) \leqq c \max\{-1, \inf_{y \in K} D\phi(x_i; y - x_i)\}$$

(observe that $\inf_{y \in K} D\phi(x_i; y-x_i) \leqq D\phi(x_i; x_i - x_i) = 0$), then choose some $\theta_i \in [0, 1]$ such that

$$\phi(x_i + \theta_i(y_i - x_i)) - \phi(x_i) \leqq c\left[\min_{\theta \in [0, 1]} \phi(x_i + \theta(y_i - x_i)) - \phi(x_i)\right],$$

and finally select as x_{i+1} any point in K such that

$$\phi(x_{i+1}) \leqq \phi(x_i + \theta_i(y_i - x_i)).$$

* Received by the editors July 2, 1976, and in revised form August 30, 1976.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115. This work was supported in part by the National Science Foundation under Grant MPS 73-08532A02.

We shall later prove

THEOREM I. *Let* $\{x_0, x_1, x_2, \cdots\}$ *be constructed by Procedure* A *and* x_∞ *be the limit of some subsequence of* $(x_0, x_1, \cdots)$. *Then*

$$D\phi(x_\infty; y - x_\infty) \geqq 0 \qquad (y \in K).$$

The above relation is a necessary condition for x_∞ to minimize ϕ. We refer to a point satisfying this condition as an *extremal* point of ϕ. We observe that in the special case where ϕ is a convex function, the point x_∞ of Theorem I actually minimizes ϕ on K, and we obtain a useful estimate of the error at the ith step by applying

THEOREM II. *If* ϕ *is convex and* $x_0, x_1, \cdots$ *are as described in Theorem* I *then*

$$\phi_{\min} \triangleq \min \phi(K) \leqq \phi(x_i) \leqq \phi_{\min} - \inf_{y \in K} D\phi(x_i; y - x_i) \qquad (i = 0, 1, 2, \cdots)$$

and

$$\lim_i \inf_{y \in K} D\phi(x_i; y - x_i) = 0.$$

If $\mathscr{X}$ is a Hilbert space, $b \in \mathscr{X}$ and $\phi(x) = |x - b|^2$ then Procedure A determines the nearest point in K to b and Theorems I and II yield results previously obtained by Gilbert [2] for the case $\mathscr{X} = \mathbb{R}^n$ (and related to the algorithm of Frank and Wolfe [1]).

If ϕ is not convex then Procedure A yields an extremal point of ϕ but not necessarily a minimizing one. The question then arises—and it was posed by the referee—about the range of applicability of this procedure in "nonconvex" problems (and, in particular, in optimal control problems involving nonconvex cost functionals or nonlinear differential equations). A second question concerns the possibility of implementing this procedure. To both of these questions our answer will be subject to many "if's" and "maybe's" and may therefore appear unsatisfactory. However, we must point out that no method presently exists—or will likely ever exist—for determining a minimizing point of an arbitrary continuous, or even "differentiable", function defined over a compact and convex set. This remark even applies to functions over a compact real interval. Thus a method of the type under investigation must be considered as a general scheme whose practical value and power can only be tested for special classes of applications. The same remark applies to "implementation". In Procedure A, in order to find an extremal point of ϕ, we must at each step come within a certain "range" of the minimum of a subsidiary function, namely $y \to D\phi(x_i, y - x_i)$. As we shall see later in the special case of optimal control problems, the determination of the minimum of the latter function, defined over a convex and compact subset of a normed vector space, can be reduced to the problem of minimizing functions over control sets that are usually finite-dimensional. This appears to be—and often is—a great simplification. However, no general assertions appear possible even for such "finite-dimensional" problems. Furthermore, all the objections that may be directed at a "steepest descent" procedure like ours are even more applicable to all the indirect methods of the calculus of variations, of optimal control and of optimization that are based on necessary (but not sufficient) conditions for

minimum. Thus, in summary, "beggars can't be choosers" as is evident by considering the applicability of most approximation methods designed for difficult mathematical problems.

We shall now indicate a few applications of Procedure A (and of its optimal control version, Procedure B below) that come immediately to mind: (1) as will be apparent from the proof of Theorem I, Procedure A yields a lower value of ϕ with each iteration and can be used therefore to improve on known admissible results; (2) if one can determine from the known structure of the problem (or by physical arguments) a finite number of "small regions" that contain all the extremal points of ϕ then Procedure A, with suitable initial guesses in those "regions", might yield all the extremal points and therefore also all the minimizing ones; (3) we have recently derived, in [6], certain "second order" conditions stronger than "extremality" and applied them to optimal control. Some preliminary results indicate that these conditions may perhaps be used "constructively" to find an argument of ϕ that yields a lower value than a given extremal point which violates these "second order" conditions. Thus Procedure A, in conjunction with these auxiliary constructions, may yield several extremal points, each of them better than the preceding one.

We next describe a class of relaxed optimal control problems to which we shall apply the interative procedure. Let $T \triangleq [t_0, t_1] \subset \mathbb{R}$, V be an open subset of $\mathbb{R}^n$, A_0 a convex compact subset of V, R a compact metric space, $\mathcal{K}(R)$ the collection of closed nonempty subsets of R with the Hausdorff metric, and $R^{\#}: T \to \mathcal{K}(R)$ a measurable set-valued mapping [5, I.7, p. 146]. We define $\mathcal{R}^{\#}$ as the collection of all measurable selections of $R^{\#}(\cdot)$ and the space $\mathcal{S}^{\#}$ of *relaxed control functions* as the compact metric space of all measurable $\sigma: T \to \text{rpm}\,(R)$ such that $\sigma(t)(R^{\#}(t)) = 1$ a.e., where rpm (R) is the set of all Radon probability measures on R with the relative weak star topology of $C(R)^*$ (see [5, Chap. IV] for details) and the topology of $\mathcal{S}^{\#}$ is the relative weak star topology of $L^1(T, C(R))^*$ [5, pp. 272 and 287]. For any continuous $\phi: R \to \mathbb{R}^n$, we write $\phi(\sigma(t))$ for $\int \phi(r)\sigma(t)(dr)$.

Let $f: T \times V \times R \to \mathbb{R}^n$ and $h_0: V \to \mathbb{R}$ be given functions. We denote by h_0' the derivative (gradient) of h_0 and by f_v the partial derivative of f with respect to its argument in V. We assume that

(a) h_0, h_0', $f(t, \cdot, \cdot)$ and $f_v(t, \cdot, \cdot)$ exist and are continuous for all $t \in T$;

(b) $f(\cdot, v, r)$ is measurable for all $(v, r) \in V \times R$; and

(c) there exist a compact $D \subset V$ and an integrable $\psi: T \to \mathbb{R}$ such that the differential equation

$$y(t) = a + \int_{t_0}^{t} f(\tau, y(\tau), \sigma(\tau))\, d\tau \qquad (t \in T)$$

has a unique solution $y(\sigma, a)$, with $y(\sigma, a)(t) \in D$ for all $\sigma \in \mathcal{S}^{\#}$, $a \in A_0$ and $t \in T$, and we have

$$|f(t, v, r)| \leqq \psi(t), \qquad |f_v(t, v, r)| \leqq \psi(t) \qquad (t \in T, v \in D, r \in R).$$

We set

$$\phi(\sigma, a) \triangleq h_0(y(\sigma, a)(t_1)) \qquad ((\sigma, a) \in \mathcal{S}^{\#} \times A_0).$$

Before describing the procedure in detail, we recall [5, VI.3.2, p. 370] that for every $(\sigma, a) \in \mathcal{S}^{\#} \times A_0$ there exists a "Gamkrelidze control" $\sigma_G \in \mathcal{S}^{\#}$, represented by some $\rho_j \in \mathcal{R}^{\#}$ and measurable $\alpha_j : T \to [0, 1]$ $(j = 0, 1, \cdots, n)$, with

$$\sum_{j=0}^{n} \alpha_j(t) = 1 \qquad (t \in T),$$

$$\sigma_G(t)(\{\rho_j(t)\}) = \alpha_j(t) \qquad (t \in T, j = 0, \cdots, n),$$

and such that

$$y(\sigma_G, a) = y(\sigma, a).$$

We denote by $\mathcal{S}_G^{\#}$ the collection of Gamkrelidze controls and say that $(\sigma_G, a) \in \mathcal{S}_G^{\#} \times A_0$ is *f-equivalent* to $(\sigma, a) \in \mathcal{S}^{\#} \times A_0$ if $y(\sigma_G, a) = y(\sigma, a)$. In a computational procedure, there may be an advantage in using Gamkrelidze controls because they require the storage of "only" the $2(n+1)$ functions ρ_j, α_j instead of a measure-valued function σ.

Finally, we ought to mention at this point that Procedure B, which we describe below, bears a certain relation to some algorithms investigated by Mayne and Polak. In a few of them [3], use is made of a function defined like $\bar{\rho}$ of Procedure B and, in another [4], the authors consider convergence in the topology of relaxed controls.

Procedure B (for the determination of an extremal in "free" problems of optimal control).

Let $(\sigma_0, a_0) \in \mathcal{S}_G^{\#} \times A_0$ be arbitrary. If $(\sigma_i, a_i) \in \mathcal{S}_G^{\#} \times A_0$ is known, we determine (σ_{i+1}, a_{i+1}) as follows: we set $y_i \triangleq y(\sigma_i, a_i)$ and determine the unique absolutely continuous solution $z : T \to \mathbb{R}^n$ of

$$z(t)^T = h_0(y_i(t_i)) + \int_t^{t_1} z(\tau)^T f_v(\tau, y_i(\tau), \sigma_i(\tau))\, d\tau \qquad (t \in T),$$

where the superscript ^{T} denotes transposition or a row vector. We next determine any $\bar{a} \in A_0$ that minimizes

$$a \to z(t_0)^T a$$

over A_0 and any $\bar{\rho} \in \mathcal{R}^{\#}$ such that

$$z(t)^T f(t, y_i(t), \bar{\rho}(t)) = \min_{r \in R^{\#}(t)} z(t)^T f(t, y_i(t), r) \quad \text{a.e. } in\ T.$$

If

$$z(t_0)^T a_i = \min_{a \in A_0} z(t_0)^T a \quad \text{and} \quad z(t)^T f(t, y_i(t), \sigma_i(t)) = z(t)^T f(t, y_i(t), \bar{\rho}(t))$$

a.e. in T,

we set $(\sigma_j, a_j) = (\sigma_i, a_i)$ $(j = i+1, i+2, \cdots)$ and terminate the procedure. Otherwise, we observe that for all $\theta \in [0, 1]$ we can solve the equation

$$y(t) = (1-\theta)\left[a_i + \int_{t_0}^{t} f(\tau, y(\tau), \sigma_i(\tau))\, d\tau\right] + \theta\left[\bar{a} + \int_{t_0}^{t} f(\tau, y(\tau), \bar{\rho}(\tau))\, d\tau\right] \qquad (t \in T)$$

to obtain the unique solution $\bar{y}(\theta)(t)$. We determine (by a search on $[0, 1]$) a number $\theta_i \in [0, 1]$ that minimizes $\theta \to h_0(\bar{y}(\theta)(t_1))$. Finally, we set $a_{i+1} = (1-\theta_i)a_i + \theta_i \bar{a}$ and choose as σ_{i+1} any element of $\mathcal{S}^{\#}_{G}$ such that

$$h_0(y(\sigma_{i+1}, a_{i+1})(t_1)) \leqq h_0(\bar{y}(\theta_i)(t_1)).$$

In particular, we may choose σ_{i+1} so that

$$y(\sigma_{i+1}, a_{i+1}) = y((1-\theta_i)\sigma_i + \theta_i \delta_{\bar{\rho}}, a_{i+1}),$$

where $\delta_{\bar{\rho}}(t)$ is the Dirac measure concentrated at $\bar{\rho}(t)$.

Remark. For the sake of simplicity and clarity, we have adapted Procedure A in its simplest form to optimal control. However, Theorem III below, and its proof, remain valid if we adapt Procedure A in its general form. This can be done by choosing a priori a number $c \in (0, 1]$ and, at the ith step, computing $(\bar{\rho}, \bar{a}) \in \mathcal{R}^{\#} \times A_0$ and $\theta_i \in [0, 1]$ so that

$$z(t_0)^T \bar{a} + \int_{t_0}^{t_1} z(\tau)^T f(\tau, y_i(\tau), \bar{\rho}(\tau))\, d\tau \leqq c \max\left\{-1, \min_{a \in A_0} z(t_0)^T a + \int_{t_0}^{t_1} \left[\min_{r \in R^{\#}(\tau)} z(\tau)^T f(\tau, y_i(\tau), r)\right] d\tau\right\}$$

and

$$h_0(\bar{y}(\theta_i)(t_1)) - h_0(y_i(t_1)) \leqq c \min_{\theta \in [0, 1]} [h_0(\bar{y}(\theta)(t_1)) - h_0(y_i(t_1))],$$

everything else remaining the same.

THEOREM III. *Let $(\sigma_\infty, a_\infty)$ be the limit in (the compact metric space) $\mathcal{S}^{\#} \times A_0$ of some subsequence of $((\sigma_i, a_i))$. Then $(\sigma_\infty, a_\infty)$ is an extremal of the optimal control problem, i.e.*

$$z_\infty(t_0) a_\infty = \min_{a \in A_0} z_\infty(t_0) a,$$

$$z_\infty(t)^T f(t, y_\infty(t), \sigma_\infty(t)) = \min_{r \in R^{\#}(t)} z_\infty(t)^T f(t, y_\infty(t), r) \quad \text{a.e. in } T,$$

where

$$y_\infty \triangleq y(\sigma_\infty, a_\infty)$$

and

$$z_\infty(t)^T = h_0'(y_\infty(t_1)) + \int_t^{t_1} z_\infty(\tau)^T f_v(\tau, y_\infty(\tau), \sigma_\infty(\tau))\, d\tau \qquad (t \in T).$$

Examples. The following very simple examples illustrate the use of Procedure B.

Example I.

$$\dot{y}_1(t) = y_2(t)^2 - \int r^2 \sigma(t)(dr), \qquad \dot{y}_2(t) = \int r\sigma(t)(dr),$$

$y_1(0) = y_2(0) = 0$, $T = [0, 1]$, $R^{\#}(t) \equiv R = [-1, 1]$, $h_0(v_1, v_2) = v_1$. (Thus we minimize $y_1(1)$.)

The optimal solution is $\sigma(t) = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$, where δ_r is the Dirac measure at r.

$i = 0$. We choose $\sigma_0 = \delta_0$ yielding $\phi(\sigma_0) = 0$, $z_1(t) \equiv 1$, $z_2(t) \equiv 0$.

We choose $\bar{\rho}(t) \equiv 1$, yielding $\theta_1 = 1$.

$i = 1$. $\sigma_1 = \delta_1$, $\phi(\sigma_1) = -0.66 \cdots$, $z_1(t) \equiv 1$, $z_2(t) \equiv 1 - t^2$, $\bar{\rho}(t) \equiv -1$, $\theta_2 = \frac{1}{2}$.

$i = 2$. $\sigma_2 = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, $\phi(\sigma_2) = -1$, σ_2 optimal.

Example II. The problem is the same example as in I except that

$$\dot{y}_1(t) = y_2(t)^2 + \int r^2 \sigma(t)(dr), \qquad y_2(0) = \tfrac{1}{2}.$$

The optimal solution is the point-valued function

$$\sigma(t) = \delta_{u(t)}, \quad \text{with } u(t) \equiv -\tfrac{1}{2}\frac{\sinh(1-t)}{\cosh 1},$$

yielding optimal $\phi_{\min} = 0.19038$.

$i = 0$. $\sigma_0 = \delta_0$, $\phi(\sigma_0) = 0.25$, $z_1 \equiv 1$, $z_2(t) \equiv 1 - t$, $\bar{\rho}(t) \equiv -\frac{1}{2}(1-t)$, $\theta_1 = 1$.

$i = 1$. If we set $\sigma_1 = (1-\theta_1)\sigma_0 + \theta_1\delta_{\bar{\rho}} = \delta_{\bar{\rho}}$ then $\phi(\sigma_1) = 0.2$.

However, Procedure B allows us to pick any σ_1 for which $\phi(\sigma_1) \leqq \phi((1-\theta_1)\sigma_0 + \theta_1\delta_{\bar{\rho}})$ and, because of the convexity of the problem, this will be the case if we choose $\sigma_1 = \delta_{u_1}$, where

$$u_1 = (1-\theta)\cdot 0 + \theta \cdot \bar{\rho}$$

with θ chosen optimally. This yields $\theta_{\text{optimal}} = \frac{10}{14}$ and $\phi(u_1) = 0.19047$—an improvement over 0.2.

Proof of Theorem I. Assume, by way of contradiction, that there exists $\bar{y} \in K$ such that

$$D\phi(x_\infty; \bar{y} - x_\infty) = -\gamma < 0.$$

Let $\beta \triangleq c \min(1, \gamma/2)$ and let $\{i_1, i_2, \cdots\}$ be such that $\lim_j x_{i_j} = x_\infty$.

We first observe that, for each $i = 0, 1, 2, \cdots$,

$$\phi(x_{i+1}) \leqq \phi(x_i + \theta_i(y_i - x_i)) \leqq \phi(x_i).$$

Thus $\phi(x_{i_{j+1}}) \leqq \phi(x_{i_j})$ $(j = 1, 2, \cdots)$. Since the functions $x \to D\phi(x; y_{i_j} - x)$ $(j = 1, 2, \cdots)$ are equicontinuous, there exist $j_0 \in \{1, 2, \cdots\}$ and $\theta_0 \in (0, \frac{1}{2}]$ such

that, for all $j \geqq j_0$ and $\theta \in [0, \theta_0]$, we have

$$\begin{aligned} &D\phi(x_{i_j}+\theta(y_{i_j}-x_{i_j}); y_{i_j}-[x_{i_j}+\theta(y_{i_j}-x_{i_j})]) \\ &\quad \leqq D\phi(x_{i_j}; y_{i_j}-x_{i_j})+\beta/2 \\ &\quad \leqq c \max[-1, D\phi(x_{i_j}; \bar{y}-x_{i_j})]+\beta/2 \\ &\quad \leqq c \max[-1, D\phi(x_\infty; \bar{y}-x_\infty)+\alpha/2]+\beta/2 \\ &\quad \leqq -\beta/2 < 0. \end{aligned}$$

If we set $\psi_j(\theta) \triangleq \phi(x_{i_j}+\theta(y_{i_j}-x_{i_j}))$ then above relation implies that

$$\begin{aligned} \lim_{\alpha \to 0+} \frac{1}{\alpha}[\psi_j(\theta+\alpha)-\psi_j(\theta)] &= \frac{1}{1-\theta} D\psi_j(\theta; 1-\theta) \\ &= \frac{1}{1-\theta} D\phi(x_{i_j}+\theta(y_{i_j}-x_{i_j}); (1-\theta)(y_{i_j}-x_{i_j})) \\ &\leqq -\beta/2 \qquad (\theta \leqq \theta_0, j \geqq j_0). \end{aligned}$$

It follows that $\psi_j(\theta_0) \leqq \psi_j(0)-\theta_0\beta/2$, that is,

$$\phi(x_{i_j}+\theta_0(y_{i_j}-x_{i_j})) \leqq \phi(x_{i_j})-\theta_0\beta/2 \qquad (j \geqq j_0);$$

hence

$$\begin{aligned} \phi(x_{i_{j+1}}) \leqq \phi(x_{i_j+1}) &\leqq \phi(x_{i_j}+\theta_{i_j}(y_{i_j}-x_{i_j})) \\ &\leqq (1-c)\phi(x_{i_j})+c\phi(x_{i_j}+\theta_0(y_{i_j}-x_{i_j})) \\ &\leqq (1-c)\phi(x_{i_j})+c[\phi(x_{i_j})-\theta_0\beta/2] \\ &= \theta(x_{i_j})-c\theta_0\beta/2 \qquad (j \geqq j_0). \end{aligned}$$

Thus $\lim_j \phi(x_{i_j}) = -\infty$, which is absurd because the continuous function ϕ must be bounded on the sequentially compact set K. Q.E.D.

Proof of Theorem II. Since ϕ is convex, we have

$$\frac{1}{\theta}[\phi(x+\theta(y-x))-\phi(x)] \leqq \phi(y)-\phi(x) \qquad (x, y \in K, \theta \in [0, 1]),$$

hence

$$D\phi(x; y-x) \leqq \phi(y)-\phi(x)$$

and

$$\inf_{y \in K} D\phi(x; y-x) \leqq \min \phi(K)-\phi(x) \qquad (x \in K).$$

Now assume that there exist $\varepsilon > 0$ and sequences $J \subset (1, 2, \cdots)$ and $(y_j)_{j \in J}$ in K such that $D\phi(x_j; y_j-x_j) < -\varepsilon$ for $j \in J$. We may assume that $(x_j)_{j \in J}$ converges to some $\bar{x}$ in the sequentially compact set K and it follows, by Theorem I, that

$$D\phi(\bar{x}; y_j-\bar{x}) \geqq 0 \qquad (j \in J);$$

hence, by the equicontinuity of $x \to D\phi(x; y_j - x)$, we have $D\phi(x_j; y_j - x_j) \geqq -\varepsilon/2$ for sufficiently large j in J, a contradiction. Q.E.D.

Proof of Theorem III. Let $L^1(T, C(R))$ be the Banach space of Lebesgue integrable functions $\phi: T \to C(R)$ with the usual norm, and let $\mathcal{N}$ denote its conjugate space $L^1(T, C(R))^*$ with a "weak" norm $|\cdot|_w$ [5, IV.1.9, p. 272], defined by

$$|\nu|_w \triangleq \sum_{j=1}^{\infty} 2^{-j} \frac{|\nu(\phi_j)|}{1+|\phi_j|},$$

where $\{\phi_1, \phi_2, \cdots\}$ is some dense denumerable subset of $L^1(T, C(R))$. We recall [5, IV.3.11, p. 287] that $\mathcal{S}^{\#}$ can be identified with a compact convex subset of $(\mathcal{N}, |\cdot|_w)$. We set

$$\mathcal{X} \triangleq \mathcal{N} \times \mathbb{R}^n, \quad K \triangleq \mathcal{S}^{\#} \times A_0, \quad \phi(\sigma, a) \triangleq h_0(y(\sigma, a)(t_1)).$$

For any $(\sigma, a), (\nu, b) \in K$, we have

$$D\phi((\sigma, a); (\nu, b) - (\sigma, a)) = h_0'(y(\sigma, a)(t_1))\eta(t_1),$$

where $\eta: T \to \mathbb{R}^n$ is defined by

$$\eta(t) = b - a + \int_{t_0}^{t} [f_v(\tau, y(\sigma, a)(\tau), \sigma(\tau))\eta(\tau) + f(\tau, y(\sigma, a)(\tau), \nu(\tau) - \sigma(\tau))]\, d\tau \qquad (t \in T).$$

It follows easily that

$$D\phi((\sigma, a); (\nu, b) - (\sigma, a)) = z(t_0)^T(b - a) + \int_{t_0}^{t_1} z(\tau)^T f(\tau, y(\sigma, a)(\tau), \nu(\tau) - \sigma(\tau))\, d\tau, \tag{1}$$

where $z: T \to \mathbb{R}^n$ is the solution of

$$z(t)^T = h_0'(y(\sigma, a)(t_1)) + \int_{t}^{t_1} z(\tau)^T f_v(\tau, y(\sigma, a)(\tau), \sigma(\tau))\, d\tau \quad (t \in T).$$

We can verify by standard techniques (such as in, e.g., [5, VI.1.1, p. 348]) that $\phi: K \to \mathbb{R}$ and the function

$$((\sigma, a), (\nu, b)) \to D\phi((\sigma, a); (\nu, b) - (\sigma, a)): K \times K \to \mathbb{R}$$

are continuous. Thus the assumptions of Theorem I are satisfied. Relation (1) now shows that for any $(\sigma_i, a_i) \in K$ the element $(\bar{\rho}, \bar{a}) \in K$ minimizes the function

$$(\nu, b) \to D\phi((\sigma_i, a_i); (\nu, b) - (\sigma_i, a_i)).$$

Thus Procedure B is a special case of Procedure A, and our conclusion follows from Theorem I and relation (1). Q.E.D.

Acknowledgement. I wish to acknowledge stimulating discussions with Professor E. Polak about algorithms for solving optimal control problems.

REFERENCES

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

[2] E. G. GILBERT, *An iterative procedure for computing the minimum of a quadratic form on a convex set*, this Journal, 4 (1966), pp. 61–80.

[3] D. Q. MAYNE AND E. POLAK, *First-order strong variation algorithms for optimal control*, J. Optimization Theory Appl., 16 (1975), pp. 277–301.

[4] E. POLAK AND D. Q. MAYNE, *A feasible directions algorithm for optimal control problems with control and terminal inequality constraints*, Preprint.

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

[6] ———, *A second order condition that strengthens Pontryagin's maximum principle*, submitted for publication.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

NECESSARY OPTIMALITY CONDITIONS WITH APPLICATION TO A VARIATIONAL PROBLEM*

NORBERT CHRISTOPEIT†

Abstract. For an abstract optimization problem with operator constraints and vector-valued objective function, necessary optimality conditions of Kuhn–Tucker type are developed, and are then used to obtain a generalized Euler–Lagrange condition for the problem of Bolza with phase restrictions in the nonconvex and nondifferentiable case.

Introduction. Consider the following abstract optimization problem:

$$\text{(P)} \qquad \begin{aligned} &\text{minimize } F(x) \\ &\text{subject to} \\ &\qquad G(x) \leqq_Y 0, \\ &\qquad H(x) = 0, \\ &\qquad x \in X_0, \end{aligned}$$

where

(i) F, G and H are operators defined on a locally convex space X with values in locally convex spaces W, Y and Z, respectively;

(ii) the pre-ordering "$\leqq_Y$" in Y is defined by

$$y \leqq_Y 0: \Leftrightarrow \langle y, \lambda_\alpha \rangle \leqq 0 \quad \text{for all } \alpha \in A,$$

where $\Lambda = \{\lambda_\alpha\}_{\alpha \in A}$ is a given family of continuous linear functionals on Y;

(iii) the minimum is meant to be a generalized Pareto-optimum with respect to the pre-ordering "$\leqq_W$" in W generated by a family $\Omega = \{\omega_\beta\}_{\beta \in B}$ of continuous linear functionals on W in the same way as described in (ii), i.e. for $Q \subset X$

$$F(\hat{x}) = \operatorname*{v\text{-}min}_{x \in Q} F(x): \Leftrightarrow \text{there is no } x \in Q \text{ such that } \langle F(x), \omega_\beta \rangle \leqq \langle F(\hat{x}), \omega_\beta \rangle \text{ for all } \beta \in B \text{ and strict inequality holds for some } \beta' \in B;$$

(iv) X_0 is a nonempty subset of X.

The pre-orderings in Y and W correspond to convex cones $K_Y = -\Lambda^0 = \{y \in Y: \langle y, \lambda \rangle \geqq 0 \text{ for all } \lambda \in \Lambda\}$ (the negative polar) and $K_W = -\Omega^0$ via the relations $y \leqq_Y 0 \Leftrightarrow y \in -K_Y$ and $w \leqq_W 0 \Leftrightarrow w \in -K_W$, respectively. It is easy to show that if $\hat{x}$ is an efficient point of $F(Q)$ with respect to the cone K_W, i.e. if $F(Q) \cap \{F(\hat{x}) - K_w\} = \{F(\hat{x})\}$, then $F(\hat{x}) = \text{v-min}_{x \in Q} F(x)$. The converse is not true in general.

Necessary optimality conditions for problems like (P) with scalar-valued objective function have been obtained by Halkin [5] and Neustadt [11] for finite dimensional equality constraints, by Bazaraa and Goode [1] and Virsan [16] in the infinite dimensional case. In this paper we shall derive a multiplier rule under fairly mild differentiability conditions, including both the convex and the differentiable case. Further, the assumption that K_Y possesses interior points will be weakened.

* Received by the editors April 23, 1976, and in revised form August 12, 1976.

† Institut für Angewandte Mathematik der Universität Bonn, Abteilung für Wahrscheinlichkeitstheorie und Mathematische Statistik, Bonn, West Germany. This research was supported by the Sondersforschungsbereich 72 at the University of Bonn.

1. Convex approximations. Let us introduce the sets

$$Q_0 = \{x \in X : F(x) \leqq_W F(x_0) \text{ and } \langle F(x), \omega_\beta \rangle \text{ for some } \beta \in B\},$$

$$Q_1 = \{x \in X : G(x) \leqq_Y 0\},$$

$$Q_2 = \{x \in X : H(x) = 0\},$$

$$Q = Q_1 \cap Q_2 \cap X_0.$$

Then problem (P) can be written in the form

$$\text{find} \quad \text{v-min}\{F(x) : x \in Q\}.$$

Adopting the notation of Laurent [9] we define the cone of interior directions of a set $Q \subset X$ at the point $x_0 \in X$:

$$\Gamma(Q; x_0) = \{x \in X : \text{there exist a neighborhood } U \text{ of } x \text{ and a positive number } \varepsilon \text{ such that } x_0 + \bigcup_{0<\eta<\varepsilon} \eta U \subset Q\}$$

and the cone of tangent directions

$$\Gamma^*(Q; x_0) = \{x \in X : \text{for every neighborhood } U \text{ of } x \text{ and every positive number } \varepsilon \text{ there exist } y \in U \text{ and } 0 < \eta < \varepsilon \text{ such that } x_0 + \eta y \in Q\}.$$

In the classical theorems of Kuhn–Tucker type the cone of interior directions of the set Q_1 is characterized by means of the derivative of G. Following this line of thought we introduce a generalized concept of differentiability.

Let $A(x_0) = \{\alpha \in A : \langle G(x_0), \lambda_\alpha \rangle = 0\}$ denote the set of active constraints at the point x_0. The corresponding family $\Lambda(x_0) = \{\lambda_\alpha\}_{\alpha \in A(x_0)}$ of functionals generates a convex cone $K_Y' = -\Lambda(x_0)^0$ inducing a pre-ordering "$\leqq_Y'$" in Y. Obviously $K_Y \subset K_Y'$. Now assume:

(G) a) A is a topological space, and there exists a K_Y'-convex mapping $\hat{G} : X \to Y$ with $\hat{G}(0) = 0$ such that for all $\bar{x} \in X$ and all $\bar{\alpha} \in A(x_0)$ the following is true:

For every $\varepsilon > 0$ there exists an open neighborhood $\theta_{\bar{\alpha}}$ of $\bar{\alpha}$ and a positive number $\eta_{\bar{\alpha}}$ as well as a neighborhood $U_{\bar{x}}$ of $\bar{x}$ such that

$$\frac{G_\alpha(x_0 + \eta x) - G_\alpha(x_0)}{\eta} - \hat{G}_\alpha(\bar{x}) < \varepsilon \tag{1.1}$$

for all $x \in U_{\bar{x}}$, $0 < \eta < \eta_{\bar{\alpha}}$, $\alpha \in \theta_{\bar{\alpha}}$.

b) There is a neighborhood $\mathcal{U}$ of x_0 such that

(i) $(\alpha, x) \mapsto G_\alpha(x) : A \times \mathcal{U} \to \mathbb{R}$ and

(ii) $(\alpha, x) \mapsto \hat{G}_\alpha(x) : A(x_0) \times X \to \mathbb{R}$

are continuous mappings.

Here $G_\alpha(x)$ and $\hat{G}_\alpha(x)$ stand for $\langle G(x), \lambda_\alpha \rangle$ and $\langle \hat{G}(x), \lambda_\alpha \rangle$, respectively.

Let us compare the above concept with the notion of differentiability introduced by Neustadt [11]:

(G′) There exists a K_Y'-convex mapping $\hat{G}\colon X \to Y$, $\hat{G}(0)=0$, such that for every $\bar{x} \in X$ and every 0-neighborhood V in Y, a neighborhood U of $\bar{x}$ and a positive number δ can be found such that

$$\frac{G(x_0+\eta x)-G(x_0)}{\eta} \in G(\bar{x})-K_Y'+V \tag{1.2}$$

for all $x \in U$, $0<\eta<\delta$.

It can be shown that if the mapping $(\alpha, y)\mapsto\langle y, \lambda_\alpha\rangle\colon A(x_0)\times Y \to \mathbb{R}$ is continuous then (G′) implies (G) part a). For compact $A(x_0)$ this amounts to requiring the equicontinuity of $\Lambda(x_0)$. Let us remark in this context that every convex cone K containing interior points can be represented in the form $K=\Lambda^0$ where Λ is an equicontinuous family of functionals.

Introducing the notation

$$K_\alpha=\{x \in X\colon \hat{G}_\alpha(x)<0\},$$

$$K_1=\begin{cases}\bigcap_{\alpha\in A(x_0)} K_\alpha & \text{if } A(x_0)\neq\varnothing,\\ X & \text{if } A(x_0)=\varnothing,\end{cases}$$

we can now give a characterization of the cone $\Gamma(Q_1; x_0)$ where x_0 is supposed to be feasible for the inequality constraints, i.e. $G(x_0) \leqq_Y 0$.

PROPOSITION 1. *Let G satisfy condition* (G). *Suppose further that the following assumptions hold*:

(A1) *$A(x_0)$ is compact and*

(A2) *for every open neighborhood θ of $A(x_0)$ there is a 0-neighborhood U_θ in X such that $G_\alpha(x_0+x)<0$ for all $x \in U_\theta$ and all $\alpha \in A\setminus\theta$.*

Then

$$\Gamma(Q_1; x_0)\supset K_1. \tag{1.3}$$

Proof. (I) Consider first the case $A(x_0)=\varnothing$. For $\theta=\varnothing$, (A2) guarantees the existence of a 0-neighborhood $U_\varnothing$, which may be supposed to be convex, such that $G_\alpha(x_0+x)<0$ for all $x \in U_\varnothing$, $\alpha \in A$. Now for arbitrary $\bar{x} \in X$ choose $\varepsilon>0$ such that $\varepsilon\bar{x} \in U_\varnothing$ and take a neighborhood U of $\bar{x}$ that lies in $(1/\varepsilon)U_\varnothing$. Then $G_\alpha(x_0+\eta x)<0$ for all $x \in U$ and all $0<\eta<\varepsilon$, i.e. $\bar{x} \in \Gamma(Q_1; x_0)$.

(II) Suppose $\bar{x} \in K_1$, $\bar{\alpha} \in A(x_0)$. By (G) b) (ii) an open neighborhood $\theta_{\bar{\alpha}}^{(1)}$ of $\bar{\alpha}$ can be found such that

$$\hat{G}_\alpha(\bar{x}) \leqq -\delta<0 \quad \text{for all } \alpha \in \theta_\alpha^{(1)} \tag{1.4}$$

where $\delta=-\hat{G}_{\bar{\alpha}}(\bar{x})/2$. Let $U_{\bar{\alpha}}$ and $\theta_{\bar{\alpha}}^{(2)}$ be the neighborhoods of $\bar{x}$ and $\bar{\alpha}$, respectively, and $\eta_{\bar{\alpha}}$ the positive number whose existence is postulated in (G) part a) for $\varepsilon=\delta/2$. Then by (1.1) and (1.4)

$$\frac{G_\alpha(x_0+\eta x)-G_\alpha(x_0)}{\eta}<\hat{G}_\alpha(\bar{x})+\frac{\delta}{2}\leqq-\frac{\delta}{2}<0 \quad \text{for all } x \in U_{\bar{\alpha}},$$

$$\alpha \in \theta_{\bar{\alpha}} := \theta_{\bar{\alpha}}^{(1)}\cap\theta_{\bar{\alpha}}^{(2)}, \quad 0<\eta<\eta_{\bar{\alpha}},$$

and $G_\alpha(x_0) \leqq 0$ implies

$$G_\alpha(x_0+\eta x)<0 \quad \text{for all } x \in U_{\bar\alpha}, \quad \alpha \in \theta_{\bar\alpha}, \quad 0<\eta<\eta_{\bar\alpha}.$$

$A(x_0)$ being compact we can select a finite number of neighborhoods $\theta_{\bar\alpha_1}, \ldots, \theta_{\bar\alpha_n}$ from the open covering $\{\theta_{\bar\alpha}\}_{\bar\alpha \in A(x_0)}$ of $A(x_0)$ such that

$$A(x_0) \subset \bigcup_{i=1}^{n} \theta_{\bar\alpha_i} =: \theta.$$

Setting $U_1 := \bigcap_{i=1}^{n} U_{\bar\alpha_i}$, $\eta_1 := \min_{i=1,\ldots,n} \eta_{\bar\alpha_i}$ we obtain

$$G_\alpha(x_0+\eta x)<0 \quad \text{for all } x \in U_1, \quad \alpha \in \theta, \quad 0<\eta<\eta_1. \tag{1.5}$$

Now proceeding as in the first part of the proof choose $\eta_2>0$ and a neighborhood U_2 of $\bar x$ such that

$$G_\alpha(x_0+\eta x)<0 \quad \text{for all } x \in U_2, \quad \alpha \in A\setminus\theta, \quad 0<\eta<\eta_2. \tag{1.6}$$

Then with $U := U_1 \cap U_2$, $\varepsilon := \min(\eta_1, \eta_2)$, (1.5) and (1.6) imply

$$G(x_0+\eta x) \leqq_Y 0 \quad \text{for all } x \in U, \quad 0<\eta<\varepsilon.$$

This completes the proof. □

Proposition 1 is especially useful in cases where Y is a function space over a noncompact time interval with "$\leqq_Y$" representing the natural (pointwise) ordering. Let, for example, $G(x(t))$, $t \in I$, be a continuous function describing phase restrictions in some control problem. Then, from the structure of the problem, it may be known that, for an optimal solution x_0, $G(x_0(t)) = 0$ on some compact set of t-values and that there is no asymptotic movement to zero outside this set.

COROLLARY 1. *If A is compact and G satisfies* (G), (1.3) *holds.*

Proof. A straightforward compactness argument shows that (A1) and (A2) hold. □

The cone $\Gamma(Q_0; x_0)$ can be dealt with in a similar manner. Let (F) denote the assumption obtained from (G) by substituting $W, F, \hat F, B, \beta, \omega$ for $Y, G, \hat G, A, \alpha, \lambda$, respectively. Introducing the mapping $\phi(x) = F(x) - F(x_0)$ and the sets

$$\tilde Q_0 = \{x \in X: \phi(x) \leqq_W 0\},$$

$$Q_\beta = \{x \in X: \phi_\beta(x) < 0\},$$

$$K_\beta = \{x \in X: \hat\phi_\beta(x) < 0\},$$

$$K_0 = \bigcap_{\beta \in B} K_\beta$$

($\phi_\beta(x) = \langle \phi(x), \omega_\beta \rangle$, $\hat\phi_\beta(x) = \langle \hat\phi(x), \omega_\beta \rangle$) we have

$$Q_0 = \tilde Q_0 \cap \bigcup_{\beta \in B} Q_\beta.$$

Making use of the relations (compare [9])

$$\Gamma\Big(\bigcap_{i \in I} Q_i; x_0\Big) = \bigcap_{i \in I} \Gamma(Q_i; x_0) \quad \text{for finite } I,$$

$$\Gamma\Big(\bigcup_{j \in J} Q_j; x_0\Big) \supset \bigcup_{j \in J} \Gamma(Q_j; x_0) \quad \text{for arbitrary } J$$

we obtain

$$\Gamma(Q_0; x_0) \supset \Gamma(\tilde{Q}_0; x_0) \cap \bigcup_{\beta \in B} \Gamma(Q_\beta; x_0). \tag{1.7}$$

Now if F is differentiable in the sense of (F), so is ϕ with $\hat{\phi} = \hat{F}$. Observing $B(x_0) = B$ we apply Proposition 1 to obtain

PROPOSITION 2. *Let F satisfy the differentiability condition* (F) *and assume*:

(B1) *B is compact.*

Then

$$\Gamma(Q_0; x_0) \supset K_0. \tag{1.8}$$

Proof. By Proposition 1, $\Gamma(\tilde{Q}_0; x_0) \supset K_0$, and, as is easily shown, $\Gamma(Q_\beta; x_0) \supset K_\beta$. From (1.7) we get $\Gamma(Q_0; x_0) \supset K_0 \cap \bigcup_{\beta \in B} K_\beta = K_0$. □

Finally, with regard to H, let us impose the following conditions.

(H) For all $\bar{x} \in X$ the limit

$$\hat{H}(\bar{x}) = \lim_{\substack{x \to \bar{x} \\ \eta \searrow 0}} \frac{H(x_0 + \eta x) - H(x_0)}{\eta}$$

exists, and $\hat{H}(x)$ is continuous and linear in x.

(C1) $\ker \hat{H} \subset \Gamma^*(Q_2; x_0)$.

This leads us to the following almost trivial result:

PROPOSITION 3. (H) *and* (C1) *imply* $\ker \hat{H} = \Gamma^*(Q_2; x_0)$.

A sufficient condition for (C1) is given by Liusternik [10].

2. A factorization theorem. Having found subsets of the approximating cones which can be expressed in terms of the generalized derivatives $\hat{F}$ and $\hat{G}$, we will proceed to the next step which will be to apply a separation theorem to these sets. This will establish the existence of zero-sum functionals belonging to the polar cones. The Lagrange multipliers are found as the first factors in the decompositions $l = \omega \circ \hat{F}$ etc. of these functionals.

To see when such a factorization is possible we need a result from the theory of ordered topological vector spaces (Peressini [12]).

LEMMA 1. *Let E be a locally convex space, $\{e_\gamma\}_{\gamma \in \Gamma}$ and $\{r_\gamma\}_{\gamma \in \Gamma}$ families in E and $\mathbb{R}$ respectively. Then a sufficient condition for the existence of a continuous linear functional φ on E satisfying $\langle e_\gamma, \varphi \rangle \geqq r_\gamma$ for all $\gamma \in \Gamma$ is the existence of a 0-neighborhood V in E such that $\sum_{i=1}^n \rho_i e_{\gamma_i} \in V$, $\rho_i \geqq 0$, $\gamma_i \in \Gamma$, implies*

$$\sum_{i=1}^n \rho_i r_{\gamma_i} \leqq M$$

for some positive number M.

The cones K_Y and K'_Y induce dual cones $K_{Y^*} = -K_Y^0$ and $K'_{Y^*} = -(K'_Y)^0$ in Y^* (the topological dual). The corresponding pre-orderings in Y^* will be denoted by "$\leqq_{Y^*}$" and "$\leqq'_{Y^*}$". Clearly $K'_{Y^*} \subset K_{Y^*}$.

PROPOSITION 4. *Let G satisfy* (G). *Assume further*:

(A3) *$\Lambda(x_0)$ is w^*-compact and* $\operatorname{int}(K'_Y) \neq \emptyset$.

Then if $K_1 \neq \emptyset$, for every $l \in K_1^0$ there exists $\lambda \in K'_{Y^*}$ such that

$$\langle \hat{G}(x), \lambda \rangle \geqq \langle x, l \rangle \quad \text{for all } x \in X, \tag{2.1}$$

$$\langle G(x_0), \lambda \rangle = 0. \tag{2.2}$$

Proof. We apply Lemma 1 with $E = Y$ and $\Gamma = X \cup K'_Y$, the families $\{e_\gamma\}$ and $\{r_\gamma\}$ being defined by

$$e_\gamma = \begin{cases} \hat{G}(x) & \text{for } \gamma = x \in X, \\ y & \text{for } \gamma = y \in K'_Y \end{cases}$$

and

$$r_\gamma = \begin{cases} \langle x, l \rangle & \text{for } \gamma = x \in X, \\ 0 & \text{for } \gamma = y \in K'_Y \end{cases}$$

respectively. Take $x_1 \in K_1$ and set $m = \max_{\alpha \in A(x_0)} \hat{G}_\alpha(x_1)$. Then $m < 0$ (we suppose $A(x_0) \neq \varnothing$, the case $A(x_0) = \varnothing$ being trivial). Choose $y_1 \in \operatorname{int}(K'_Y)$ such that $-m > m' = \max_{\alpha \in A(x_0)} \langle y_1, \lambda_\alpha \rangle > 0$. Then $V = y_1 - K'_Y$ is a 0-neighborhood in Y. Denote by $\hat{G}'_{\alpha_+}(0; \cdot)$ the one-sided Gâteaux derivative of $\hat{G}_\alpha$ at 0. Then $\sum_{i=1}^n \rho_i \hat{G}(x_i) \in V$ $(\rho_i \geqq 0)$ implies

$$\sum_{i=1}^n \rho_i \hat{G}'_{\alpha_+}(0; x_i) \leqq \sum_{i=1}^n \rho_i \hat{G}_\alpha(x_i) \leqq \langle y_1, \lambda_\alpha \rangle \quad \text{for all } \alpha \in A(x_0)$$

and so

$$\sum_{i=1}^n \rho_i \hat{G}'_{\alpha_+}(0; x_i) \leqq m' < -m.$$

From the subadditivity of the one-sided Gâteaux derivative (Holmes [7]) it follows that

$$\begin{aligned} \hat{G}'_{\alpha_+}\Big(0; \sum_{i=1}^n \rho_i x_i + x_1\Big) &\leqq \sum_{i=1}^n \rho_i \hat{G}'_{\alpha_+}(0; x_i) + \hat{G}_\alpha(x_1) \\ &\leqq \sum_{i=1}^n \rho_i \hat{G}'_{\alpha_+}(0; x_i) + m < 0 \quad \text{for all } \alpha \in A(x_0), \end{aligned}$$

which means that $\sum_{i=1}^n \rho_i x_i + x_1 \in \bigcap_{\alpha \in A(x_0)} K'_\alpha =: K'_1$ with $K'_\alpha = \{x : \hat{G}'_{\alpha_+}(0; x) < 0\}$. Because $\hat{G}(0) = 0$, $\hat{G}'_{\alpha_+}(0; x) < 0$ implies $\hat{G}_\alpha(\eta_\alpha x) < 0$ for some $\eta_\alpha > 0$; a compactness argument then shows that for every $x \in K'_1$ there is a positive number η such that $\eta x \in K_1$. Hence $K'^0_1 = K^0_1$, and for $l \in K^0_1$ we have

$$\sum_{i=1}^n \rho_i \langle x_i, l \rangle \leqq -\langle x_1, l \rangle =: M \quad \text{for all } \rho_i \geqq 0, \quad x_i \in X \text{ such that } \sum_{i=1}^n \rho_i G(x_i) \in V. \tag{2.3}$$

Now consider $\sum_{i=1}^n \rho_i \hat{G}(x_i) + \sum_{i=1}^k \sigma_i y_i \in V$, $\rho_i, \sigma_i \geqq 0$, $y_i \in K'_Y$. Setting $y := \sum_{i=1}^k \sigma_i y_i \in K'_Y$ we obtain $\sum_{i=1}^n \rho_i \hat{G}(x_i) \in V - y = y_1 - (K'_Y + y) \subset y_1 - K'_Y = V$. From (2.3) and the definition of $\{r_\gamma\}_{\gamma \in \Gamma}$ it follows that the sufficient condition of Lemma 1 is fulfilled. Hence there is a continuous linear functional λ such that $\langle \hat{G}(x), \lambda \rangle \geqq \langle x, l \rangle$ for all $x \in X$ and $\langle y, \lambda \rangle \geqq 0$ for all $y \in K'_Y$, i.e. $\lambda \geqq_{Y^*} 0$. $\langle G(x_0), \lambda \rangle = 0$ is obvious. □

Applying the above result to F yields:

PROPOSITION 5. *Let assumptions* (F) *and* (B1) *be satisfied. Assume further*:

(B2) Ω *is* w^**-compact and* $\operatorname{int}(K_W) \neq \emptyset$.

Then if $K_0 \neq \emptyset$, *for every* $l \in K_0^0$ *there is a functional* $\omega \in K_{W^*} = -K_W^0$ *such that*

$$\langle \hat{F}(x), \omega \rangle \geqq \langle x, l \rangle \quad \text{for all } x \in X. \tag{2.4}$$

Finally let us impose a condition on H which will be useful in proving factorization properties of functionals in $\Gamma^*(Q_2; x_0)$:

(C2) $\hat{H}: X \to \hat{H}(X)$ is an open mapping.

PROPOSITION 6. *Under condition* (C2) *every* $l \in (\ker \hat{H})^0$ *can be written in the form* $l = \mu \circ \hat{H}$ *with* $\mu \in Z^*$.

3. The multiplier theorem.

LEMMA 2. *Let* x_0 *be a solution of* (P). *Then*

$$\Gamma(Q_0; x_0) \cap \Gamma(Q_1; x_0) \cap \Gamma(X_0; x_0) \cap \Gamma^*(Q_2; x_0) = \emptyset.$$

Proof. Assume that the intersection contains a point $\bar{x}$. Then $\bar{x} \in \Gamma(Q_0 \cap Q_1 \cap X_0; x_0)$, and by definition there exist a neighborhood U of $\bar{x}$ and a positive number ε such that $x_0 + \eta x \in Q_0 \cap Q_1 \cap X_0$ for all $x \in U$, $0 < \eta < \varepsilon$. According to the definition of Γ^* we can find $x' \in U$ and $0 < \eta' < \varepsilon$ such that $\hat{x} = x_0 + \eta' x' \in Q_2$. Then $\hat{x}$ satisfies all the constraints and $F(\hat{x}) \leqq_W F(x_0)$ with $F_\beta(\hat{x}) < F_\beta(x_0)$ for at least one $\beta \in B$. This is a contradiction to the optimality of x_0. □

LEMMA 3. *Let* $C_0, \cdots, C_{n-1}$ *be nonempty open convex sets in* X, C_n *a nonvoid convex set with the property that* $0 \in \operatorname{cl}(C_i)$, $i = 0, \cdots, n$. *Then a necessary and sufficient condition for* $\bigcap_{i=1}^n C_i = \emptyset$ *is the existence of functionals* $l_i \in C_i^0$, $i = 0, \cdots, n$, *not all zero, such that* $\sum_{i=1}^n l_i = 0$.

The proof is based on a separation theorem (compare Laurent [9]).

On the basis of these two lemmas the following multiplier rule can be proved:

THEOREM 1. *Suppose* $x_0 \in Q$. *Assume that* F, G *and* H *satisfy* (F), (G) *and* (H), *respectively, and that conditions* (A1)–(A3), (B1), (B2), (C1) *and* (C2) *are fulfilled. Further assume that* $\Gamma(X_0; x_0)$ *contains an open convex cone* K *with vertex* 0. *Then a necessary condition for* x_0 *to be a solution of* (P) *is the existence of functionals* $\omega_0 \in K_{W^*}$, $\lambda_0 \in K'_{Y^*}$, $\mu_0 \in Z^*$, *not all zero, such that*

$$\langle \hat{F}(x), \omega_0 \rangle + \langle \hat{G}(x), \lambda_0 \rangle + \langle \hat{H}(x), \mu_0 \rangle \geqq 0 \quad \text{for all } x \in K, \tag{3.1}$$

$$\langle G(x_0), \lambda_0 \rangle = 0. \tag{3.2}$$

Proof. From Lemma 2 and Propositions 1–3 it follows that

$$K_0 \cap K_1 \cap K \cap \ker \hat{H} = \emptyset. \tag{3.3}$$

Suppose first that $K_0 \neq \emptyset$, $K_1 \neq \emptyset$. Then the sets in (3.3) satisfy the conditions of Lemma 3 (note that by (G) b) (ii) and (A1) K_1 is open; the same is true for K_0). Hence there exist functionals $l_0 \in K_0^0$, $l_1 \in K_1^0$, $l_2 \in \Gamma(K; x_0)^0 = K^0$, $l_3 \in (\ker \hat{H})^0$, not all zero, such that $l_0 + l_1 + l_2 + l_3 = 0$. Application of Propositions 3–5 yields functionals $\omega_0 \in K_{W^*}$, $\lambda_0 \in K'_{Y^*}$, $\mu_0 \in Z^*$ such that

$$\langle \hat{F}(x), \omega_0 \rangle + \langle \hat{G}(x), \lambda_0 \rangle + \langle \hat{H}(x), \mu_0 \rangle + \langle x, l_2 \rangle \geqq 0$$

for all x and $\langle G(x_0), \lambda_0\rangle = 0$. Equation (3.3) then follows from the definition of l_2.

Now take the case $K_0 = \varnothing$. Then the set $\hat{F}(X)^+ := \{w \in W: w \geqq_W \hat{F}(x)$ for some $x \in X\}$ is convex and has the property $\hat{F}(X)^+ \cap \operatorname{int}(-K_W) = \varnothing$. By a well known separation theorem there is a closed hyperplane with slope $\omega_0 \neq 0$ separating $\hat{F}(X)^+$ and $-K_W$:

$$\langle -K_W, \omega_0\rangle \leqq 0,$$

$$\langle \hat{F}(x), \omega_0\rangle \geqq 0 \quad \text{for all } x \in X.$$

The argument in the case $K_1 = \varnothing$ follows the same lines. □

In nonlinear control problems a direct application of Theorem 1—treating the differential equation as equality constraint $H = 0$—leads only to a differential form of the maximum principle; the proof of the proper maximum principle by means of this theorem requires a certain transformation of the original problem (compare [4]). More versatility may be obtained by amalgamating part of the equality constraints in the set X_0.

THEOREM 1'. *Suppose $x_0 \in Q$. Assume that F and G satisfy* (F) *and* (G), *respectively, and that conditions* (A1)–(A3), (B1) *and* (B2) *are fulfilled. Define $X_0' = X_0 \cap Q_2$; assume that $\Gamma^*(X_0'; x_0)$ contains a convex cone K' with vertex* 0. *Then for x_0 to be a solution of* (P) *it is necessary that there exist functionals $\omega_0 \in K'_{W^*}$, $\lambda_0 \in K'_{Y^*}$, not both zero, such that*

$$\langle \hat{F}(x), \omega_0\rangle + \langle \hat{G}(x), \lambda_0\rangle \geqq 0 \quad \textit{for all } x \in K', \tag{3.1'}$$

$$\langle G(x_0), \lambda_0\rangle = 0. \tag{3.2'}$$

Proof. An obvious modification of Lemma 2 leads to the necessary condition

$$\Gamma(Q_0; x_0) \cap \Gamma(Q_1; x_0) \cap \Gamma^*(X_0'; x_0) = \varnothing.$$

Then similar considerations as above apply. □

As to the scope of these theorems it should be noted that no constraint qualification is needed and so part of the equality constraints could also be incorporated in Q_1 if there is a suitable way of defining the order cone. For finite dimensional constraints this is always possible.

Theorem 1 may be compared to the result obtained by Virsan [16], who derives a multiplier rule for operatorial equality constraints; however, only finite dimensional inequality constraints are admitted. More closely related to our results are those of Bazaraa and Goode [1]; they are obtained for normed spaces under stronger differentiability conditions and the assumption that K_Y itself contains interior points. The results of Neustadt are of the form presented in Theorem 1' with G consisting of the finite dimensional equality constraints φ and the (possibly) infinite dimensional inequality constraints ϕ in [11]. K_Y—corresponding to the cone $\bar{Z}$—is required to contain interior points, whereas K' plays the role of the convex approximating set $\mathcal{M}$.

Up to now we know only that at least one multiplier must be nonzero. We shall now give a sufficient condition for $\omega_0 \neq 0$.

PROPOSITION 7. *If in addition to the assumptions made in Theorem* 1

$$K_1 \cap \Gamma(X_0; x_0) \cap \ker \hat{H} \neq \varnothing$$

holds, then $\omega_0 \neq 0$.

Proof. We can confine ourselves to the case $K_0 \neq \varnothing$. Going back to the proof of Theorem 1, suppose that $\omega_0 = 0$. Then it follows that $l_1 + l_2 + l_3 = 0$ and $l_1 \neq 0$ or $l_2 \neq 0$. To be definite, suppose $l_1 \neq 0$. Then for $x \in K_1 \cap \Gamma(X_0; x_0) \cap \ker \hat{H}$ we find $\langle x, l_1 \rangle < 0$, $\langle x, l_2 \rangle \leqq 0$, $\langle x, l_3 \rangle = 0$; hence $\langle x, l_1 + l_2 + l_3 \rangle < 0$, which leads to a contradiction. □

4. Application to a variational problem. We shall be concerned with the following variational problem of Bolza type:

$$\text{minimize} \quad l(x(0), x(1)) + \int_0^1 L(t, x(t), \dot{x}(t))\, dt$$

$$\text{subject to} \quad g_i(t, x(t)) \leqq 0 \quad \text{for all } t \in I_i, \quad i = 1, \cdots, k,$$

where x ranges in the class of absolutely continuous functions from $[0, 1]$ to $\mathbb{R}^n$, $l: \mathbb{R}^{2n} \to \mathbb{R}$, $L: [0, 1] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and $g: [0, 1] \times \mathbb{R}^n \to \mathbb{R}^k$ are given functions and the I_i are compact subsets of $[0, 1]$.

As is shown in [14], constrained problems of this kind can be transformed into an unconstrained variational problem of the form

$$\text{minimize} \quad \tilde{l}(z(0), z(1)) + \int_0^1 \tilde{L}(t, z(t), \dot{z}(t))\, dt$$

where the minimum is to be taken over all absolutely continuous functions z taking on values in an appropriately augmented state space. Here $\tilde{l}$ and $\tilde{L}$ are extended real valued functions taking on the value $+\infty$ outside certain sets which are suitably chosen so as to describe the problem constraints. By means of the theory of conjugate convex functions and an appropriate extension to the nonconvex case, Rockafellar and Clarke have obtained generalized versions of the Euler–Lagrange equation in the convex case [13] and under the assumption that L satisfies a Lipschitz condition [2].

In the approach developed in the preceding sections the constraints enter the problem explicitly, i.e. in the form of functional equalities or inequalities. We restrict our attention to the case where l and L are locally Lipschitz (in a sense to be defined below)—which is a stronger assumption than in [2], where l may be lower semi-continuous—and the g_i are convex functions. Additional constraints such as convex differential inclusions or convex mixed phase-control constraints, which can be treated in the framework of [13], could be included. The former—without offering any major difficulties—would produce an additional multiplier playing the role of an adjoint variable; the latter would lead to specific problems concerning the well-behavior of the corresponding multipliers, which come up as bounded finitely additive measures. To keep things simple we shall restrict ourselves to the problem stated above. However, at the end of this section, we shall indicate the modifications caused by constraints of the form $(x(0), x(1)) \in S$.

In the case where both l and L are convex—which is also covered because of the finiteness of l and L—our results may be compared to those obtained in [15]. In the nonconvex case the approach via extended real valued functions will lead to troubles concerning the Lipschitz properties of L when the phase constraints are integrated into the objective function.

We begin with some definitions and propositions taken from [3]. Let (t, x, u), (x, y) and (t, x) denote the arguments of L, l and g, respectively.

DEFINITION 1. Let $x(\cdot)$ be an absolutely continuous function with derivative $\dot{x}(\cdot)$. L is called *Lipschitz* near $x(\cdot)$ if there exist an integrable function $k(\cdot)$ on $[0, 1]$ and a positive number r such that whenever (x_1, u_1), (x_2, u_2) lie within r of $(x(t), x(t))$ the inequality

$$|L(t, x_1, u_1) - L(t, x_2, u_2)| \leqq k(t)|(x_1 - x_2, u_1 - u_2)|$$

holds.

Let F be a functional on a normed space X satisfying a local Lipschitz condition at some point x_0, i.e.

$$|F(x) - F(y)| \leqq K\|x - y\| \tag{4.1}$$

for all x, y close enough to x_0.

DEFINITION 2. The functional

$$F^0(x_0; x) = \limsup_{\substack{h \to 0 \\ \delta \searrow 0}} \frac{F(x_0 + h + \delta x) - F(x_0 + h)}{\delta}$$

is called the *generalized directional derivative* of F at x_0 in the x direction.

DEFINITION 3. The *generalized gradient* of F at x_0, denoted $\partial^0 F(x_0)$, is the convex hull of the set of points of the form $\lim_{i \to \infty} \nabla F(x_i)$, where we consider all sequences x_i converging to x_0 such that $\nabla F(x_i)$ exists and converges.

For functionals on $\mathbb{R}^n$ the following properties are shown in [3]:

PROPOSITION 8. (i) *$F^0(x_0; x)$ is a finite convex function of x.*

(ii) *If F is convex, then $\partial^0 F(x_0) = \partial F(x_0)$, where $\partial F(x_0)$ is the usual set of subgradients.*

(iii) $\partial^0 F^0(0) = \partial^0 F(x_0)$.

We now come to a result relating the notion of generalized directional derivative with the concept of differentiability developed in § 1.

PROPOSITION 9. *$F^0(x_0; \cdot)$ satisfies the differentiability condition* (F).

Proof. With the notation of § 1 we have $W = \mathbb{R}$ and $\{\omega_\beta\}_{\beta \in B} = \{1\}$. The Lipschitz property implies that $F^0(x_0; x)$ is bounded above on an open set; hence, by convexity, it is continuous.

Now, for $\varepsilon > 0$, we can find positive numbers $h(\varepsilon)$ and $\delta(\varepsilon) \leqq 1$ such that

$$\frac{F(x_0 + h + \delta x) - F(x_0 + h)}{\delta} \leqq F^0(x_0; x) + \varepsilon \quad \text{for all } h \leq h(\varepsilon), \quad 0 < \delta \leqq \delta(\varepsilon).$$

Let $B_{h(\varepsilon)}(x)$ denote the $h(\varepsilon)$-ball centered at x. Then every $\eta x'$ with $x' \in B_{h(\varepsilon)}(x)$ and $0 < \eta \leqq \delta(\varepsilon)$ can be written in the form $\eta x' = \eta x + h$ with $\|h\| \leqq h(\varepsilon)$. Hence, if in addition h is chosen so small that (4.1) is valid for the pair $(x_0, x_0 + h)$,

$$\frac{F(x_0 + \eta x') - F(x_0)}{\eta} = \frac{F(x_0 + \eta x + h) - F(x_0 + h)}{\eta} + \frac{F(x_0 + h) - F(x_0)}{\eta}$$

$$\leqq (F^0(x_0; x) + \varepsilon) + \frac{K\|h\|}{\eta} \leqq F^0(x_0; x) + \varepsilon + Kh(\varepsilon).$$

According to the remark following (1.2) this proves the assertion. □

In order to apply Theorem 1 we must write our problem as an abstract optimization problem. Restricting the admissible solutions to absolutely continuous functions with essentially bounded derivative, let us suppose that x_0 is an optimal solution. In particular this means that $L(t, x_0(t), \dot{x}_0(t))$ is a measurable and integrable function. If we assume that L is measurable in t and Lipschitz in (x, u) near x_0, then $L(t, x(t), u(t))$ is measurable and integrable (with finite value of the integral) for all $(x, u) \in C^n \times L^n_\infty$ lying within r of $(x_0(t), \dot{x}_0(t))$ almost everywhere. Thus the functional

$$F_2(x, u) = \int_0^1 L(t, x(t), u(t))\, dt$$

is well defined and finite on a ball $B_r(x_0, \dot{x}_0)$ in $C^n \times L^n_\infty$. We extend its domain of definition to the whole space by setting it equal to a positive constant outside $B_r(x_0, \dot{x}_0)$.

Further we introduce mappings F_1, G and H on $C^n \times L^n_\infty$ with values in $\mathbb{R}$, C^k and C^n, respectively, defined by

$$F_1(x, u) = l(x(0), x(1)),$$

$$G(x, u)(t) = g(t, x(t)),$$

$$H(x, u)(t) = x(t) - \int_0^t u(s)\, ds.$$

Then $(x_0, \dot{x}_0)$ is a solution to the problem

$$\begin{aligned} \text{find} \quad & \min \{F_1(x, u) + F_2(x, u) : (x, u) \in B_r(x_0, \dot{x}_0), \\ & H(x, u) = 0,\ G(x, u) \leqq_{K_I} 0\}, \end{aligned} \tag{4.2}$$

which is of the form discussed in the preceding sections. Here we have set $K_I = \{x \in C^n : x_i(t) \leqq 0 \text{ for all } t \in I_i,\ i = 1, \cdots, k\}$, or, in the notation of the Introduction $\{\lambda_\alpha\}_{\alpha \in A} = \{\delta^i_t\}_{t \in I_i; i = 1, \cdots, k}$, where $\langle x, \delta^i_t \rangle = x_i(t)$.

LEMMA 4. *Under the assumptions made above F_2 is locally Lipschitz at $(x_0, \dot{x}_0)$. Furthermore, $F_2^0(x_0, \dot{x}_0; x, u)$ exists for all $(x, u) \in C^n \times L^n_\infty$ and $F_2^0(x_0, \dot{x}_0; x, u) \leqq \int_0^1 L^0(t, x_0(t), \dot{x}_0(t); x(t), u(t))\, dt$.*

Proof. For (x, u), $(x', u') \in B_r(x_0, \dot{x}_0)$ we have

$$\begin{aligned} &|F_2(x, u) - F_2(x', u')| \\ &\quad \leqq \int_0^1 k(t) |(x(t) - x'(t), u(t) - u'(t))|\, dt \\ &\quad \leqq \int_0^1 k(t) [|x(t) - x'(t)|^2 + |u(t) - u'(t)|^2]^{1/2}\, dt \\ &\quad \leqq \int_0^1 k(t) \cdot 2^{1/2} \sup (|x(t) - x'(t)|, |u(t) - u'(t)|)\, dt \\ &\quad \leqq 2^{1/2} \int_2^1 k(t) [|x(t) - x'(t)| + |u(t) - u'(t)|]\, dt \\ &\quad \leqq 2^{1/2} \|k\| (\|x - x'\| + \|u - u'\|) = K' \|(x - x', u - u')\|. \end{aligned}$$

The second part of the assertion is a consequence of Fatou's lemma. □

Concerning l and g we make the following assumptions: l is locally Lipschitz at $(x_0(0), x_0(1))$; g is continuous in (t, x) and convex in x. Then it is easy to see that

$$F_1^0(x_0, u_0; x, u) = l^0(x_0(0), x_0(1); x(0), x(1))$$

and that

$$\hat{G}(x, u) = G(x_0 + x, u) - G(x_0, u)$$

satisfies (G).

Finally let us suppose that there exists an absolutely continuous function x_1 such that $g_i(t, x_0(t) + x_1(t)) < 0$ for all t satisfying $g_i(t, x_0(t)) = 0$, $i = 1, \cdots, k$. Then by Theorem 1 and Proposition 7 there exist n functions μ_i with bounded variation and k nondecreasing functions λ_i with the following properties:

(4.3) $\lambda_i(1) = 0$, and $\lambda_i(t)$ is continuous from the right and constant on every subinterval having an empty intersection with the set $\{y: g_i(t, x_0(t)) = 0\}$,

such that

$$\begin{aligned} J(x, u) := {} & l^0(x_0(0), x_0(1); x(0), x(1)) \\ & + \int_0^1 L^0(t, x_0(t), x_0(t); x(t), u(t))\, dt \\ & + \sum_{i=1}^{k} \int_0^1 [g_i(t, x_0(t) + x(t)) - g_i(t, x_0(t))] \\ & + \sum_{i=1}^{n} \int_0^1 [x_i(t) - \int_0^t u_i(s)\, ds]\, d\mu_i(t) \geqq 0 \end{aligned} \tag{4.4}$$

for all $(x, u) \in B_r(0, 0)$. Hence J has a local minimum at the point $(0, 0)$, which is also a global one because J is a convex function. From ordinary subdifferential calculus we obtain that $(0, 0) \in \partial J(0, 0) = \sum_{i=1}^{4} \partial J_i(0, 0)$, where J_i stands for the ith term in the sum (4.4). Our main task is now to calculate the subdifferentials. As to J_2 and J_3 this task is achieved by using the fundamental results of Ioffe and Levin in [8]. They imply that for every pair $(\rho_2, \sigma_2) \in J_2(0, 0)$ there are measurable $\mathbb{R}^n$-valued functions $p_2(t)$, $q_2(t)$ with integrable components satisfying

$$\langle (x, u), (\rho_2, \sigma_2) \rangle = \int_0^1 [\langle x(t), p_2(t) \rangle + \langle u(t), q_2(t) \rangle]\, dt$$

for all $(x, u) \in C^n \times L_\infty^n$ and

$$(p_2(t), q_2(t)) \in \partial L^0(t, x_0(t), \dot{x}_0(t); 0, 0) \quad \text{a.e.} \tag{4.5}$$

Similarly, taking into account that every pair $(\rho_3, \sigma_3) \in \partial J_3(0, 0)$ must have its second component equal to zero, we find that the subgradients of J_3 can be represented in the form

$$\langle (x, u), (\rho_3, 0) \rangle = \sum_{i=1}^{k} \int_0^1 \langle x(t), p_3^i(t) \rangle\, d\lambda_i(t),$$

where the p_3^i are measurable $\mathbb{R}^n$-valued functions with λ_i-integrable components satisfying

$$p_3^i(t) \in \partial_x(g_i(t, x_0(t))(0), \quad \lambda_i\text{-a.e.;} \tag{4.6}$$

the right hand side means the subdifferential of the function $x \mapsto g_i(t, x_0(t)+x)$ at 0.

In order to calculate $\partial J_1(0, 0)$ we write J_1 in the form

$$J_1(x, u) = l^0(x_0(0), x_0(1); \cdot, \cdot) \circ (E_0, E_1)(x, u),$$

where $(E_0, E_1)(x, u) = (x(0), x(1))$. Then, from subdifferential calculus, $\partial J_1(0, 0) = (E_0, E_1)^*(\partial l^0(x_0(0), x_0(1); 0, 0))$, the asterisk denoting transposition. Hence, for every pair $(\rho_1, \sigma_1) \in \partial J_1(0, 0)$ there are n-vectors p_1, q_1 such that

$$\langle x, \rho_1\rangle + \langle u, \sigma_1\rangle = \langle x(0), p_1\rangle + \langle x(1), q_1\rangle$$

and

$$(p_1, q_1) \in \partial l^0(x_0(0), x_0(1); 0, 0). \tag{4.7}$$

Collecting all these results we see that $(0, 0) \in \partial J(0, 0)$ implies

$$\begin{aligned} &\langle x(0), p_1\rangle + \langle x(1), q_1\rangle + \int_0^1 [\langle x(t), p_2(t)\rangle + \langle u(t), q_2(t)\rangle]\, dt \\ &+ \sum_{i=1}^{k} \int_0^1 \langle x(t), p_3^i(t)\rangle\, d\lambda_i(t) + \sum_{i=1}^{n} \int_0^1 \left[x_i(t) - \int_0^t u_i(s)\, ds\right] d\mu_i(t) = 0 \end{aligned} \tag{4.8}$$

for all $(x, u) \in C^n \times L_\infty^n$. For pairs $(x, \dot{x})$ with $x(0) = 0$, $x(1) = 0$, this equation reduces to

$$\begin{aligned} &\int_0^1 \langle x(t), p_2(t)\rangle\, dt + \sum_{i=1}^{k} \int_0^1 \langle x(t), p_3^i(t)\rangle\, d\lambda_i(t) \\ &+ \int_0^1 \langle \dot{x}(t), q_2(t)\rangle\, dt = 0. \end{aligned} \tag{4.9}$$

By using a straightforward modification of Hestenes' proof of the fundamental lemma in the calculus of variation it can be shown that (4.9) implies

$$q_2^j(t) = \int_0^t p_2^j(s)\, ds + \sum_{i=1}^{k} \int_0^t p_{3,j}^i(s)\, d\lambda_i(s) + c_j \quad \text{a.e.,} \tag{4.10}$$

where

$$c_j = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon} \int_0^\varepsilon q_2^j(t)\, dt. \tag{4.11}$$

Now consider functions x_j^ε, $j = 1, \cdots, n$, defined by

$$x_{j,j}^\varepsilon(t) = \begin{cases} \dfrac{1}{\varepsilon}(\varepsilon - t) & \text{for } 0 \leqq t \leqq \varepsilon, \\ 0 & \text{for } \varepsilon \leqq t \leqq 1, \end{cases}$$

$$x_{j,i}^\varepsilon \equiv 0 \quad \text{for } i \neq j.$$

Inserting the pairs $(x_j^\varepsilon, \dot{x}_j^\varepsilon)$ into (4.8) we obtain

$$p_1^j + \int_0^\varepsilon x_{j,j}^\varepsilon(t) p_2^j(t)\, dt - \frac{1}{\varepsilon}\int_0^\varepsilon q_2^j(t)\, dt + \sum_{i=1}^k \int_0^\varepsilon x_{j,j}^\varepsilon(t) p_{3,j}^i(t)\, d\lambda_i(t) = 0$$

and, passing to the limit,

$$p_1^j = c_j - \sum_{i=1}^k p_{3,j}^i(0)\lambda_i(\{0\}). \tag{4.12}$$

Similar considerations concerning the right endpoint lead to the equation

$$q_1^j = c_j' - \sum_{i=1}^k p_{3,j}^i(1)\lambda_i(\{1\}), \tag{4.13}$$

where

$$c_j' = -\lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon}\int_{1-\varepsilon}^1 q_2^j(t)\, dt. \tag{4.14}$$

Let us sum up these results in

THEOREM 2. *Suppose that the following assumptions hold*:

(i) *L is measurable in t and Lipschitz in (x, u) near x_0;*

(ii) *l is locally Lipschitz at $(x_0(0), x_0(1))$;*

(iii) *g is continuous in (t, x) and convex in x, and there exists an absolutely continuous function x_1 such that $g_i(t, x_0(t) + x_1(t)) < 0$ for all $t \in I_i$ satisfying $g_i(t, x_0(t)) = 0$, $i = 1, \cdots, k$.*

Then a necessary condition for $(x_0, \dot{x}_0)$ to be an optimal solution to the problem (4.2) *is the existence of n-vectors p_1, q_1, $\mathbb{R}^n$-valued measurable functions $p_2(t), q_2(t)$ with integrable components, k nondecreasing functions λ_i satisfying* (4.3) *and k measurable $\mathbb{R}^n$-valued functions $p_3^i(t)$, each with λ_i-integrable components, such that*

$$(p_1, q_1) \in \partial^0 l(x_0(0), x_0(1)), \tag{4.15}$$

$$(p_2(t), q_2(t)) \in \partial^0 L(t, x_0(t), \dot{x}_0(t)) \quad a.e., \tag{4.16}$$

$$p_3(t) \in \partial_x(g(t, x_0(t)), \quad \lambda\text{-}a.e., \tag{4.17}$$

$$q_2(t) = \int_0^t p_2(s)\, ds + \int_0^t P_3(s)\, d\lambda(s) + c \quad a.e. \tag{4.18}$$

and the "transversality conditions"

$$p_1 = c - P_3(0)\lambda(\{0\}), \tag{4.19}$$

$$q_1 = c' - P_3(1)\lambda(\{1\}) \tag{4.20}$$

hold. Here $P_3(t)$ is the $n \times k$ matrix whose i-th column is $p_3^i(t)$.

Proof. Expressions (4.17)–(4.20) are just symbolic notations for the corresponding relations derived above. Expressions (4.15) and (4.16) follow from (4.7) and (4.5), respectively, by considering Proposition 8. □

In case there are no phase restrictions ($\lambda = 0$), (4.18) shows that $q_2(t)$ is absolutely continuous; therefore 0 and 1 are Lebesgue-points of q_2, and (4.19) and (4.20) reduce to $p_1 = q_2(0)$ and $q_1 = -q_2(1)$, respectively (compare [2]).

Now suppose that there are additional constraints of the form $(x(0), x(1)) \in S$, where the set S is given by the relations

$$l_i(x(0), x(1)) \leqq 0, \qquad i = 2, \cdots, r,$$

with real valued functions l_i. Assuming that these functions are locally Lipschitz at $(x_0(0), x_0(1))$ and regarding them as additional components of the operator G we find that there exist a vector $(\beta_1, \cdots, \beta_r)$ with nonnegative components and r pairs of n-vectors

$$(p_1^i, q_1^i) \in \beta_i \partial^0 l_i(x_0(0), x_0(1)) \tag{4.21}$$

(with $l_1 = l$) such that

$$\beta_i l_i(x_0(0), x_0(1)) = 0, \qquad i = 2, \cdots, r,$$

and (4.8) holds with

$$\sum_{i=1}^{r} (\langle x(0), p_1^i \rangle + \langle x(1), q_1^i \rangle)$$

instead of the first two terms of the sum.

Note that now the multiplier belonging to l_1 may vanish because no constraint qualification concerning the l_i has been made as yet (of course the constraint qualification for the g_i can then be dispensed with, too), thus maintaining the possibility of including equalities. Proceeding as above it is shown that all the results of Theorem 2 hold with (4.15) replaced by (4.21), (4.19) and (4.20) by

$$\sum_{i=1}^{r} p_1^i = c - P_3(0)\lambda(\{0\}) \tag{4.22}$$

and

$$\sum_{i=1}^{r} q_1^i = c' - P_3(1)\lambda(\{1\}), \tag{4.23}$$

respectively. For convex l_i, (4.21) is equivalent to

$$\langle s_0 - x_0(0), p_1^i \rangle + \langle s_1 - x_0(1), q_1^i \rangle \leqq 0, \qquad i = 1, \cdots, r,$$

for all $(s_0, s_1) \in S$.

REFERENCES

[1] M. S. BAZARAA AND J. J. GOODE, *Necessary optimality criteria in mathematical programming in normed linear spaces*, J. Optimization Theory Appl., 3 (1973), pp. 235–244.

[2] F. H. CLARKE, *The Euler–Lagrange differential inclusion*, J. Differential Equations, 19 (1975), pp. 80–90.

[3] ———, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[4] J. V. GIRSANOV, *Lectures on Mathematical Theory of Extremum Problems*, Springer Lecture Notes in Economics and Mathematical Systems, 67, Springer-Verlag, Berlin, 1972.

[5] H. HALKIN, *A satisfactory treatment of equality and operator constraints in the Dubovitskii–Milyutin formalism*, J. Optimization Theory Appl., 2 (1970), pp. 138–149.

[6] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
[7] R. B. HOLMES, *A Course on Optimization and Best Approximation*, Springer-Verlag, New York, 1972.
[8] A. D. IOFFE AND V. L. LEVIN, *Subdifferentials of Convex Functions*, Trans. Moscow Math. Soc., 26 (1972), pp. 1–72.
[9] P. J. LAURENT, *Approximation et Optimisation*, Hermann, Paris, 1972.
[10] L. A. LIUSTERNIK AND V. J. SOBOLEV, *Elemente der Funktionalanalysis*, Akademie-Verlag, Berlin, 1965.
[11] L. W. NEUSTADT, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.
[12] A. L. PERESSINI, *Ordered Topological Vector Spaces*, Harper and Row, New York, 1967.
[13] R. T. ROCKAFELLAR, *Conjugate convex functions in optimal control and the calculus of variation*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.
[14] ———, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–39.
[15] ———, *State constraints in convex control problems of Bolza*, this Journal, 4 (1972), pp. 691–715.
[16] C. VIRSAN, *Necessary conditions for optimization problems with operatorial constraints*, this Journal, 4 (1970), pp. 527–554.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 4, July 1977

A COMPARISON OF THE FORCING FUNCTION AND POINT-TO-SET MAPPING APPROACHES TO CONVERGENCE ANALYSIS*

R. R. MEYER†

CONTENTS

Abstract. A "forcing function" approach is developed for the analysis of convergence properties of "monotonic" mathematical programming algorithms. This approach differs from rather more traditional analyses based on point-to-set mappings in that it does *not* require point-to-set mapping concepts. A comparison is given between the forcing function and point-to-set mapping approaches that indicates that they are essentially mathematically equivalent for two major categories of algorithms, but that only the forcing function approach is readily extended to a third category of algorithms involving anti-jamming parameters.

1. Introduction. The point-to-set mapping approach to the qualitative analysis of convergence of mathematical programming algorithms is well-known, having been considered and promoted in many papers and books. (The reader entirely unfamiliar with the field can get a "feel" for the area by consulting [5], [7], [10], [11], [15] or [20] or the excellent survey article by Hogan [4].) On the other hand, analyses of convergence that do *not* rely on properties of point-to-set mappings have also been numerous (see, for example, [2], [8], [12], [13], [16], [18], [21]), but have not considered algorithms at the level of generality of the point-to-set mapping approach. This report proposes a "forcing function" approach to convergence analysis that not only is more general than the best-known point-to-set mapping results, but moreover, is often easier to apply and may be readily extended to handle a class of algorithms that do not appear to be amenable to analysis by a straightforward point-to-set mapping approach. (These are constrained optimization methods involving "anti-jamming" parameters.) The forcing function approach has the further pedagogical advantage of requiring only certain continuity properties of real-valued functions rather than of point-to-set mappings.

In the results to be obtained below, the sequence of iterates $\{x_i\}$ should be thought of as resulting from the application of an iterative algorithm. Such sequences will always be assumed to be contained in a *closed* set $G \subseteq \mathbb{R}^n$. Unless otherwise specified, the domain of all functions and mappings considered below

* Received by the editors December 29, 1975, and in revised form September 10, 1976.

† Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706. This work was supported in part by the National Science Foundation under Grant DCR74-20584.

will be G.(For unconstrained optimization algorithms an appropriate G might be $\mathbb{R}^n$, whereas for constrained optimization, the set G is often taken as the feasible region. Many of the results below also remain valid when G is a closed subset of an arbitrary topological space, and this will be pointed out when it is the case.)

The notation $q_i \xrightarrow{I} q$ is to be understood to mean that I is an infinite increasing sequence of nonnegative integers and that $\lim_{i\to\infty, i\in I} q_i$ is q.

2. The forcing function approach. Roughly speaking, the *nonnegative* function δ to be considered below will play the role of an "optimality indicator" in the class of algorithms to be described, in that an iterate x_i may be repeated only if $\delta(x_i)=0$. When $\delta(x_i)>0$, the next iterate, x_{i+1}, will be required to have a "value improvement" (in terms of a particular function ϕ to be introduced below) of at least $\delta(x_i)$.

The term "forcing function approach" is used since, on the one hand, δ "forces" an improvement if $\delta(x_i)>0$, and, on the other hand, convergence of a sequence of function values of ϕ will be seen to "force" the sequence $\{\delta(x_i)\}$ to converge to 0. (The term "forcing function" is used by Ortega and Rheinboldt [13] to describe a related but slightly different property. Specifically, they define a mapping $\sigma\colon [0,\infty)\to[0,\infty)$ to be a forcing function if, for any sequence $\{t_k\}\subset[0,\infty)$, the property $\sigma(t_k)\to 0$ implies $t_k\to 0$; they then analyze a family of optimization methods in which $\delta(x_i)$ is equal to a forcing function of a certain scalar function of x_i. See the Appendix for further comparisons.)

Let $\delta\colon G\to R^1_+$ and define

$$\Omega^L \equiv \{x \mid \exists\{y_i\}\subset G \text{ with } y_i\to x,\ \delta(y_i)\to 0\}, \tag{2.1}$$

$$\Omega^* \equiv \{x \mid x\in G,\ \delta(x)=0\}, \tag{2.2}$$

$$\Omega^+ \equiv \{x \mid x\in G,\ \delta(x)>0\}. \tag{2.3}$$

Since δ is assumed to be nonnegative, note that $G=\Omega^*\cup\Omega^+$.

Thinking of δ as an "optimality indicator" that is 0 at points satisfying some optimality condition, it is clearly desirable for an iterative algorithm to have the property that its iterates must converge to a point in Ω^*. In order to achieve this strong result (given as Corollary 2.5 below) a number of hypotheses are needed. By assuming only some of the hypotheses of Corollary 2.5, however, weaker results that are of some interest in themselves are obtained, and we shall first develop these weaker results.

LEMMA 2.1. *Let $\phi\colon G\to R^1$ be a function that is* lower semi-continuous (*l.s.c.*) *on G, and let $\{x_i\}$ be a sequence with the property that*

$$\phi(x_i)-\phi(x_{i+1})\geqq\delta(x_i), \qquad i=0,1,2,\cdots. \tag{2.4}$$

If $\{x_i\}$ has an accumulation point $\bar{x}$, then $\delta(x_i)\to 0$ and $\bar{x}\in\Omega^L$.

Proof. If $\bar{x}$ is an accumulation point of $\{x_i\}$, it follows from (2.4), the nonnegativity of δ, and the l.s.c. of ϕ, that $\phi^*\equiv\lim\phi(x_i)$ exists and that $\phi(x_i)\geqq\phi(\bar{x})$ for all i. Thus $\phi(x_i)-\phi(x_{i+1})\to 0$, $\delta(x_i)\to 0$, and if I is the index set corresponding to the subsequence of $\{x_i\}$ converging to $\bar{x}$, then $x_i\xrightarrow{I}\bar{x}$ and $\delta(x_i)\xrightarrow{I}0$, proving that $\bar{x}\in\Omega^L$. □

It might be noted that the sole continuity hypothesis, namely, l.s.c. of ϕ on G, could be replaced by the hypothesis that *ϕ is bounded from below on G*, since the l.s.c. of ϕ is only needed to establish convergence of $\{\phi(x_i)\}$. However, in most applications the corresponding ϕ is at least continuous, and often continuously differentiable, so the l.s.c. hypothesis appears preferable to the boundedness hypothesis. Since no special properties of $\mathbb{R}^n$ were used, note that Lemma 2.1 remains true if G is a subset of some topological space. The same observation also applies to Lemma 2.2 and Theorem 2.3 below.

It might also be noted that Lemma 2.1 is similar in some respects to some results of Eaves and Zangwill [3]. They develop a theory of *cutting plane* algorithms by assuming that the *distance* between an iterate and certain prior iterates is bounded from below by a nonnegative "separator" function δ, that has the property that if $z_i \to z$ and $\delta(z_i) \to 0$, then z must be in what is termed "the goal set". In Lemma 2.1, δ is a lower bound for the *change* in an *arbitrary* function ϕ, and Ω^L itself plays the role of the "goal" set, rather than being a subset of a "goal" set that is never characterized. Since convergence to a point outside of Ω^L is also impossible under the hypotheses made by Eaves and Zangwill, it seems inappropriate to describe any point outside of Ω^L as being a "goal" of the algorithm.

A comparison of Lemma 2.1 with the more closely related results of Zangwill [20] and Polak [15] is given the Appendix. Again the conclusion that the accumulation points must lie in Ω^L turns out to be a sharper result than membership in a so-called "solution" or "desirable" set.

In most specific applications, the function δ turns out to have certain continuity properties that allow a strengthening of the conclusion of Lemma 2.1. In particular, the weak continuity property that we will now introduce turns out to be satisfied by most optimality indicators that arise in practice (the only significant exception seems to arise from optimality indicators associated with certain feasible direction methods; this point will be taken up in § 5).

A scalar-valued function ω is said to be *null-continuous* or *C_0 at a point z* if the existence of a sequence $\{y_i\}$ with $y_i \to z$ and $\omega(y_i) \to 0$ implies $\omega(z) = 0$. The function is said to be C_0 on a set if it is C_0 at each point of the set. (As will be seen, this continuity concept for scalar functions will replace point-to-set mapping continuity properties in the convergence theorems to be developed. From a pedogogical standpoint, it also appears preferable to introduce convergence analysis through the use of continuity properties of functions rather than continuity properties of point-to-set mappings, since students often have difficulties in obtaining a feeling for point-to-set mappings.)

Note that null-continuity of δ on G is a weaker property than lower semi-continuity of δ on G (assuming $G \neq \Omega^*$), but a stronger property than lower semi-continuity on Ω^*. (It is, in fact, *equivalent* to the relation $\Omega^L = \Omega^*$ and also equivalent to l.s.c. of δ on Ω^L.) For many well-known constrained optimization algorithms, the corresponding optimality indicator is *not* lower semi-continuous on all of G, but is a function for which $\Omega^L = \Omega^*$ may be established.

(It might also be noted that if δ is nonnegative and null-continuous on G, then there exists a function δ_L defined on G such that (i) $0 \leqq \delta_L(x) \leqq \delta(x)$ for all $x \in G$, (ii) δ_L is l.s.c. on G, and (iii) $\{x | \delta_L(x) = 0\} = \Omega^*$. In fact, δ_L may be defined at each $x \in G$ by the equation $\delta_L(x) \equiv \inf\{\theta | \exists \{y_i\}, y_i \to x, \delta(y_i) \to \theta\}$, and properties (i), (ii),

and (iii) are easily verified. From the standpoint of application, however, null-continuity is a weaker requirement than l.s.c. and one that may often be verified more easily than l.s.c.)

As an immediate consequence of Lemma 2.1 and the property that $\Omega^L = \Omega^*$ when δ is null-continuous we have:

LEMMA 2.2. *Let ϕ be l.s.c. on G, let δ be a C_0 function on G, and let $\{x_i\}$ be a sequence satisfying* (2.4). *If $\{x_i\}$ has an accumulation point $\bar{x}$, then $\bar{x} \in \Omega^*$.*

Example 1. As a simple example of the forcing function approach we will consider the method of steepest descent with an Armijo-type step-size. (Examples of the application of this approach to constrained optimization methods may be found in [8] and in Chung [1], where an interesting application to exact penalty methods is made.) We will assume that ϕ is continuously differentiable on all of $\mathbb{R}^n$, that $G = \mathbb{R}^n$, and that, given a point z, its successor z' is uniquely determined by the relations:

$$z' = z - \lambda \nabla \phi(z)^T,$$

where

$$\lambda = \max\{\lambda \mid \lambda = 2^{-i},\ i = 0, 1, 2, \cdots,\ \phi(z) - \phi(z - \lambda \nabla \phi(z)^T) \geqq \tfrac{1}{2}\lambda \|\nabla \phi(z)\|^2\}.$$

It is easily shown that a suitable optimality indicator for this example is obtained by setting

$$\delta(z) = L(z)\|\nabla \phi(z)\|^2 = \phi(z) - \phi(z'),$$

where L has the property that if $y_i \to z$, and $\nabla \phi(z) \neq 0$, then $\liminf L(y_i) > 0$. Here, δ is nonnegative and null-continuous on $\mathbb{R}^n$, since $y_i \to z$ and $\delta(y_i) \to 0$ imply $\nabla \phi(y_i) \to 0$, and thus $\nabla \phi(z) = 0$. Note that $\Omega^L = \Omega^* = \{x \mid \nabla \phi(x) = 0\}$. Lemma 2.2 thus guarantees that if the sequence $\{x_i\}$ has an accumulation point $\bar{x}$, then $\nabla \phi(\bar{x}) = 0$. (Boundedness of $\{x_i\}$ may be guaranteed by appropriate level set compactness hypotheses on ϕ.) □

The preceding example also illustrates that the application of the convergence theorems of this section to a given algorithm may require the determination of an appropriate function δ. In order that the conclusions of the theorems be as sharp as possible, the sets in which the accumulation points are contained should be as small as possible, which means that the best choice for δ is the supremum of all functions for which (2.4) will be satisfied. In the case of Example 1, this is the motivation for the definition of $\delta(z)$. A general procedure for obtaining an appropriate δ for algorithms in which δ is not given explicitly is discussed in § 3.

Of course, if $\{x_i\}$ has no accumulation point, then there is no guarantee that Ω^* is nonempty or that $\delta(x_i) \to 0$. Examples are easily constructed with these properties. Furthermore, even if $\{x_i\}$ is bounded, there is no guarantee that the $\{x_i\}$ will converge to a unique accumulation point, and, in fact, "oscillatory" behavior of the $\{x_i\}$ between a finite number of accumulation points may occur (see [10]). (Indeed, under certain weak hypotheses, the existence of a sequence satisfying the hypotheses of Lemma 2.2 yet displaying oscillatory behavior can be guaranteed, as will be shown in Theorem 4.3). From a practical point of view, however, oscillatory behavior is quite rare, suggesting that some additional properties are

generally satisfied which prevent such bad behavior. Thus, we are led to consider additional hypotheses under which convergence of the sequence $\{x_i\}$ may be demonstrated. The most obvious (but least applicable) such hypothesis is given in Corollary 2.3 below, and a more useful approach is presented in Theorem 2.4 and Corollary 2.5. (In all of the remaining results of this section, the compactness of closed, bounded subsets of $\mathbb{R}^n$ is exploited, so a direct extension to more general spaces is not possible.)

COROLLARY 2.3. *Let the hypotheses of Lemma* 2.2 *hold, and assume in addition that ϕ is continuous, that $\{x_i\}$ is bounded, and that, for each $\theta \in R$, the set $\Omega^* \cap \{x | x \in G, \phi(x) = \theta\}$ contains at most one element; then there exists an $x^* \in \Omega^*$ such that $x_i \to x^*$.*

Proof. If the result were false, then there would be two subsequences with index sets I and J such that $x_i \xrightarrow{I} x'$ and $x_i \xrightarrow{J} x''$ with $x' \neq x''$. By Lemma 2.2, x' and x'' are in Ω^*, and by the monotonicity of $\{\phi(x_i)\}$, $\phi(x') = \phi(x'')$, a contradiction. □

The disadvantage to the above approach to proving convergence is that it requires what amounts to a global uniqueness hypothesis. This type of hypothesis is usually not verifiable except under strict convexity hypotheses. On the other hand, since most algorithms perform only a local search at each iteration, their convergence properties are generally determined by the local behavior of the function to be minimized. To make these notions precise, we will introduce in Theorem 2.4 below a "stability" hypothesis (2.5) that, in effect, "damps" the step-length $\|x_i - x_{i+1}\|$ when x_i is near Ω^*.

THEOREM 2.4. *Let the hypotheses of Lemma* 2.1 *hold, let $\{x_i\}$ be bounded, and,* in addition, *assume that there exist functions ρ and μ such that $\mu: \mathbb{R}^n \to \mathbb{R}^1_+$, μ is a C_0 function defined on $\mathbb{R}^n$ with $\mu(x) > 0$ for $x \neq 0$, ρ is a function from G into $\mathbb{R}^1_+$ such that $\delta(x_i) \to 0$ implies $\rho(x_i) \to 0$, and for $i = 0, 1, \cdots$,*

$$\rho(x_i) \geqq \mu(x_i - x_{i+1}). \tag{2.5}$$

Then $\|x_i - x_{i+1}\| \to 0$, and the set of accumulation points of $\{x_i\}$ consists of a single point or *a continuum.*

Proof. Suppose that $\|x_i - x_{i+1}\| \not\to 0$. Then there exists an index set J such that $x_i \xrightarrow{J} x^*$, $x_{i+1} \xrightarrow{J} x^{**}$ with $x^* \neq x^{**}$. By Lemma 2.1, $\delta(x_i) \to 0$, so (2.5) implies $\mu(x_i - x_{i+1}) \to 0$, and thus $\mu(x^* - x^{**}) = 0$. However, $\|x^* - x^{**}\| > 0$ implies $\mu(x^* - x^{**}) > 0$, a contradiction. Thus $\|x_i - x_{i+1}\| \to 0$, and because of the boundedness of $\{x_i\}$, the remaining conclusion is a well-known result of Ostrowski [14]. □

Example 2. If the sequence $\{x_i\}$ was derived according to the procedure described in Example 1, then by taking $\mu(z) = \|z\|$ and $\rho(z) = \|\nabla\phi(z)\|$, it is easily seen that the iterates of Example 1 satisfy (2.5) and that ρ and μ satisfy the hypotheses of Theorem 2.4. □

The following Corollary is an immediate consequence of Lemma 2.2 and Theorem 2.4 and establishes sufficient conditions for convergence of the *entire* sequence $\{x_i\}$ to a point in Ω^*.

COROLLARY 2.5. *Let $\{x_i\}$ be a* bounded *sequence satisfying*:

(a) $\phi(x_i) - \phi(x_{i+1}) \geqq \delta(x_i)$ $(i = 0, 1, 2, \cdots)$, *where ϕ is l.s.c. on G and δ is C_0 on G, and*

(b) $\rho(x_i) \geqq \mu(x_i - x_{i+1})$ $(i = 0, 1, 2, \cdots)$, *where* $\rho(x) = 0$ *for* $x \in \Omega^*$ *and* ρ *is continuous at each* $x \in \Omega^*$, *and* μ *is* C_0 *on* R^n *and satisfies* $\mu(x) > 0$ *for* $x \neq 0$.

If for each $x^* \in \Omega^*$, *there exists an open set* $N(x^*)$ *containing* x^* *such that* $N(x^*) \cap \Omega^* = \{x^*\}$, *then* $\{x_i\}$ *converges to a point of* Ω^*.

Proof. By Theorem 2.4, *either* $\{x_i\}$ converges to a point in Ω^* *or* its accumulation points form a continuum contained in Ω^*. However, by hypothesis, Ω^* consists of isolated points and hence cannot contain a continuum. □

Example 3. Again let $\{x_i\}$ be as in Example 1, and suppose that $\phi \in C^2$ and that $\nabla^2\phi(x)$ is nonsingular if $x \in \Omega^*$. Then for each $x \in \Omega^*$, there exists an open set $N(x)$ such that $N(x) \cap \Omega^* = \{x\}$. (For, otherwise, there would be an $x^* \in \Omega^*$ and a sequence $\{y_i\}$ with $y_i \to x^*$ and $\nabla\phi(y_i) = 0$. Without loss of generality, we may assume the sequence $\{(y_i - x^*)/\|y_i - x^*\|\}$ converges to d, where $\|d\| = 1$, and it is then easily shown that $\nabla^2\phi(x^*)d = 0$, contradicting the nonsingularity of $\nabla^2\phi$ on Ω^*.) □

Note that Corollary 2.5 is a *global* convergence theorem, i.e., it guarantees convergence to a point in Ω^* from an *arbitrary* starting point x_0 of G provided that the monotonicity and localization hypotheses are satisfied by the iterates and by Ω^*. Point-of-attraction theorems establishing *local* convergence under somewhat weaker hypotheses as well as convergence theorems that make use of the properties of accumulation points of $\{x_i\}$ may be found in [10], but it should be recognized that for global convergence, global hypotheses are required. The main results of this section are summarized in Table 1, which also indicates the results to be obtained in §§ 3 and 4.

TABLE 1

Properties of the point-to-set mapping S	Properties of the functions δ, ρ, μ	Convergence results
Monotonicity	Non-negativity of δ	Accumulation points are in Ω^L.
$x^* \in S(x^*)$(x^* a GFP)	$\delta(x^*) = 0$	x^* can be an accumulation point.
Monotonicity plus sequential monotonicity at non-GFP's	Nonnegativity and null-continuity of δ	Each accumulation point $\hat{x}$ is a GFP and $\delta(\hat{x}) = 0$ $(\hat{x} \in \Omega^*)$.
$\{x^*\} = S(x^*)$(x^* an SFP)	$\delta(x^*) = 0$; $\rho(x^*) = 0$	If $x_0 = x^*$, then $x_i = x^*$ for all i.
Monotonicity plus sequential monotonicity at non-SFP's plus u.s.c. at SFP's	Nonnegativity and null-continuity of δ; ρ is continuous and equals 0 at points x such that $\delta(x) = 0$; μ is null-continuous and positive-definite	If $\{x_i\}$ is bounded and the SFP's do not form a continuum, then $\{x_i\}$ *converges* to an $\hat{x}$ such that $\delta(\hat{x}) = \rho(\hat{x}) = 0$ ($\hat{x}$ is an SFP).

3. Relationships between forcing functions and point-to-set mappings. Given a pair of functions ϕ, δ defined on G, the *algorithm corresponding to ϕ and δ* is defined to be the algorithm given by:

(3.1) choose an arbitrary $x_0 \in G$,

and

(3.2) given x_i, choose x_{i+1} such that
$\phi(x_i) - \phi(x_{i+1}) \geqq \delta(x_i)$, $i = 0, 1, 2, \cdots$.

Clearly, this algorithm will be well-defined if and only if the set defined by

(3.3) $$S(x) \equiv \{y \mid y \in G, \phi(x) - \phi(y) \geqq \delta(x)\}$$

is *nonempty* for *all* $x \in G$. Having so defined the point-to-set mapping S, this algorithm could also be thought of as *the algorithm corresponding to S* since (3.2) could be replaced by the statement

(3.4) given x_i, choose $x_{i+1} \in S(x_i)$, $i = 0, 1, 2, \cdots$.

Thus, we could attempt to analyze this algorithm *either* by considering the properties of ϕ and δ and applying the results of the previous section, *or* by considering properties of S, and applying point-to-set mapping convergence theorems. In this section we will discuss what properties of S are, in some sense, equivalent to certain properties of ϕ and δ, and develop point-to-set mapping convergence theorems analogous to the convergence theorems of § 2. (The results to be established in this section are summarized in Table 1.)

As shown in the previous section, some convergence properties can be proved if we merely assume that ϕ is l.s.c. on G and that δ is a nonnegative C_0 function on G.

These properties of ϕ and δ, however, do *not* imply semi-continuity properties for S. S, for example, may fail to be upper semi-continuous (or "closed" as this property is sometimes described) as a result of discontinuities in ϕ and/or δ. (A mapping T from G into the subsets of G will be said to be *u.s.c. at a point x* if $\{x_i\} \subseteq G$, $x_i \to x$, $y_i \in T(x_i)$, and $y_i \to y$ imply $y \in T(x)$, *and u.s.c. on a subset of G* if it is u.s.c. at every point in that subset.) In order to see what properties may be claimed for S in this case, some additional notation will be introduced.

Let T be a point-to-set mapping from G into the subsets of G. A point x^* is defined to be a *generalized fixed-point* (GFP) of T if $x^* \in T(x^*)$, and a *strong fixed-point* (SFP) of T if $T(x^*) = \{x^*\}$. (Clearly every SFP is also a GFP, but the converse will not hold if $\{x^*\}$ is a proper subset of $T(x^*)$. As will be seen below, there is a correspondence between the set of GFP's and the set of points on which a related optimality indicator vanishes, and a correspondence between the set of SFP's and the set of points on which both an optimality indicator and a certain distance majorant vanish.)

T is said to be *monotonic* on G w.r.t. a function $\omega: G \to R^1$ if $y \in T(x)$ implies $\omega(y) \leqq \omega(x)$. T will be said to be *sequentially monotonic* w.r.t. ω on a set $M \subseteq G$ if $x \in M$, $x_i \to x$, $y_i \in T(x_i)$, $\omega(x_i) \to \omega^*$, and $\omega(y_i) \to \bar{\omega}$ imply $\bar{\omega} < \omega^*$. (Note that if x^* is

a GFP or SFP of T, then x^* cannot be in the set M on which T is sequentially monotonic, because we may take $x_i \equiv y_i \equiv x^*$, and the strict inequality in the definition of sequential monotonicity is not satisfied. In the convergence results below, the mappings considered will be sequentially monotonic at all points other than GFP's or SFP's. Note also that if $x \in M$ and $y \in T(x)$, then $\omega(y) < \omega(x)$, but that sequential monotonicity is a *stronger* property than the simple requirement that $\omega(y) < \omega(x)$ whenever $x \in M$ and $y \in T(x)$.)

Our first result using these definitions indicates the properties induced by requiring δ to be a nonnegative C_0 function on G.

THEOREM 3.1. *If δ is a nonnegative C_0 function on G, then* (a) S *is* monotonic *on G w.r.t. ϕ, and* (b) S *is* sequentially monotonic *w.r.t. ϕ on G/Ω^*, and* (c) Ω^* *is the* set of GFP's *of S.*

Proof. Properties (a) and (c) follow directly from the definitions, so we will exhibit the proof for (b) only. Let $x \in \Omega^+$, $x_i \to x$, $y_i \in S(x_i)$, $\phi(x_i) \to \phi^*$, $\phi(y_i) \to \bar{\phi}$. By the nonnegativity of δ, $\bar{\phi} \leqq \phi^*$. If $\bar{\phi} = \phi^*$, then $\delta(x_i) \to 0$ and thus $x \in \Omega^*$, a contradiction, so $\bar{\phi} < \phi^*$. □

Let the functions ρ and μ be defined on G and $\mathbb{R}^n$ respectively, and define the point-to-set mapping

$$\hat{S}(x) \equiv \{y \mid y \in S(x), \rho(x) \geqq \mu(x-y)\}. \tag{3.5}$$

Note that S is a *restriction* of $\hat{S}$, by which we mean that $\hat{S}(x) \subseteq S(x)$ for all $x \in G$. As a restriction of S, it is easily seen that conclusions (a) and (b) of Theorem 3.1 must continue to hold when S is replaced by $\hat{S}$. We will now show that conclusion (c) may be strengthened in a useful manner if appropriate properties are assumed for ρ and μ. (These properties are essentially those used in Theorem 2.4.)

THEOREM 3.2. *Let δ be a nonnegative C_0 function on G, let ρ be a nonnegative function on G such that ρ is continuous on Ω^* and $\rho(x^*) = 0$ if $x^* \in \Omega^*$, and let μ be a nonnegative C_0 function on $\mathbb{R}^n$ with $\mu(z) > 0$ if $z \neq 0$. Then* (a) $\hat{S}$ *is sequentially monotonic w.r.t. ϕ on G/Ω^*, and* (b) S *is* u.s.c. *on Ω^*, which is the* set of SFP's *of $\hat{S}$.*

Proof. As noted previously, conclusion (a) follows from the observation that $\hat{S}$ is a restriction of S; we thus need only prove (b). If $x^* \in \Omega^*$, then $\rho(x^*) = 0$, so the inequality $\rho(x^*) \geqq \mu(x^* - y)$ and the positive-definite property of μ force $y = x^*$ when $y \in \hat{S}(x^*)$, so x^* must be a SFP of $\hat{S}$. Conversely, if x^* is a SFP of $\hat{S}$, then $x^* \in \hat{S}(x^*)$ implies $\delta(x^*) = 0$. Now suppose also that $z_i \to x^*$, $y_i \in \hat{S}(z_i)$, $y_i \to y^*$. Since $\rho(x^*) = 0$ and ρ is continuous on Ω^*, $\rho(z_i) \to 0$, and thus $\mu(z_i - y_i) \to 0$. Since μ is a C_0 function, $\mu(x^* - y^*) = 0$, and thus $y^* = x^* \in S(x^*)$. □

Having derived properties of S and $\hat{S}$ that are induced by properties of the functions appearing in their definitions, we will now take the opposite point of view, and, given a point-to-set mapping T with certain properties, we will show that related functions δ, ρ, and μ with the properties introduced in § 2 may be constructed.

Let T be a point-to-set mapping from G into its subsets. If T is monotonic on G w.r.t. a function ϕ, we define the nonnegative *optimality indicator* corresponding to T and ϕ to be

$$\delta^*(x) = \inf_{y \in T(x)} (\phi(x) - \phi(y)). \tag{3.6}$$

Also, we define the *distance majorant* associated with T to be the *extended real-valued* function on G defined by

$$\rho^*(x) = \sup_{y \in T(x)} \|y - x\|. \tag{3.7}$$

The set of GFP's of T is denoted by Γ^* and the set of SFP's of T is denoted by Γ^{**}.

Our first result employing these definitions gives sufficient conditions for the optimality indicator to be C_0 on G.

THEOREM 3.3. *Let T be* monotonic on G *w.r.t.* ϕ, *and let T be* sequentially monotonic on G/Γ^* *w.r.t.* ϕ. *If* ϕ *is continuous* or *if* ϕ *is l.s.c. and bounded from above on G, then* δ^* *is a nonnegative* C_0 *function on G, and* $\{x \mid \delta^*(x) = 0\} = \Gamma^*$.

Proof. Clearly if $x^* \in \Gamma^*$, then $\delta^*(x^*) = 0$. On the other hand, if $\delta^*(\bar{x}) = 0$, then by choosing $x_i = \bar{x}$ for all i, letting $\{y_i\}$ be such that $\{y_i\} \subseteq T(\bar{x})$ and $\phi(y_i) \to \phi(\bar{x})$, and exploiting the sequential monotonicity property of T on G/Γ^*, we conclude that $\bar{x} \notin G/\Gamma^*$, so $\bar{x} \in \Gamma^*$.

If $\{z_i\} \subseteq G$ and $z_i \to \hat{x}$, then because of the hypotheses on ϕ, there exists a subsequence of $\{z_i\}$ such that $\phi(z_i)$ is convergent. So if $x_i \to x$ and $\delta^*(x_i) \to 0$, without loss of generality we may assume that there exists a ϕ^* such that $\phi(x_i) \to \phi^*$ and a sequence $\{y_i\}$ such that $y_i \in T(x_i)$ for each i and $\phi(y_i) \to \phi^*$. Thus $x \notin G/\Gamma^*$ and δ^* is a C_0 function on G. □

The following theorem shows that if T is u.s.c. on Γ^{**}, and has a weak boundedness property "near" Γ^{**}, then ρ^* is continuous at each point of Γ^{**}.

THEOREM 3.4. *Let T be u.s.c. on* Γ^{**}. *If for each* $x^{**} \in \Gamma^{**}$ *there exist positive constants K and K' such that* $\|x - x^{**}\| \leqq K$ *implies that* $\rho^*(x) \leqq K'$, *then* ρ^* (*defined by* (3.7)) *vanishes and is continuous at each point of* Γ^{**}.

Proof. If $x^{**} \in \Gamma^{**}$, then by definition $\rho^*(x^{**}) = 0$. Let $\{x_i\} \subseteq G$ with $x_i \to x^{**}$. For i sufficiently large, $\rho^*(x_i) \leqq K'$, so choose an index set I and a sequence $\{y_i\}$ such that $y_i \in T(x_i)$ for all i, $y_i \xrightarrow{I} y$, and $\|y_i - x_i\| \xrightarrow{I} \limsup_{i \to \infty} \rho^*(x_i)$. But by the u.s.c. of T on Γ^{**}, $y \in T(x^{**})$ so $y = x^{**}$ and $\limsup_{i \to \infty} \rho^*(x_i) = 0$. Since $\limsup \rho^*(x_i) \geqq \liminf \rho^*(x_i) \geqq 0$, we have $\limsup \rho^*(x_i) = \liminf \rho^*(x_i) = 0$ and thus ρ^* is continuous at x^{**}. □

The following corollary summarizes the results of the previous two theorems:

COROLLARY 3.5. *Let T satisfy the hypotheses of Theorems* 3.3 *and* 3.4. *Then there exist functions δ, ρ, and μ such that $y \in T(x)$ implies $\phi(x) - \phi(y) \geqq \delta(x)$ and $\rho(x) \geqq \mu(x - y)$, where δ and ρ are nonnegative on G, δ is C_0 on G, ρ vanishes and is continuous at each point of Γ^{**}, and μ is C_0 and positive-definite on $\mathbb{R}^n$.*

Proof. Let $\delta \equiv \delta^*$, $\rho \equiv \rho^*$, and $\mu(x - y) \equiv \|x - y\|$, and apply Theorems 3.3 and 3.4. □

It should be noted that the defining relation (3.6) provides the supremum over all functions δ for which the relation (2.4) would be satisfied for arbitrary choices of the successor point x_{i+1}, and, in this sense, is the best choice for a δ to be used in applying the theorems of § 2 to an algorithm corresponding to a point-to-set mapping. In some applications, however, it may be more convenient to derive an *estimate* for a lower bound on $\phi(x_i) - \phi(x_{i+1})$, and a δ developed by such an estimation procedure will yield the same results as δ^* provided that δ is null continuous and has the same set of zeros as δ^*.

4. Convergence theorems for point-to-set mappings. By using the convergence results of § 2 and Theorems 3.3 and 3.4, it is possible to develop convergence theorems for algorithms based on point-to-set mappings. However, rather than constructing proofs for such theorems by applying previous theorems, it turns out to be somewhat simpler and more illuminating to give direct proofs.

The first result of this type is the analogue of Lemma 2.2 suggested by Theorem 3.3.

THEOREM 4.1. *Let T be a point-to-set mapping from G into the* nonempty *subsets of G, and let T be* monotonic *on G w.r.t. some l.s.c. function ϕ, and* sequentially monotonic *w.r.t. ϕ on G/Γ^* (where Γ^* is the set of GFP's of T). If a sequence generated by the algorithm corresponding to T has an accumulation point x^*, then $x^* \in \Gamma^*$.*

Proof. We will suppose that $x^* \notin \Gamma^*$, and show a contradiction. Since ϕ is assumed l.s.c., $\phi^* \equiv \lim \phi(x_i) \geqq \phi(x^*)$. Let I be such that $x_i \xrightarrow{I} x^*$. Then $\phi(x_i) \xrightarrow{I} \phi^*$, and $\phi(x_{i+1}) \xrightarrow{I} \phi^*$, contradicting the sequential monotonicity property at x^*. □

A similar analogue of Theorem 2.4 may be stated, but for the sake of brevity we will state and prove the analogue of Corollary 2.5 suggested by Theorem 3.5.

THEOREM 4.2. *Let T satisfy the hypotheses of Theorem* 4.1, *and let $\{x_i\}$ be generated by the algorithm corresponding to T. If* (a) *$\{x_i\}$ is bounded,* (b) *the set of SFP's of T coincides with Γ^* and does* not *contain a continuum, and* (c) *T is u.s.c. at each SFP, then $x_i \to x^*$, where x^* is an SFP of T.*

Proof. Let x^* be an accumulation point of $\{x_i\}$. By the previous theorem $x^* \in \Gamma^*$ so x^* is an SFP of T by hypothesis (b). We will assume $\|x_{i+1} - x_i\| \not\to 0$ and show a contradiction. If $\|x_{i+1} - x_i\| \not\to 0$, there exists an I and a $\delta > 0$ such that $\|x_{i+1} - x_i\| \geqq \delta$ for $i \in I$ and $x_i \xrightarrow{I} x'$, $x_{i+1} \xrightarrow{I} x''$. Since x' must be an SFP, $x'' = x'$ and thus $\|x_{i+1} - x_i\| \xrightarrow{I} 0$, a contradiction. Since $\{x_i\}$ is bounded and $\|x_{i+1} - x_i\| \to 0$, if $\{x_i\}$ did not converge, its accumulation points would form a continuum contained in Γ^*, which is impossible. □

The crucial role of the SFP's in the preceding theorem is illustrated by the following result, which shows that, if Γ^* is finite but contains no SFP's, it is always possible to generate a *divergent* sequence by using the algorithm corresponding to T.

THEOREM 4.3. *Let the hypotheses of Theorem* 4.1 *hold. If Γ^* is a finite set containing* no *SFP's then there exists an x_0 and a corresponding sequence $\{x_i\}$ generated by the algorithm corresponding to T such that $\{x_i\}$ does* not *converge.*

Proof. Let $\Phi \equiv \{x' | x' \in \Gamma^*, \phi(x') \leqq \phi(x)$ for all $x \in \Gamma^*\}$ and let $x_0 \in \Phi$. Given x_i, choose $x_{i+1} \neq x_i$. (This is always possible, for otherwise x_i would be an SFP.) We will suppose that $x_i \to x^*$, and show a contradiction. By a preceding theorem, $x^* \in \Gamma^*$, and since $\phi(x^*) \leqq \phi(x_0)$, we have $x^* \in \Phi$ and thus $\phi(x^*) = \phi(x_i)$ for all i. Thus $\delta(x_i) = 0$ for all i and $x_i \in \Phi$ for all i. Since Φ is a finite set, the relations $\{x_i\} \subseteq \Phi$ and $x_i \to x^*$ imply $x_i = x^*$ for all i sufficiently large, contradicting the fact that $x_{i+1} \neq x_i$. □

5. Methods involving anti-jamming parameters. We now wish to extend the convergence analysis approach developed in previous sections to algorithms for which the optimality indicator depends not only on x but also on a scalar

parameter ε. This extension allows the analysis of constrained optimization methods employing a so-called "anti-jamming" parameter. The conditions to be given below were previously stated in [12], but here they will be presented as a natural extension of the results in §§ 2–4.

The simplest approach to an appropriate extension of the previously developed theory is to replace the relation $\phi(x_i)-\phi(x_{i+1})\geqq\delta(x_i)$ by

$$\phi(x_i)-\phi(x_{i+1})\geqq\hat{\delta}(\varepsilon_i, x_i) \tag{5.1}$$

in order to reflect the dependence of the change in ϕ on the ith value of the parameter, ε_i. Note that by defining the composite variable $z=(\varepsilon, x)$ and the function $\hat{\phi}(z)\equiv\phi(x)$, the relation (5.1) can be written so that the same variable appears on both sides, i.e., in the form $\hat{\phi}(z_i)-\hat{\phi}(z_{i+1})\geqq\hat{\delta}(z_i)$. If $\hat{\delta}$ is nonnegative on $R^1\times G$ and ϕ is l.s.c. on G (so that $\hat{\phi}$ will also be l.s.c. on $R^1\times G$), then Lemma 2.1 may be applied to establish properties of the accumulation points of the sequence $\{z_i\}$. (More generally, any set of relations of the form $f(u_i)-f(u_{i+1})\geqq g(v_i)$, where the u_i are in some space U and the v_i are in some space V may be converted in an obvious fashion to a new set of relations in which variables from $U\times V$ appear on both sides and the functions involved have the same continuity and nonnegativity properties as f and g, so that the results of § 2 may be applied.) As in § 2, we would then like to go a bit further and exploit additional properties of $\hat{\delta}$ in order to sharpen the characterization of the accumulation points obtained from Lemma 2.1. Unfortunately, while an analogue of Lemma 2.2 may be established if $\hat{\delta}$ is C_0 on $R^1\times G$, for many well-known feasible direction methods the corresponding function $\hat{\delta}$ is *not* null-continuous on $R^1_+\times G$ (for a simple example of this phenomenon see p. 24 of [10].)

Thus the properties of $\hat{\delta}$ that we will exploit are *weaker* than null-continuity, but will nonetheless be *strong enough* to guarantee that the accumulation points satisfy an optimality condition. These properties will also, of course, be such that they are satisfied by the well-known feasible direction methods. The appropriate additional properties of $\hat{\delta}$ are as follows:

$$\hat{\delta}(\varepsilon_i, x_i)\geqq\delta_2(\omega_i, x_i)\cdot\min\{\omega_i, \|x_i-x_{i+1}\|\}, \tag{5.2}$$

where

$$\omega_i\equiv\delta_3(\varepsilon_i, x_i), \tag{5.3}$$

and

$$\|x_i-x_{i+1}\|\geqq\delta_1(\min\{\varepsilon_i, \omega_i\}, x_i), \tag{5.4}$$

where δ_1 and δ_2 are nonnegative and have the *generalized forcing function property* (for $j=1, 2$) on $R^1_+\times G$:

$$\delta_j(\eta_i, y_i)\to 0 \quad\text{and}\quad y_i\to\bar{y} \quad\text{imply}\quad \eta_i\to 0 \tag{5.5}$$

and δ_3 is nonnegative and null-continuous on $R^1_+\times G$. By assuming (5.1)–(5.5) and making an assumption on the relationship of $\{\varepsilon_i\}$ to $\{\omega_i\}$, the following Theorem shows that δ_3, which plays the role of an *optimality indicator*, vanishes at the accumulation points of $\{(\varepsilon_i, x_i)\}$.

THEOREM 5.1. *Let ϕ be l.s.c. on G, let* (5.1)–(5.5) *hold, and let δ_3 be null-continuous on $R_+^1 \times G$. Let $\{\varepsilon_i\}$ be such that the existence of a* subsequence *of $\{\varepsilon_i\}$ converging to* 0 *implies* (i) *that $\varepsilon_i \to 0$ and* (ii) *that $\{\omega_i\}$ also contains a subsequence converging to* 0. *If $\{x_i\}$ contains an accumulation point $\bar{x}$, then $\delta_3(0, \bar{x}) = 0$.*

Proof. This result is proved in Theorem 1, p. 7, of [10]. □

It is shown in [12] that this theorem may be applied to the analysis of the feasible direction methods of Zoutendijk [21], Topkis and Veinott [18], and Mangasarian [8], and that it also suggests new and, in some cases, more efficient parameter generation schemes. While it is thus possible to extend the forcing function approach to algorithms with an anti-jamming parameter, it does not appear possible to similarly extend the point-to-set mapping approach in a *natural* way to cover this situation, since the convergence proof depends on the special structure of $\hat{\delta}(\varepsilon_i, x_i)$. The results of this section may also be compared with similar results of Klessig [6]. (Although Klessig's results are stated in a rather different format involving point-to-set mappings, the continuity properties that are the cornerstone of his hypotheses can be formulated as continuity properties of single-valued functions analogous to the δ_i above.) Theorem 5.1 provides (i) a more general lower bound on the decrease $\phi(x_i) - \phi(x_{i+1})$, which permits a wider variety of step-sizes to be used, and (ii) a more general relation between the sequence of values, $\{\varepsilon_i\}$, of the anti-jamming parameter and the sequence of values, $\{\omega_i\}$, of the optimality indicator, so that new, *nonmonotonic* procedures for anti-jamming parameter generation can be handled. (For more details on these points, see Meyer [12].)

It might be noted that Zangwill [20] treats the convergence of the feasible direction methods that he considers by a direct argument rather than by the application of his general point-to-set mapping convergence theorems. (There is also some question as to whether Zangwill's results are correct, since the proof given for his Theorem 13.2 is erroneous and cannot be corrected in a straightforward manner.)

The interested reader may also refer to [12] for extensions of Theorem 5.1 that give sufficient conditions for the convergence of the full sequence $\{x_i\}$ to a point x^* such that $\delta_3(0, x^*) = 0$. These conditions are analogous to those assumed in Corollary 2.5.

6. Conclusions. A general convergence theory for monotonic mathematical programming algorithms has been developed via the forcing function approach. This approach has the pedagogical advantage of avoiding the use of point-to-set mappings, but is nevertheless shown to be equivalent to a development relying on point-to-set mapping properties for two of the three classes of algorithms considered. The forcing function approach has some advantages in terms of providing a framework for the analysis of the third class of algorithms, feasible direction methods. On the other hand, there are classes of algorithms involving contraction mappings [10], cyclic or "restart" policies [11], and linearization procedures [16] for which a point-to-set mapping approach appears quite suitable whereas the forcing function approach would be somewhat unnatural. The point-

to-set mapping approach also offers geometric insights not as easily obtained from the forcing function approach.

Appendix. In order to obtain globally convergent mathematical programming algorithms, it is customary in practice to introduce step-size procedures that guarantee a "sufficient decrease" in some function. In terms of the theory described above, "sufficient decrease" means that the function δ determining a lower bound for the decrease should be null-continuous and that the set of points on which δ vanishes should coincide with the set of points satisfying an appropriate optimality condition. For purposes of comparison with the results of Zangwill and Polak, however, we must allow for the possibility of a "worsening" (or increase) in the value of ϕ, or an empty set of successors, even though for nonlinear minimization algorithms used in practice it is always possible to let $x_{i+1} = x_i$ if computations at the ith iteration have not yielded a point with a smaller objective value.

A comparison with Polak's Theorem 1.3.10 (see [15]). In Polak [15], it is assumed that a set $T^* \subseteq G$ has been designated a priori as the set of desirable points, and the following algorithm and theorem are given:

ALGORITHM (Polak). Let A be a point-to-set mapping from G into the nonempty subsets of G.

Step 0. Compute a $z_0 \in G$.

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 $\phi(z_{i+1}) \geqq \phi(z_i)$ stop; else, set $i = i+1$ and go to Step 2.

THEOREM 1.3.10 (Polak). *Suppose that* (i) ϕ *is either continuous at all nondesirable points or* ϕ *is bounded from below on* G; (ii) *for every* $z \in G$ *which is not desirable, there exist an* $\varepsilon(z) > 0$ *and a* $\bar{\delta}(z) < 0$ *such that* $\phi(z'') - \phi(z') \leqq \bar{\delta}(z) < 0$ *for all* $z' \in G$ *such that* $\|z' - z\| \leqq \varepsilon(z)$ *and for* $z'' \in A(z')$. *Then, either the sequence* $\{z_i\}$ *constructed by the algorithm is finite and its next to last element is desirable, or else it is infinite and every accumulation point of* $\{z_i\}$ *is desirable.*

In order to compare our approach with that of Polak, we first define

$$\delta(x) \equiv \max\{0, \inf\{\phi(x) - \phi(x') | x' \in A(x)\}\}. \tag{A.1}$$

Note that δ is nonnegative on G. We will now show how Lemma 2.1 may be used to obtain a strengthened version of Polak's theorem.

LEMMA A.1. *Let* ϕ *be either l.s.c. or bounded from below on* G, *and let hypothesis* (ii) *of Polak's theorem hold. If the set of* z_i *constructed by Polak's algorithm is finite, then its next-to-last element is in* Ω^*. *If the set of* z_i *has an accumulation point* $\bar{x}$, *then* $\bar{x} \in \Omega^L$, *which is contained in the set of desirable points.*

Proof. In the finite termination case, the conclusion is obvious. In the infinite case, since $\phi(z_i) - \phi(z_{i+1}) \geqq \delta(z_i)$ for all i, Lemma 2.1 applies (recall that the proof requires only that ϕ is bounded from below). To see that Ω^L is a *subset* of the desirable points, suppose that $y_i \to x$ and $\delta(y_i) \to 0$, but that $x \in T'$, the set of *nondesirable points*. Since $\delta(y_i) \to 0$, there exists a sequence $\{y_i'\}$, with $y_i' \in A(y_i)$ such that $\limsup[\phi(y_i) - \phi(y_i')] \leqq 0$, contradicting Polak's hypothesis (ii). □

Note that Lemma A.1 yields a stronger result than Polak's theorem, since Ω^L may be a *proper* subset of the desirable points, as the following example shows:

Example. Let $G \equiv \{1/n | n = 1, 2, \cdots\} \cup \{0\} \cup \{-1\}$, $\phi(x) \equiv x$, $T' \equiv \{1/n | n = 2, 3, \cdots\}$, and

$$A(x) \equiv \begin{cases} \{-1\} & \text{if } x = -1, \text{ or } 0, \\ \{1/(n+1)\} & \text{if } x = 1/n, \quad n = 1, 2, \cdots. \end{cases}$$

Then all of the hypotheses of Polak's theorem are satisfied, and $G/T' = \{-1, 0, 1\}$, $\Omega^L = \{-1, 0\}$, and $\Omega^* = \{-1\}$. Note that although the point 1 has been classed as a "desirable" point, the algorithm can neither terminate at 1 nor converge to 1. In this case, then, Lemma A.1 is a sharper result than Polak's, since it restricts the terminal and accumulation points to smaller sets.

The difference between the sharpness of the two results is essentially a result of the fact that Ω^L and Ω^* are completely determined by A and ϕ via (A.1), whereas Polak's "desirable set" is determined independently of the algorithm. Note, however, that if the points in Ω^L are designated as the desirable points, then hypothesis (ii) of Polak's Theorem 1.3.10 is unnecessary, since the set of nondesirable points then becomes G/Ω^L, which is precisely the set of points for which the algebraic conditions of Polak's hypothesis (ii) hold; in this instance the conclusions of Theorem 1.3.10 and Lemma A.1 are essentially equivalent.

Comparison with Zangwill's Theorem A (see [20]). In Theorem A the algorithm is given a point z_1 and generates the sequence $\{z_k\}$ by use of the recursion $z_{k+1} \in A(z_k)$.

CONVERGENCE THEOREM A (Zangwill). *Let the point-to-set map $A: G \to G$ determine an algorithm that given a point $z_1 \in G$ generates the sequence $\{z_k\}$. Also let a* solution set $\Omega \subset G$ *be given.*

Suppose

1) *All points z_k are in a compact set $X \subset G$.*
2) *There is a continuous function $Z: G \to E^1$ such that:*
   (a) *if z is not a solution, then for any $y \in A(z)$*

$$Z(y) > Z(z);$$

   (b) *if z is a solution, then either the algorithm terminates or for any $y \in A(z)$,*

$$Z(y) \geqq Z(z).$$

and

3) *The map A is closed at z if z is not a solution.*

Then either the algorithm stops at a solution, or the limit of any convergent subsequence is a solution.

The statement of Zangwill's theorem is a bit unclear, since the suggestion of the possibilities that "the algorithm terminates" or that "the algorithm stops" cannot be reconciled with the hypothesis that the algorithm generates an infinite sequence $\{z_i | i = 1, 2, \cdots\}$. On the basis of some of Zangwill's other results we will assume that the statement that "the algorithm terminates" at x_i is equivalent to $A(x_i)$ being empty. Hence, we will again define a δ to take this into account, and

apply Lemma 2.1 to obtain a strengthened result. Let $\phi \equiv -Z$ and let δ be defined as follows on G:

$$\text{(A.2)} \qquad \delta(x) \equiv \begin{cases} \inf\{\phi(x)-\phi(y) | y \in A(x)\}, & \text{if } A(x) \text{ is nonempty,} \\ 0 & \text{if } A(x) \text{ is empty.} \end{cases}$$

(Note that δ is nonnegative on G.)

LEMMA A.2. *Let* 2) *and* 3) *of Theorem* A *hold, and let* $\{z_i | i \in I\}$ *be a set of points generated by Zangwill's algorithm. If* I *is finite, then the last element* $z_i \in \Omega^* \cap \Omega$. *If* I *is infinite and the sequence* $\{z_i\}$ *has an accumulation point* $\bar{x}$, *then* $\bar{x} \in \Omega^L$. *If* I *is infinite and the sequence* $\{z_i\}$ *is bounded, then each accumulation point* $x^* \in \Omega^L \cap \Omega$.

Proof. The proof of the first two conclusions is analogous to the proof of Lemma A.1. If the sequence $\{z_i\}$ is bounded, then by Zangwill's theorem the accumulation points belong to Ω, and by the second conclusion of the lemma they also belong to Ω^L. □

It should be noted that Ω^L may contain points not in Ω and vice-versa, as the following example shows:

Example. Let

$$N \equiv \{n | n = 1, 2, \cdots\},$$

$$G \equiv N \cup \{1/n | n = 1, 2, \cdots\} \cup \{0\} \cup \{-1\},$$

$$A(x) \equiv \begin{cases} -1 & \text{if } x = -1 \text{ or } 0, \\ (1/x)+1 & \text{if } x = 1/n, \quad n = 2, 3, \cdots, \\ 1/(x+1) & \text{if } x = n, \quad n = 1, 2, 3, \cdots, \end{cases}$$

$$\phi(x) \equiv \begin{cases} x & \text{if } x \leqq 1, \\ 1/x & \text{if } x > 1, \end{cases}$$

$$\Omega \equiv \{-1, 1\}.$$

Note that hypotheses 2) and 3) of Zangwill's theorem are satisfied, and that $\Omega^L = \{-1, 0\}$ and $\Omega^* = \{-1\}$. Thus, by Lemma A.2, if the algorithm terminates in a finite number of steps, it must terminate at the point -1; if the algorithm yields an infinite set of iterates, then either of -1 or 0 could be accumulation points; and, if it yields an infinite set of iterates contained in a *bounded* set, then they must converge to -1. Note, in fact, that these results are in this case the best possible, since -1 will be the unique accumulation point if the algorithm starts with $z_1 = -1$ or 0, and 0 will be the unique accumulation point for any other starting point in G. By comparison, Zangwill's Theorem A does *not* apply to the case in which the z_i are unbounded (which occurs unless $z_1 = -1$ or 0), and in the bounded iterate case, Theorem A narrows the candidates for accumulation points only to the set $\{-1, 0\}$. Analogues of the comments made regarding the sharpness of Polak's theorem apply here, since Ω is *not* uniquely determined by A and Z, and thus may be taken to be larger than is really necessary, whereas Ω^* and Ω^L are uniquely determined by A and Z. Of course, in applying Theorem A one would like to choose the set Ω as small as possible, i.e., as the intersection of all Ω for which hypotheses 2) and 3) were satisfied. However, this choice has the disadvantage of

providing a rather complex definition of a "solution set" Ω (in comparison with the definition of Ω^*), and the resulting Ω need not contain all the possible accumulation points in the unbounded iterate case.

Other forcing function approaches. In the approach to convergence analysis used by Ortega and Rheinboldt [13] (a similar approach is also used by Daniel [2]), two main hypotheses are made regarding the decrease in ϕ at each iteration:

$$\phi(x_i) - \phi(x_{i+1}) \geqq \sigma_1(|\nabla\phi(x_i)p_i|/\|p_i\|),$$

and

$$|\nabla\phi(x_i)p_i|/\|p_i\| \geqq \sigma_2(\|\nabla\phi(x_i)\|),$$

where p_i is a search direction and the σ_j have the property that, for $j = 1$ or 2, $\lim_{k\to\infty} \sigma_j(t_k) = 0$ implies $\lim_{k\to\infty} t_k = 0$ for any nonnegative sequence $\{t_k\}$. (By considering the above two inequalities, the step-size and the direction-generation techniques of an algorithm may be analyzed separately, so that the potential independence of those two techniques is emphasized.) If we assume *in addition* that σ_1 is monotone nondecreasing (a hypothesis that is satisfied by all the algorithms considered in Ortega and Rheinboldt [13], where, in fact, in most cases $\sigma_1(t) = Mt^2$ for some $M > 0$), then we have

$$\phi(x_i) - \phi(x_{i+1}) \geqq \sigma_1(\sigma_2(\|\nabla\phi(x_i)\|)).$$

Letting $\delta(x_i) \equiv \sigma_1(\sigma_2(\|\nabla\phi(x_i)\|))$ and assuming that ϕ is continuously differentiable and $\sigma_1(0) = \sigma_2(0) = 0$, we may conclude that δ is C_0 and that the set of points on which δ vanishes is the set of points on which $\nabla\phi$ vanishes. Thus, under these hypotheses Lemma 2.2 may be applied, and we may conclude that $\nabla\phi$ will vanish at the accumulation points of $\{x_i\}$. Finally Ortega and Rheinboldt also establish, for each algorithm they consider, conditions analogous to (2.5) to guarantee $\|x_i - x_{i+1}\| \to 0$ (these are generally of the form $\bar{M}|\nabla\phi(x_i)p_i|/\|p_i\| \geqq \|x_i - x_{i+1}\|$, where $\bar{M} > 0$), or prove the relation $\|x_i - x_{i+1}\| \to 0$ by utilizing properties of the step-size techniques together with a hypothesis (hemivariateness) on ϕ.

REFERENCES

[1] S. M. CHUNG, *Globally and superlinearly convergent algorithms for nonlinear programming*, Ph.D. Thesis, Computer Sciences Dept., Univ. of Wisconsin, Madison, 1975.

[2] J. W. DANIEL, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[3] B. C. EAVES AND W. I. ZANGWILL, *Generalized cutting plane algorithms*, this Journal, 9 (1971), pp. 529–542.

[4] W. W. HOGAN, *Point-to-set maps in mathematical programming*, SIAM Rev., 15 (1973), pp. 591–603.

[5] P. HUARD, *Optimization algorithms and point-to-set maps*, Math. Prog., 8 (1975), pp. 308–331.

[6] R. KLESSIG, *A general theory of convergence for constrained optimization Algorithms that use antizigzagging provisions*, this Journal, 12 (1974), 598–608.

[7] DAVID G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA., 1973.

[8] O. L. MANGASARIAN, *Dual feasible direction methods*, Techniques of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972, pp. 67–88.

[9] R. R. MEYER, *The validity of a family of optimization methods*, this Journal, 8 (1970), pp. 41–54.

[10] ———, *Sufficient conditions for the convergence of monotonic mathematical programming algorithms*, J. Comput. System Sci., 12 (1976), pp. 108–121.
[11] ———, *On the convergence of algorithms with restart*, Univ. of Wisconsin Computer Sciences Tech. Rep. 225, Madison, 1974; SIAM J. Numer. Anal. 13 (1976), pp. 696–704.
[12] ———, *A convergence theory for a class of anti-jamming strategies*, Univ. of Wisconsin Mathematics Research Center Rep. 1481, Madison, 1975; J. Optimization Theory Appl., to appear.
[13] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[14] A. M. OSTROWSKI, *Solution of Equations and Systems of Equations*, Academic Press, New York, 1966.
[15] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York, 1971.
[16] B. T. POLYAK, *Gradient methods for the minimization of functionals*, U.S.S.R. Computational Math. and Math. Phys., 3 (1963), pp. 864–878.
[17] S. M. ROBINSON AND R. R. MEYER, *Lower semicontinuity of multivalued linearization mappings*, this Journal, 11 (1973), pp. 525–533.
[18] D. M. TOPKIS AND A. F. VEINOTT, *On the convergence of some feasible direction algorithms for nonlinear programming*, this Journal, 5 (1967), pp. 268–279.
[19] PHILIP WOLFE, *On the convergence of gradient methods under constraint*, IBM Res. Paper RC-1752, Yorktown Heights, NY, 1967.
[20] W. I. ZANGWILL, *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
[21] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

OPTIMAL PERIODIC CONTROL: A GENERAL THEORY OF NECESSARY CONDITIONS*

ELMER G. GILBERT†

Abstract. Does time-dependent periodic control yield better process performance than optimal steady-state control? This paper examines exhaustively the role of first order necessary conditions in answering this question. For processes described by autonomous, ordinary differential equations, a very general optimal periodic control problem (OPC) is formulated. By considering control and state functions which are constant, a finite-dimensional optimal steady-state problem (OSS) is obtained from OPC. Three solution sets are introduced: $\mathcal{S}$(OSS)—the solutions of OSS, $\mathcal{S}$(OPC)—the solutions of OPC, $\mathcal{S}$(SSOPC)—the solutions of OPC which are constant. Necessary conditions for elements of each of these sets are derived; their solution sets are denoted, respectively, by $\mathcal{S}$(NCOSS), $\mathcal{S}$(NCOPC), and $\mathcal{S}$(NCSSOPC). The relationship between these six solutions sets is a central issue. Under various hypotheses certain pair-wise inclusions of the six sets are determined and it is shown that no others can be obtained. Tests which imply that time-dependent periodic control is better than optimal steady-state control ($(\mathcal{S}(\text{SSOPC}) = \varnothing,\ \mathcal{S}(\text{OSS}) \neq \varnothing)$, including those based on relaxed steady-state control, are investigated and limits to what tests exist are established. The results integrate and amplify results which have appeared in the literature. Examples provide insight which supports the theory.

1. Introduction. Since the 1967 paper by Horn and Lin [13] there has been an increasing interest in the mathematical theory of periodic processes. The motivations for this theory came initially from the optimization of chemical processes [3], but there are other areas of potential application such as vehicle cruise [10]. The essence of most applications is the optimization of a "continuing process," a process which is fixed in its characteristics and is expected to operate continuously over an indefinitely long period of time. The traditional approach to such problems is to minimize process cost by selecting constant controls subject to the constraint that the (dynamic) process is in static equilibrium. Although this "steady-state" approach is simple (time does not appear) and has intuitive appeal, it is not necessarily best. It may be possible to exploit the process dynamics and obtain even lower cost. Experiments with actual processes have shown that this can indeed be the case. The theory has helped to explain some of the mechanisms for such improvement and suggests situations where "time-dependent" control may improve performance. Much of the literature on periodic control has been reviewed by Bailey [3] and Guardabassi, Locatelli and Rinaldi [11].

The natural starting point for a theoretical investigation of continuing processes is the formulation of a dynamic optimization problem. It is clear from the preceding discussion that this optimal control problem should satisfy certain requirements: 1. the system dynamics and control constraints should not depend explicitly on time, 2. the system state and control functions should be defined on the time interval $(-\infty, +\infty)$, 3. a meaningful "optimal steady-state" problem,

* Received by the editors December 8, 1975, and in revised form August 30, 1976.

† Department of Aerospace Engineering, Program in Computer, Information and Control Engineering, University of Michigan, Ann Arbor, Michigan 48109. This research was completed while the author was on leave at the Department of Electrical Engineering, Johns Hopkins University, Baltimore, Maryland. It was supported in part by the United States Air Force, Air Force Office of Scientific Research, Air Force Systems Command, under Grants 73-2517 and 77-3158.

which does not involve time, should result when the system state and control functions are assumed to be constant. This is the attitude taken in this paper; everything is based on the optimal control problem (OPC) which is stated in § 2. The structure of this problem is chosen so that requirements 1 and 3 are met directly. Requirement 2 is imposed indirectly by assuming that the system state and control functions are periodic. Although this is not absolutely essential it is consistent with the previous literature, is a practical constraint, and avoids certain mathematical difficulties. The problem OPC, which assumes the system dynamics are represented by ordinary differential equations, is quite general and includes most of the problems which have appeared to date as special cases.

Because of the special form of OPC there are three notions of optimality (solutions of OPC, solutions of OPC which are constant, solutions of the steady-state problem) and, correspondingly, three sets of necessary conditions. Hence many potential relationships exist between the necessary conditions and the various optima. The investigation of these relationships is the central theme of this paper. Apart from its intrinsic interest this investigation is valuable for a number of other reasons: it puts together in a larger, more consistent framework many of the scattered results in the literature; it produces stronger tests for optimality and properness (time-dependent control better than optimal-steady-state control); it establishes certain limits to what can be proved concerning these tests; it sheds new light on the role of relaxed steady-state controls.

The paper is organized as follows. Section 2 states the problem OPC and introduces notation for the three sets of solutions. In § 3 the necessary conditions are derived. The developments are restricted to the "first variation" and are, for the most part, applications of well established theory. Section 4 introduces notation for the sets of solutions of the necessary conditions and relates these sets to the three sets of optima. Section 5 presents a number of examples which show that it is not possible to obtain more set inclusions than those obtained in § 4. Tests for properness are considered in § 6 and it is shown that under certain reasonable conditions no other tests exist. Tests for optimality and relative optima are also discussed. Section 7 treats relaxed steady-state optima; one of the main consequences is an extension of the well known results of Bailey and Horn [1].

It is worth noting that the concept of a continuing process seems essential to much of what follows. While it is possible to pose optimal periodic control problems which do not satisfy requirements 1 and 3, the results concerning the comparison of time-dependent and steady-state optima are greatly weakened.

2. Formulation of the problem. In this section a problem of optimal periodic control is formulated which meets the general requirements of the previous section. It models a wide class of continuing processes and subsumes a meaningful steady-state problem. Solution sets related to the two optimization problems are defined and some simple facts concerning them are noted.

Before stating the optimal periodic control problem it is necessary to introduce the following notation and assumptions: j and k are nonnegative integers, $T \in R$ is positive, $U \subset R^m$ is an arbitrary set, $X \subset R^n$ and $Y \subset R^l$ are open sets, for $i = -j, \cdots, k$ the functions g_i: $Y \times X \to R$ are continuously differen-

tiable, the functions $f: X \times U \to R^n$ and $\tilde{f}: X \times U \to R^l$ are continuous and for each $u \in U$ are continuously differentiable in x.

Optimal periodic control problem (OPC). Find $u(\cdot)$, $x(\cdot)$ and τ which minimize J subject to

$$J = g_0(y, x(0)), \tag{2.1-1}$$

$$g_i(y, x(0)) \leqq 0, \qquad i = -j, \cdots, -1, \tag{2.1-2}$$

$$g_i(y, x(0)) = 0, \qquad i = 1, \cdots, k, \tag{2.1-3}$$

$$y = \frac{1}{\tau} \int_0^\tau \tilde{f}(x(t), u(t))\, dt \in Y, \tag{2.1-4}$$

$$\dot{x}(t) = f(x(t), u(t)) \quad \text{almost all } t \in [0, T], \quad x(0) = x(\tau), \tag{2.1-5}$$

$$u(\cdot) \in \mathcal{U} = \{u(\cdot): u(\cdot) \text{ measurable and essentially bounded on } [0, T],\ u(t) \in U \text{ for all } t \in [0, T]\}, \tag{2.1-6}$$

$$x(\cdot) \in \mathcal{X} = \{x(\cdot): x(\cdot) \text{ absolutely continuous on } [0, T],\ x(t) \in X \text{ for all } t \in [0, T]\}, \tag{2.1-7}$$

$$\tau \in (0, T]. \tag{2.1-8}$$

Some general comments are in order. Equations (2.1-5) represent the dynamics of the process and the constraints that $x(\cdot)$ and $u(\cdot)$ are periodic on $(-\infty, +\infty)$ when appropriate extensions of their definitions are made: $x(t + \nu\tau) = x(t)$, $u(t + \nu\tau) = u(t)$, $t \in [0, \tau)$, $\nu =$ integer. The components of $\tilde{f}(x(t), u/t))$ are quantities of interest in the optimization problem, e.g., rates of process fuel consumption, material flow rates, overhead cost rates, value measures of process products. It is the *average* of these quantities y, as given by (2.1-4), which appear in the actual optimization of the process, i.e., the minimization of (2.1-1) subject to (2.1-2) and (2.1-3). The dependence of the g_i on $x(0)$ allows consideration of factors relating to the "start-up" of each cycle of operation. It also allows constraints to be imposed on $x(0) = x(\tau)$. Note that f and $\tilde{f}$ and the control constraint set U do not depend on t and the g_i do not depend on τ. This is essential if the requirements 1 and 3 of § 1 are to be satisfied. The bound (2.1-8) is consistent with the assumption of periodic operation. While $T = +\infty$ is not allowed, arbitrarily large T is permitted. Thus the quasi-stationary approximation treated in the literature [3], [11] can be extended to OPC. This is not done here. The convention $j = 0$ is used to denote the absence of inequality constraints; similarly $k = 0$ denotes absence of equality constraints.

By appropriate changes in notation problem formulations considered previously in the literature become special cases of OPC. For example, the problem of Guardabassi, Locatelli and Rinaldi [11] requires $j = 0$ and g_i, $i = 0, \cdots, k$ equal to the components of y; the problem of Bailey and Horn [1] requires $j = k = 0$ and g_0 equal to a general function of y. The problem in [1] is somewhat more general than it may first appear because a simple substitution of variables allows it to include the case $j = 0$, $k > 0$ when the functions g_i, $i > 0$, are components of y [2]. However, when restricted to the context of continuing systems, none of the previous formulations have the full generality of OPC.

The steady-state problem is obtained from OPC by adding the constraint that $x(\cdot)$ and $u(\cdot)$ are constant. As expected, this yields a finite-dimensional optimization problem which does not depend on τ.

Optimal steady-state problem (OSS). Find u and x which minimize J subject to

$$J = g_0(y, x), \tag{2.2-1}$$

$$g_i(y, x) \leqq 0, \qquad i = -j, \cdots, -1, \tag{2.2-2}$$

$$g_i(y, x) = 0, \qquad i = 1, \cdots, k, \tag{2.2-3}$$

$$y = \tilde{f}(x, u) \in Y, \tag{2.2-4}$$

$$f(x, u) = 0, \tag{2.2-5}$$

$$u \in U, \tag{2.2-6}$$

$$x \in X. \tag{2.2-7}$$

It is of interest to compare the solutions of OPC with the solutions of OSS. This can be done conveniently by introducing the following solution sets, all of which are subsets of $\mathcal{U} \times \mathcal{X} \times (0, T]$:

$$\mathcal{S}(\text{OPC}) = \{(u(\cdot), x(\cdot), \tau)\text{: } (u(\cdot), x(\cdot), \tau \text{ solves OPC}\}, \tag{2.3}$$

$$\mathcal{S}(\text{SS}) = \{(u(\cdot), x(\cdot), \tau)\text{: (2.1-2)–(2,1-8) are satisfied and } u(\cdot) \text{ and } x(\cdot) \text{ are constant}\}, \tag{2.4}$$

$$\mathcal{S}(\text{SSOPC}) = \mathcal{S}(\text{OPC}) \cap \mathcal{S}(\text{SS}), \tag{2.5}$$

$$\mathcal{S}(\text{OSS}) = \{(u(\cdot), x(\cdot), \tau)\text{: } (u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{SS}) \text{ and } (u(0), x(0)) \text{ solves OSS}\}. \tag{2.6}$$

Of course, $\mathcal{S} = \varnothing$, the null set, is possible in any of the four cases. The particular circumstance $\mathcal{S}(\text{SSOPC}) = \varnothing$, $\mathcal{S}(\text{OSS}) \neq \varnothing$ implies that there exist time-dependent controls which do better than the best steady-state controls. If $\mathcal{S}(\text{SSOPC}) \neq \varnothing$ any $\psi \in \mathcal{S}(\text{SSOPC})$ is also in $\mathcal{S}(\text{OSS})$ since ψ is optimum with respect to choices in $\mathcal{U} \times \mathcal{X} \times (0, T]$ and $\mathcal{S}(\text{SS}) \subset \mathcal{U} \times \mathcal{X} \times (0, T]$. Also, it is clear that all elements of $\mathcal{S}(\text{OSS})$ and $\mathcal{S}(\text{SSOPC})$ yield identical costs J. This leads to the following.

Remark 2.1. There are three mutually exclusive possibilities:

(i) $\mathcal{S}(\text{SSOPC}) = \mathcal{S}(\text{OSS}) \neq \varnothing$;

(ii) $\mathcal{S}(\text{SSOPC}) = \varnothing$, $\mathcal{S}(\text{OSS}) \neq \varnothing$;

(iii) $\mathcal{S}(\text{SSOPC}) = \mathcal{S}(\text{OSS}) = \varnothing$.

Possibility (iii) is not apt to occur since for well posed problems it is likely that $\mathcal{S}(\text{OSS}) \neq \varnothing$. Possibility (i) implies that OPC has a steady-state solution and consequently, there is no advantage (even though OPC may also have time-dependent solutions) in using time-dependent control. Possibility (ii) implies time-dependent control can do better than steady-state control (a statement which holds true even if $\mathcal{S}(\text{OPC}) = \varnothing$). Because of the importance of possibilities (i) and (ii) the following definitions are introduced.

DEFINITION 2.1. If $\mathscr{S}(\text{SSOPC}) = \mathscr{S}(\text{OSS}) \neq \varnothing$ the problem OPC is called *steady-state.*

DEFINITION 2.2. If $\mathscr{S}(\text{SSOPC}) = \varnothing$, $\mathscr{S}(\text{OSS}) \neq \varnothing$ the problem OPC is called *proper* (compare [5]).

The study of relative minima of OPC and OSS will prove to be of value, particularly in the case of steady-state minima.

DEFINITION 2.3. $(u(\cdot), x(\cdot), \tau) \in \mathscr{S}(\text{SS})$ is a *strong* {*weak*} *relative minimum* of OPC if there exists an $\varepsilon > 0$ such that for all $(\hat{u}(\cdot), \hat{x}(\cdot), \hat{\tau})$ which satisfy (2.1-2)–(2.1-8) and $\|\hat{x}(t) - x(0)\| < \varepsilon$ $\{\|\hat{x}(t) - x(0)\| < \varepsilon,\ \|\hat{u}(t) - u(0)\| < \varepsilon\}$, $t \in [0, T]$, it follows that $g_0(y, x(0)) \leqq g_0(\hat{y}, \hat{x}(0))$.

DEFINITION 2.4. $(u(\cdot), x(\cdot), \tau) \in \mathscr{S}(\text{SS})$ is a *strong* {*weak*} *relative minimum* of OSS if there exists an $\varepsilon > 0$ such that for all $(\hat{u}, \hat{x})$ which satisfy (2.2-2)–(2.2-7) and $\|\hat{x} - x(0)\| < \varepsilon$ $\{\|\hat{x} - x(0)\| < \varepsilon,\ \|\hat{u} - u(0)\| < \varepsilon\}$ it follows that $g_0(y, x(0)) \leqq g_0(\hat{y}, \hat{x})$.

In these definitions $\|\cdot\|$ denotes any norm on R^n or R^l and $\hat{y} = y$ for $u = \hat{u}$, $x = \hat{x}$, $\tau = \hat{\tau}$. Corresponding to each of the four types of relative minima, notations for the set of minima are adopted:

$$\mathscr{S}(\text{SRMSSOPC}), \mathscr{S}(\text{WRMSSOPC}), \mathscr{S}(\text{SRMOSS}), \mathscr{S}(\text{WRMOSS}).$$

For example,

$$\mathscr{S}(\text{SRMSSOPC}) = \{(u(\cdot), x(\cdot), \tau) : (u(\cdot), x(\cdot), \tau) \in \mathscr{S}(\text{SS}) \text{ is a strong relative minimum of OPC}\}. \tag{2.7}$$

Obviously, $\mathscr{S}(\text{SSOPC}) \subset \mathscr{S}(\text{SRMSSOPC}) \subset \mathscr{S}(\text{WRMSSOPC})$ and $\mathscr{S}(\text{OSS}) \subset \mathscr{S}(\text{SRMOSS}) \subset \mathscr{S}(\text{WRMOSS})$. By using the same reasoning which led to Remark 2.1 it is easy to see that $\mathscr{S}(\text{SRMSSOPC}) \subset \mathscr{S}(\text{SRMOSS})$. However, $\mathscr{S}(\text{SRMSSOPC}) \neq \varnothing$ does not imply $\mathscr{S}(\text{SRMSSOPC}) = \mathscr{S}(\text{SRMOSS})$ because elements of $\mathscr{S}(\text{SRMSSOPC})$ do not necessarily have the same cost as elements of $\mathscr{S}(\text{SRMOSS})$. Similar reasoning applies to the case of weak relative minima. All of this is summarized in

Remark 2.2. The following conclusions are valid: $\mathscr{S}(\text{SSOPC}) \subset \mathscr{S}(\text{SRMSSOPC}) \subset \mathscr{S}(\text{WRMSSOPC})$, $\mathscr{S}(\text{OSS}) \subset \mathscr{S}(\text{SRMOSS}) \subset \mathscr{S}(\text{WRMOSS})$, $\mathscr{S}(\text{SSOPC}) \subset \mathscr{S}(\text{OSS})$, $\mathscr{S}(\text{SRMSSOPC}) \subset \mathscr{S}(\text{SRMOSS})$, $\mathscr{S}(\text{WRMSSOPC}) \subset \mathscr{S}(\text{WRMOSS})$.

3. The necessary conditions. Since explicit characterization of $\mathscr{S}(\text{OPC})$, $\mathscr{S}(\text{SSOPC})$ and $\mathscr{S}(\text{OSS})$ is generally difficult or impossible, it is essential to consider necessary conditions for the elements of these sets. The necessary conditions for OPC will be obtained by applying some necessary conditions obtained by Neustadt (summarized in Appendix A). Similarly, known conditions for finite-dimensional optimization problems (summarized in Appendix B) are applied to OSS. An entirely separate derivation starting from the necessary conditions for OPC is required to obtain necessary conditions for elements of $\mathscr{S}(\text{SSOPC})$. Relationships between the various necessary conditions and the solution sets introduced in the previous section are examined in § 4.

In what follows: let $f_x(x, u)$ and $\tilde{f}_x(x, u)$ denote respectively the Jacobian matrices of $f(x, u)$ and $\tilde{f}(x, u)$ with respect to x; for $i = -j, \cdots, k$, let $g_{iy}(y, x)$ and

$g_{ix}(y, x)$ denote respectively the Jacobian (row) matrices of $g_i(y, x)$ with respect to y and x; let a prime denote the transpose of a (column) vector or matrix.

THEOREM 3.1 (necessary conditions for OPC). *Let*

$$H(x, u, p, \tilde{p}) = p'f(x, u) + \tilde{p}'\tilde{f}(x, u) \tag{3.1}$$

where $p \in R^n$ *and* $\tilde{p} \in R^l$. *Let* $(u(\cdot), x(\cdot), \tau)$ *solve OPC. Then there exist an absolutely continuous function* $p(\cdot): [0, \tau] \to R^n$, $\tilde{p} \in R^l$ *and real numbers* $\alpha_{-j}, \cdots, \alpha_k$ *such that the following conditions are satisfied*:

$$\max_{v \in U} H(x(t), v, p(t), \tilde{p}) = H(x(t), u(t), p(t), \tilde{p}) \quad \text{almost all } t \in [0, \tau], \tag{3.2-1}$$

$$\tilde{p}' = \sum_{i=-j}^{k} \alpha_i g_{iy}(y, x(0)), \tag{3.2-2}$$

$$\dot{p}'(t) = -p'(t)f_x(x(t), u(t)) - \tilde{p}'\tilde{f}_x(x(t), u(t)) \quad \text{almost all } t \in [0, \tau], \tag{3.2-3}$$

$$\begin{aligned} &p'(\tau) - p'(0) = \tau \sum_{i=-j}^{k} \alpha_i g_{ix}(y, x(0)), \\ &\alpha_i \leqq 0, \qquad i = -j, \cdots, 0, \\ &\alpha_i g_i(y, x(0)) = 0, \qquad i = -j, \cdots, -1, \\ &(\alpha_{-j}, \cdots, \alpha_k, p'(\tau)) \neq 0. \end{aligned} \tag{3.2-4}$$

If $f(x(\cdot), u(\cdot))$ *and* $\tilde{f}(x(\cdot), u(\cdot))$ *are continuous at* τ *the following additional condition is satisfied*:

$$\begin{aligned} &\tilde{p}'y \leqq H_M \quad \text{if } \tau = T, \\ &\tilde{p}'y = H_M \quad \text{if } \tau < T, \end{aligned} \tag{3.2-5}$$

where

$$H_M = \max_{v \in U} H(x(\tau), v, p(\tau), \tilde{p}). \tag{3.2-6}$$

Proof. With the following substitution OPC can be written as GOC of Appendix A: $\hat{n} = n + l$, $\mu = j + 1$, $\nu = k + n$, $\hat{X} = X \times R^l$, $\hat{x} = (x, \tilde{x})$, $\hat{f}(\hat{x}, \mu) = (f(x, u), \tilde{f}(x, u))$; for $i = -j, \cdots, k$, $\theta_i(\hat{x}^1, \hat{x}^2, \tau) = g_i(\tau^{-1}(\tilde{x}^2 - \tilde{x}^1), x^1)$; for $i = k+1, \cdots, k+n$, $\theta_i(\hat{x}^1, \hat{x}^2, \tau) = x^2_{i-k} - x^1_{i-k}$ where the subscripts denote the components of x^2 and x^1; $\theta_{-j-1}(\hat{x}^1, \hat{x}^2, \tau) = \tau - T$; $\hat{t}$ is any real number greater than T. By choosing $\hat{X}^1$ and $\hat{X}^2$ to be appropriate neighborhoods of $\hat{x}^1(0)$ and $\hat{x}^2(\tau)$ the constraint $y \in Y$ is assured. Using the conditions from Theorem A.1, letting $\hat{p} = (p, \tilde{p})$, and replacing α_i by $\tau\alpha_i$ gives conditions (3.2). To confirm the last line of (3.2-4), note that the last condition of (A.3-4) can be written $(\tilde{p}'y - H_M, \alpha_{-j}, \cdots, \alpha_k, p'(\tau)) \neq 0$. Since $(\alpha_{-j}, \cdots, \alpha_k, p'(\tau)) = 0$ implies $\tilde{p}'y - H_M = 0$, the last line of (3.2-4) must follow.

Before stating the necessary conditions for OSS it is necessary to introduce a procedure for obtaining "perturbations" in the constraint set U. This can be done

in a variety of ways (see, e.g. [7], [17], [19]) without being very specific about the characterization of U. Here the presentation follows Canon, Cullum and Polak [7]. Let co V = convex hull of V and cl V = closure of V.

DEFINITION 3.1. A convex cone $C(u, U) \subset R^m$, $u \in U$, is a *conical approximation* to U at u if for any collection $\{\delta u_1, \cdots, \delta u_s\}$ of vectors in $C(u, U)$ there exist an $\varepsilon > 0$ and a continuous function ζ: co $\{u, u+\varepsilon\delta u_1, \cdots, u+\varepsilon\delta u_s\} \to U$, both dependent on $\{\delta u_1, \cdots, \delta u_s\}$, such that $\zeta(u+\delta u) = u+\delta u+o(\delta u)$ where $\|o(\delta u)\| \cdot \|\delta u\|^{-1} \to 0$ as $\delta u \to 0$.

When U has simple characterizations so does $C(u, U)$. For example, suppose

$$U = \{u: h_i(u) \leqq 0,\ i = 1, \cdots, q\}, \tag{3.3}$$

where the h_i are continuously differentiable on R^m with Jacobian matrices $h_{iu}(u)$. Let $I(u) = \{i: h_i(u) = 0\}$. Then

$$\text{cl } C(u, U) = \{\delta u: h_{iu}(u)\delta u \leqq 0,\ i \in I(u)\} \tag{3.4}$$

if U is convex or $\{h_{iu}(u)\}_{i \in I(u)}$ are linearly independent. For more details see [7].

Finally, the assumptions on f and $\tilde{f}$ must be strengthened. When they exist, let $f_u(x, u)$ and $\tilde{f}_u(x, u)$ denote respectively the Jacobian matrices of $f(x, u)$ and $\tilde{f}(x, u)$ with respect to u.

Assumption A1. f and $\tilde{f}$ are continuously differentiable on $X \times \hat{U}$ where $U \subset \hat{U}$ and $\hat{U} \subset R^m$ is an open set.

THEOREM 3.2 (necessary conditions for OSS). *Let f and $\tilde{f}$ satisfy Assumption* A1 *and let (u, x) solve OSS. Then there exist $p \in R^n$, $\tilde{p} \in R^l$ and real numbers $\alpha_{-j}, \cdots, \alpha_k$ such that the following conditions are satisfied for any $C(u, U)$ which is a conical approximation to U at u:*

$$(p' f_u(x, u) + \tilde{p}' \tilde{f}_u(x, u))\delta u \leqq 0 \quad \textit{for all } \delta u \in \text{cl } C(u, U), \tag{3.5-1}$$

$$\tilde{p}' = \sum_{i=-j}^{k} \alpha_i g_{iy}(y, x), \tag{3.5-2}$$

$$-p' f_x(x, u) - \tilde{p}' \tilde{f}_x(x, u) = \sum_{i=-j}^{k} \alpha_i g_{ix}(y, x), \tag{3.5-3}$$

$$\begin{aligned} &\alpha_i \leqq 0, \qquad i = -j, \cdots, 0, \\ &\alpha_i g_i(y, x) = 0, \qquad i = -j, \cdots, -1, \\ &(\alpha_{-j}, \cdots, \alpha_k, p') \neq 0. \end{aligned} \tag{3.5-4}$$

Proof. With the following substitutions OSS can be written as FDO of Appendix B: $\hat{n} = n+l$, $\mu = j$, $\nu = k+n+l$, $\hat{X} = X \times Y$, $\hat{x} = (x, y)$; for $i = -j, \cdots, k$, $\theta_i(\hat{x}, u) = g_i(y, x)$; for $i = k+1, \cdots, k+n$, $\theta_i(\hat{x}, u) = f_{i-k}(x, u)$ where the subscripts denote components of $f(x, u)$; for $i = k+n+1, \cdots, k+n+l$, $\theta_i(\hat{x}, u) = \tilde{f}_{i-k-n}(x, u) - y_{i-k-n}$ where the subscripts denote components of $\tilde{f}(x, u)$ and y. Applying the conditions from Theorem B.1, letting $p' = (\alpha_{k+1}, \cdots, \alpha_{k+n})$ and $\tilde{p}' = (\alpha_{k+n+1}, \cdots, \alpha_{k+n+l})$, gives the conditions (3.5). The last line of (3.5-4) holds because $(\alpha_{-j}, \cdots, p') = 0$ and $\tilde{p} \neq 0$ is impossible.

By changing the hypotheses other necessary conditions for OSS may be obtained.

Assumption A2. The set

$$\hat{f}(x, U) = \{(f(x, u), \tilde{f}(x, u)) : u \in U\} \subset R^{n+l} \tag{3.6}$$

is convex for all $x \in X$.

THEOREM 3.3 (maximum principle for OSS). *Let f and $\tilde{f}$ satisfy Assumption* A2. *Let (u, x) solve OSS. Then there exist $p \in R^n$, $\tilde{p} \in R^l$ and real numbers $\alpha_{-j}, \cdots, \alpha_k$ such that conditions* (3.5-2), (3.5-3), (3.5-4) *and the following condition are satisfied*:

$$\max_{v \in U} H(x, v, p, \tilde{p}) = H(x, u, p, \tilde{p}). \tag{3.5-1}$$

Proof. Make the same notational assignments as in the proof of Theorem 3.2. Applying Theorem B.2 gives (3.5-1)′ instead of (3.5-1) while everything else remains the same as in the proof of Theorem 3.2.

Remark 3.1. By applying Theorem B.1 to the maximization problem (3.5-1)′, under Assumption A1, it can be seen that (3.5-1) is a necessary condition for (3.5-1)′. Thus the conditions obtained in Theorem 3.3 are stronger than those in Theorem 3.2.

THEOREM 3.4 (necessary conditions for SSOPC). *Let* $(u(\cdot), x(\cdot), \tau) \in \mathcal{S}$(SSOPC). *Then there exist $p \in R^n$, $\tilde{p} \in R^l$ and real numbers $\alpha_{-j}, \cdots, \alpha_k$ such that conditions* (3.5-1)′, (3.5-2), (3.5-3) *and* (3.5-4) *are satisfied for $u = u(0)$ and $x = x(0)$.*

Proof. Introduce the following notation: $\alpha' = (\alpha_{-j}, \cdots, \alpha_k)$, $g'(y, x) = (g_{-j}(y, x), \cdots, g_k(y, x))$, $g_y(y, x) =$ Jacobian matrix of $g(y, x)$ with respect to y, $g_x(y, x) =$ Jacobian matrix of $g(y, x)$ with respect to x. Since $u(\cdot)$ and $x(\cdot)$ are constant let $u(t) \equiv u^*$ and $x(t) \equiv x^*$ and define: $f_x^* = f_x(x^*, u^*)$, $\tilde{f}_x^* = \tilde{f}_x^*(x^*, u^*)$, $y^* = \tilde{f}(x^*, u^*)$, $g_y^* = g_y(y^*, x^*)$, $g_x^* = g_x(y^*, x^*)$. Clearly, $(u(\cdot), x(\cdot), \sigma) \in \mathcal{S}$(SSOPC) for all $\sigma \in (0, T]$. Thus, for each σ, $(u(\cdot), x(\cdot), \sigma)$ must satisfy the conditions of Theorem 3.1. For each σ let $\alpha_i(\sigma)$, $i = -j, \cdots, k$, and $p(\sigma, \cdot)$ denote corresponding α_i and $p(\cdot)$ whose existence is guaranteed by Theorem 3.1. It is easy to show that (3.2-5) is satisfied automatically for $\tau = \sigma$ and imposes no conditions on $\alpha_i(\sigma)$ and $p(\sigma, \cdot)$. By introducing the sets

$$V^* = \{(\alpha, p) : \alpha_i \leqq 0,\ i = -j, \cdots, 0;\ \alpha_i g_i(y^*, x^*) = 0,\ i = -j, \cdots, -1\}, \tag{3.7}$$

$$C^* = \{(\alpha, p) : p'(f(x^*, v) - f(x^*, u^*)) + \alpha' g_y^*(\tilde{f}(x^*, v) - \tilde{f}(x^*, u^*)) \leqq 0 \text{ for all } v \in U\} \tag{3.8}$$

the conditions imposed by (3.2-1)–(3.2-4) on $\alpha(\sigma)$ and $p(\sigma, \cdot)$ can be written

$$(\alpha(\sigma), p(\sigma, \sigma)) \neq 0, \tag{3.9-1}$$

$$(\alpha(\sigma), p(\sigma, t)) \in V^* \cap C^* \quad \text{for all } t \in [0, \sigma], \tag{3.9-2}$$

$$\dot{p}'(\sigma, t) = -p'(\sigma, t) f_x^* - \alpha'(\sigma) g_y^* \tilde{f}_x^* \quad \text{for all } t \in [0, \sigma], \tag{3.9-3}$$

$$\sigma^{-1}(p'(\sigma, \sigma) - p'(\sigma, 0)) = \alpha'(\sigma) g_x^*. \tag{3.9-4}$$

These conditions must hold for all $\sigma \in (0, T]$; $\tilde{p}' = \alpha' g_y^*$ has been used to eliminate $\tilde{p}'$.

With the use of the variation of parameters formula condition (3.9-3) can be written

$$p'(\sigma, t) = p'(\sigma, 0)P(t) + \alpha'(\sigma)Q(t) \tag{3.10-1}$$

where the matrices $P(\cdot)$ and $Q(\cdot)$ are analytic on $[0, T]$ and satisfy the conditions: $P(0) =$ the identity matrix, $\dot{P}(0) = -f_x^*$, $Q(0) = 0$, $\dot{Q}(0) = -g_y^* \tilde{f}_x^*$. Note that if $\alpha(\sigma)$, $p(\sigma, \cdot)$ satisfy (3.9) then $\lambda\alpha(\sigma)$, $\lambda p(\sigma, \cdot)$ do also, where λ is a positive real number. Thus $\alpha(\sigma)$, $p(\sigma, \cdot)$ can always be normalized so that (3.9-1) becomes $\|\alpha(\sigma)\| + \|p(\sigma, \sigma)\| = 1$. Because of (3.10-1) and the properties of $P(\cdot)$ and $Q(\cdot)$ there therefore exists a $\hat{T} \in (0, T]$ such that (3.9-2) yields

$$\alpha(\sigma), p(\sigma, t)) \in V^* \cap C^* \cap \{(\alpha, p): .5 \leqq \|\alpha\| + \|p\| \leqq 1.5\} \quad \text{for all } t, \sigma \in [0, \hat{T}]. \tag{3.10-2}$$

Finally, by using (3.10-1) and the properties of $P(\cdot)$ and $Q(\cdot)$ it is possible to write (3.9-4) as

$$-p'(\sigma, 0)f_x^* - \alpha'(\sigma)g_y^*\tilde{f}_x^* + \gamma(\sigma) = \alpha'(\sigma)g_x^* \quad \text{for all } \sigma \in [0, \hat{T}], \tag{3.10-3}$$

where $\gamma(\sigma) \to 0$ as $\sigma \to 0$.

Now let $(\sigma_q\}$ be a sequence in $[0, \hat{T}]$ such that $\sigma_q \to 0$. From (3.7) V^* is closed and C^* is closed because it is the dual cone [20] of the set $\{(\beta, \rho): \beta = g_y^*(\tilde{f}(x^*, v) - \tilde{f}(x^*, u^*)), \rho = f(x^*, v) - f(x^*, u^*), v \in U\}$. Thus the set on the right side of (3.10-2) is compact and there exists a subsequence of $\{\sigma_q\}$, $\{\sigma_{\hat{q}}\}$, such that $\sigma_{\hat{q}} \to 0$, $\alpha(\sigma_{\hat{q}}) \to \hat{\alpha}$ and $p(\sigma_{\hat{q}}, 0) \to \hat{p}$ where $(\hat{\alpha}, \hat{p}) \in V^* \cap C^*$ and $.5 \leqq \|\hat{\alpha}\| + \|\hat{p}\| \leqq 1.5$. This shows that $\hat{\alpha}$, $\hat{p}$ satisfy (3.5-1)′ and (3.5-4) and $\hat{p}' = \hat{\sigma}' g_y^*$ satisfies (3.5-2). From (3.10-3) and $\gamma(\sigma_{\hat{q}}) \to 0$ it follows that $-\hat{p}' f_x^* - \hat{\alpha}' g_y^* \tilde{f}_x^* = \hat{\alpha} g_x^*$, which verifies (3.5-3).

Remark 3.2. The conditions in Theorems 3.3 and 3.4 are the same. Thus the reasoning used in Remark 3.1 shows that the conditions in Theorem 3.2 (with $u = u(0)$, $x = x(0)$) are necessary conditions for the elements of $\mathcal{S}$(SSOPC). However, since this (weaker) set of conditions arises from OSS it has no value in distinguishing the difference between "steady-state" and "time-dependent" control. Similar observations have been made in more restrictive circumstances by Horn and Lin [13].

Remark 3.3. It is not difficult to modify the preceding developments if τ is fixed ($\tau = T$). All the theorems are unchanged, except that condition (3.2-5) is eliminated from Theorem 3.1. The proofs are the same except: $\tau = T$ is treated as an equality constraint in the application of Theorem A.1 to the proof of Theorem 3.1, the elements of the sequence $\{\sigma_q\}$ in the proof of Theorem 3.4 are given by $\sigma_q = (q)^{-1}T$.

Several comments concerning Theorem 3.1 and its relation to previous results in the literature are in order. There are, of course, many necessary conditions which can be written. Theorem 3.1 represents a good compromise in getting strong necessary conditions with weak hypotheses. Previous derivations of necessary conditions [2], [8], [13] have required stronger assumptions, apply to more specialized problems, and have given the same or weaker conditions. It seems essential to follow a line of proof similar to that which has been taken above. The comprehensive approach taken by Bailey [2].adapts the conditions

from [18] by a change of variables. This approach applied to OPC would require the g_i, $i \neq 0$, to be twice differentiable (a hypothesis which for Bailey's problem is evident from equation (29) of [2]). Moreover, inequality constraints would be handled by the trick of Valentine which gives somewhat weaker necessary conditions ($\alpha_i \leqq 0$, $i = -j, \cdots, -1$, omitted from(3.2-4)). The requirement on the continuity of $f(x(\cdot), u(\cdot))$ and $\tilde{f}(x(\cdot), u(\cdot))$ which is needed for (3.2-5) is satisfied automatically when $u(\cdot)$ is piecewise continuous with a finite number of discontinuities. This accounts for the absence of the continuity requirement in the conditions obtained in [2]. Additional necessary conditions, e.g. the derivative condition on H expressed by equation (17) of [14], require additional hypotheses which appear to be quite strong or difficult to verify generally. The necessary conditions obtained in [5], [12] are of considerable interest, but they involve consideration of the second variation and therefore go beyond the scope of this paper.

Consider what happens if OPC is modified by replacing $g_i(y, x)$ by $g_i(y, x, \tau)$ for $i = -j, \cdots, k$. The modified OPC is not a continuing process in the sense of § 1 because requirement 3 is not satisfied. All of the preceding definitions and results can be generalized to the modified OPC, except for Theorem 3.4. The proof of Theorem 3.4 fails because $(u(\cdot), x(\cdot), \tau^*) = \mathcal{S}(\text{SSOPC})$ no longer implies $(u(\cdot), x(\cdot), \sigma) \in \mathcal{S}(\text{SSOPC})$ for all $\sigma \in (0, T]$. Since much of what follows revolves about Theorem 3.4, this shows the importance of requirement 3. A similar observation applies to the relaxation of requirement 1.

4. Relationships between the necessary conditions and the solution sets. In order to simplify references to the necessary conditions and make clearer their relationship to the solution sets introduced in § 2 it is helpful to introduce the following definitions:

$$\mathcal{S}(\text{NCOPC}) = \{(u(\cdot), x(\cdot), \tau): \text{equations (2.1-2)–(2.1-8) are satisfied and there exist } p(\cdot), \tilde{p}, \alpha_{-j}, \cdots, \alpha_k \text{ such that the conditions of Theorem 3.1 hold}\}, \tag{4.1}$$

$$\mathcal{S}(\text{NCOSS}) = \{(u(\cdot), x(\cdot), \tau): (u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{SS}) \text{ and there exist } p, \tilde{p}, \alpha_{-j}, \cdots, \alpha_k \text{ such that the conditions of Theorem 3.2 hold with } u = u(0) \text{ and } x = x(0)\}, \tag{4.2}$$

$$\mathcal{S}(\text{NCSSOPC}) = \{(u(\cdot), x(\cdot), \tau): (u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{SS}) \text{ and there exist } p, \tilde{p}, \alpha_{-j}, \cdots, \alpha_k \text{ such that the conditions of Theorem 3.4 hold}\}. \tag{4.3}$$

The set $\mathcal{S}(\text{NCOSS})$ has been defined as a subset of $\mathcal{U} \times \mathcal{X} \times (0, T]$, even though Theorem 3.2 requires $(u, x) \in U \times X$. This is done as was the case with $\mathcal{S}(\text{OSS})$ to emphasize the fact that steady-state control is a special case of time-dependent control and to allow a direct comparison of all solution sets.

With the above definitions Theorems 3.1–3.4 can be paraphrased compactly by the following inclusions: $\mathcal{S}(\text{OPC}) \subset \mathcal{S}(\text{NCOPC})$; if A1 is satisfied $\mathcal{S}(\text{OSS}) \subset \mathcal{S}(\text{NCOSS})$; if A2 is satisfied $\mathcal{S}(\text{OSS}) \subset \mathcal{S}(\text{NCSSOPC})$;

$\mathcal{S}(\text{SSOPC}) \subset \mathcal{S}(\text{NCSSOPC})$. Furthermore, if A1 is satisfied it is clear from Remark 3.2 that $\mathcal{S}(\text{NCSSOPC}) \subset \mathcal{S}(\text{NCOSS})$.

Since Theorem 3.4 was obtained from Theorem 3.1 it is tempting to surmise that $\mathcal{S}(\text{NCSSOPC}) \subset \mathcal{S}(\text{NCOPC})$. The following example shows that this conclusion is not valid.

Example 4.1. $k = j = 0$, $n = l = 1$, $X = Y = R$, $U = [-1, 1] \subset R$, $T = 1$, $f = -x + u$, $\tilde{f} = x$, $g_0 = y - \frac{3}{4}x - \frac{1}{8}x^2$. Application of the conditions in Theorem 3.4 shows that $\mathcal{S}(\text{NCSSOPC})$ is characterized by elements of the form: $u(t) \equiv x(t) \equiv 1$ or -1, $\tau \in (0, 1]$. Now consider those elements of $\mathcal{S}(\text{NCOPC})$ which also belong to $\mathcal{S}(\text{SS})$. Application of the conditions in Theorem 3.1 is more difficult because $p(\cdot)$ is not necessarily constant. However, in this example it is not difficult to integrate (3.2-3) and verify that $\mathcal{S}(\text{NCOPC}) \cap \mathcal{S}(\text{SS})$ is characterized by elements of the form: $u(t) \equiv x(t) \equiv -1$, $\tau \in (0, 1]$. The elements $u(t) \equiv x(t) \equiv 1$, $\tau \in (0, 1]$ are excluded because condition (3.2-1) requires $p(t) \leqq 0$ on $[0, \tau]$ and this turns out to be impossible. Thus $\mathcal{S}(\text{NCSSOPC}) \not\subset \mathcal{S}(\text{NCOPC})$. Under the assumption which follows it is possible to prove $\mathcal{S}(\text{NCSSOPC}) \subset \mathcal{S}(\text{NCOPC})$.

Assumption A3. The functions $g_{-j}(y, x), \cdots, g_k(y, x)$ depend only on y.

THEOREM 4.1. *Let* A3 *be satisfied. Then* $\mathcal{S}(\text{NCSSOPC}) \subset \mathcal{S}(\text{NCOPC})$.

Proof. Suppose $(u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{NCSSOPC})$ and let $u = u(0)$, $x = x(0)$. Then there exist $p \in R^n$, $\tilde{p} \in R^l$ and real numbers $\alpha_{-j}, \cdots, \alpha_k$ which satisfy (3.5-1)′, (3.5-2)–(3.5-4). Because $g_{ix}(y, x) \equiv 0$, $i = -j, \cdots, k$, this implies $p(t) \equiv p$, $\tilde{p}$, $\alpha_{-j}, \cdots, \alpha_k$ satisfy (3.2-1)–(3.2-4). Since $f(x(t), u(t)) \equiv 0$ and $y = \tilde{f}(x(\tau), u(\tau))$ condition (3.2-5) is satisfied as an equality. Thus $(u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{NCOPC})$.

Remark 4.1. For OPC problems which do not satisfy A3, Theorem 3.1 may (as Example 4.1 illustrates) offer a stronger test for $(u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{SSOPC})$ than Theorem 3.4. This is not surprising because Theorem 3.4 is obtained from Theorem 3.1 by drawing certain conclusions as $\tau \to 0$. Unfortunately, the test may be much more difficult to apply because the (constant-coefficient, linear) differential equations (3.2-3) must be considered. For OPC problems which do satisfy A3 (this includes almost all the problems which have appeared in the literature on periodic control) Theorem 4.1 shows that Theorem 3.4 provides at least as strong a test as Theorem 3.1.

Now consider a variation of Example 4.1.

Example 4.2. Same as Example 4.1, except $T = 2$. It is easy to show $\mathcal{S}(\text{NCSSOPC})$ is the same as in Example 4.1 and that $\mathcal{S}(\text{NCOPC}) \cap \mathcal{S}(\text{SS})$ is characterized by elements of the form: $u(t) \equiv x(t) \equiv -1$, $\tau \in (0, \tau^*]$. Here $\tau^* = 1.5936 \cdots$ is the positive root of $\tau = 2(1 - e^{-\tau})$. Elements of the form $u(t) \equiv x(t) \equiv -1$, $\tau \in (\tau^*, 2]$ are excluded from $\mathcal{S}(\text{NCOPC})$ because (3.2-3) shows that it is impossible for $p(t) \geqq 0$ on $(0, \tau]$ if $\tau > \tau^*$ and $p(t) \geqq 0$ is required by (3.2-1). The characterization of $\mathcal{S}(\text{NCOPC}) \cap \mathcal{S}(\text{SS})$ leads to the following observation.

Remark 4.2. Let $(u(\cdot), x(\cdot), \hat{\tau}) \in \mathcal{S}(\text{SSOPC})$. Since this implies $(u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{SSOPC})$ for all $\tau \in (0, T]$ the conditions in Theorem 3.1 apply to $(u(\cdot), x(\cdot), \tau)$ for all $\tau \in (0, T]$. If Theorem 3.1 is to be exploited fully for testing $(u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{SSOPC})$ all values of $\tau \in (0, T]$ must be considered. This is illustrated by Example 4.2. For $\tau \in (\tau^*, 2]$ there are no elements of $\mathcal{S}(\text{SS})$ which satisfy the conditions of Theorem 3.1. Thus it may be concluded that $\mathcal{S}(\text{SSOPC}) = \varnothing$ For $\tau \in (0, \tau]$ it cannot be concluded from Theorem 3.1 that $\mathcal{S}(\text{SSOPC}) = \varnothing$.

Using the results of § 2 and this section it is now possible to summarize compactly what is known about the sets $\mathcal{S}$(OPC), $\mathcal{S}$(SSOPC), $\mathcal{S}$(OSS), $\mathcal{S}$(NCOPC), $\mathcal{S}$(NCSSOPC), and $\mathcal{S}$(NCOSS).

THEOREM 4.2. (i) $\mathcal{S}(\text{SSOPC}) \subset \mathcal{S}(\text{OSS})$, (ii) $\mathcal{S}(\text{SSOPC}) \subset \mathcal{S}(\text{OPC})$, (iii) $\mathcal{S}(\text{OPC}) \subset \mathcal{S}(\text{NCOPC})$, (iv) $\mathcal{S}(\text{SSOPC}) \subset \mathcal{S}(\text{NCSSOPC})$, (v) *if* A1 *is satisfied* $\mathcal{S}(\text{OSS}) \subset \mathcal{S}(\text{NCOSS})$, (vi) *if* A1 *is satisfied* $\mathcal{S}(\text{NCSSOPC}) \subset \mathcal{S}(\text{NCOSS})$, (vii) *if* A3 *is satisfied* $\mathcal{S}(\text{NCSSOPC}) \subset \mathcal{S}(\text{NCOPC})$, (viii) *if* A2 *is satisfied* $\mathcal{S}(\text{OSS}) \subset \mathcal{S}(\text{NCSSOPC})$.

In reading the theorem it should be noted that assumptions A1 and A3 are satisfied in many applications of the theory. Assumption A2 is strong and, as will be seen later, has strong implications. Are there additional inclusions beyond those listed in the theorem? The answer is generally no, a conclusion which is made precise in the next section. The inclusions of Theorem 4.2 are summarized in Fig. 1.

$$\begin{array}{ccc}
\mathcal{S}(\text{OPC}) \subset \mathcal{S}(\text{NCOPC}) & \qquad & \mathcal{S}(\text{OPC}) \subset \mathcal{S}(\text{NCOPC}) \\
\cup \qquad\quad \cup \ \text{A3} & & \cup \qquad\quad \cup \ \text{A3} \\
\mathcal{S}(\text{SSOPC}) \subset \mathcal{S}(\text{NCSSOPC}) & & \mathcal{S}(\text{SSOPC}) \subset \mathcal{S}(\text{OSS}) \subset \mathcal{S}(\text{NCSSOPC}) \\
\cap \qquad\quad \cap \ \text{A1} & & \cap \ \text{A1} \\
\mathcal{S}(\text{OSS}) \subset \mathcal{S}(\text{NCOSS}) & & \mathcal{S}(\text{NCOSS}) \\
\text{A1} & & \\
\text{Without A2} & & \text{With A2}
\end{array}$$

FIG. 1. *Summary of Theorem* 4.2. *See* (2.3), (2.4), (2.5), (4.1), (4.2) *and* (4.3) *for definitions of solution sets.*

The results of the previous section are also related to the solution sets of relative minima. For example, let $(u(\cdot), x(\cdot), \cdot\tau) \in \mathcal{S}(\text{SRMSSOPC})$. Then if X is replaced by $X \cap \{\hat{x}: \|\hat{x} - x(0)\| < \varepsilon\}$, $\varepsilon > 0$ sufficiently small, $(u(\cdot), x(\cdot), \tau)$ is a regular minimum and the conditions of Theorem 3.4 apply without change to $(u(\cdot), x(\cdot), \tau)$. Thus $(u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{NCSSOPC})$. Similar arguments apply to weak relative minima but in the cases of Theorems 3.3 and 3.4 it is necessary to introduce a weak form of the maximum condition,

$$\max_{v \in V,\, \|v-u\| < \varepsilon} H(x, v, p, \tilde{p}) = H(x, u, p, \tilde{p}), \tag{4.4}$$

and define

$$\begin{aligned}
\mathcal{S}(\text{WNCSSOPC}) = \{&(u(\cdot), x(\cdot), \tau): \ (u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{SS}) \text{ and there} \\
&\text{exist } p, \tilde{p}, \alpha_{-j}, \cdots, \alpha_k \text{ such that conditions (4.7), (3.5-} \\
&\text{2), (3.5-3) and (3.5-4) hold with } u = u(0) \text{ and } x = x(0) \\
&\text{for some } \varepsilon > 0\}.
\end{aligned} \tag{4.5}$$

In addition the following assumption, which is not necessarily stronger than A2, must be introduced.

Assumption A4. There exists an $\bar{\varepsilon} > 0$ such that the set

$\hat{f}(x, U \cap \{u: \|u-v\| \leqq \varepsilon)\})$ (see (3.6) for notation) is convex for all $x \in X$, $v \in U$, $\varepsilon \in (0, \bar{\varepsilon}]$.

The conclusions which follow along with the inclusions of Remark 2.2 are summarized as follows.

THEOREM 4.3. *The inclusions displayed in Fig. 2 are valid.*

Some applications of these inclusions are discussed in § 6.

$$\begin{array}{ccccc}
 & & \mathscr{S}(\text{NCSSOPC}) & \subset \mathscr{S}(\text{WNCSSOPC}) & \subset \mathscr{S}(\text{NCOSS}) \\
 & & \cup & \cup & \text{A1} \\
\mathscr{S}(\text{SSOPC}) & \subset \mathscr{S}(\text{SRMSSOPC}) & & \subset \mathscr{S}(\text{WRMSSOPC}) & \\
\cap & \cap & & \cap & \\
\mathscr{S}(\text{OSS}) & \subset \mathscr{S}(\text{SRMOSS}) & & \subset \mathscr{S}(\text{WRMOSS}) & \subset \mathscr{S}(\text{NCOSS}) \\
 & \cap \text{A2} & & \cap \text{A4} & \text{A1} \\
 & & \mathscr{S}(\text{NCSSOPC}) & \subset \mathscr{S}(\text{WNCSSOPC}) & \subset \mathscr{S}(\text{NCOSS}) \\
 & & & & \text{A1}
\end{array}$$

FIG. 2. *Theorem* 4.3

5. Some examples. The examples of this section serve a number of purposes. First, they show that it is not possible to prove more inclusions than those which are contained in Theorem 4.2; this conclusion is formalized in Theorem 5.1 and extended somewhat in Theorem 5.2. Second, they delimit certain tests for optimality; this is discussed in the next section. Finally, they provide insight into the difficulties of applying and solving the various necessary conditions and into the wide variety of circumstances and phenomena which can occur in OPC problems.

Example 5.1. $k=j=0$, $n=l=1$, $X=Y=U=R$, $T>\sqrt{2}\pi$, $f=-x^2+u$, $\tilde{f}=-2x^2+u^2$, $g_0=y$. The assumption $T=\sqrt{2}\pi$ is sufficient to assure that the characterization of the solution sets is not changed by T. If $T<\sqrt{2}\pi$ one element of $\mathscr{S}$(NCOPC) disappears (d below) and everything else remains the same.

Omitting details, the conditions contained in Theorem 3.1 can be summarized as follows. From (3.2-2), $\tilde{p}=\alpha_0$. It is easy to show that for $\alpha_0=0$, (3.2) cannot have a solution and without loss of generality the case $\alpha_0<0$ can be treated as $\alpha_0=-1$. Condition (3.2-1) gives

$$u=\tfrac{1}{2}p. \tag{5.1}$$

The remaining conditions are (3.2-5) and

$$\dot{x}=-x^2+\tfrac{1}{2}p, \qquad \dot{p}=2xp-4x, \tag{5.2}$$

$$x(0)=x(\tau), \qquad p(0)=p(\tau), \qquad \tau \in (0, T]. \tag{5.3}$$

Figure 3 shows the (x, p)-phase plane for (5.2). Each characteristic curve corresponds to a fixed value of H_M in the relation $H_M=-px^2+\frac{1}{4}p^2+2x^2$. The points labeled a, b and c are constant solutions of (5.2) and (5.3) and satisfy (3.2-5) with $\tilde{p}'y=H_M$ for all $\tau \in (0, T]$. The only other solutions of (5.2) and (5.3) are d, which has period T, and all the other solutions "inside" d (excluding c) which have

periods $\tau \in (\sqrt{2}\pi, T)$. Calculation shows that for all these "time-dependent" solutions of (5.2) and (5.3), $\tilde{p}'y = -y < H_M$. Thus by (3.2-5) d is the only "time-dependent" solution of the conditions in Theorem 3.1. Because for each $(x(\cdot), p(\cdot), \tau)$ there is a corresponding $(u(\cdot), x(\cdot), \tau)$ the labels a, b, c and d can be used also to designate sets of elements in $\mathcal{U} \times \mathcal{X} \times (0, T]$. In particular, $\mathcal{S}$(NCOPC) "corresponds" to a, b, c and d, i.e., it is the union of elements designated a, b, c and d.

By using Theorems 3.2 and 3.4 it may be verified that both $\mathcal{S}$(NCOSS) and $\mathcal{S}$(NCSSOPC) correspond to a, b and c. Moreover, $\mathcal{S}$(OSS) corresponds to a and b and the cost associated with a and b is $J = -1$. Suppose there exist $u(\cdot), x(\cdot)$ and τ which satisfy (2.1-2)–(2.1-8) and give $J < -1$. This implies

$$-1 > \frac{1}{\tau}\int_0^\tau (-2x^2 + u^2)\, dt = \frac{1}{\tau}\int_0^\tau (-2x^2 + u^2 - 2\dot{x})\, dt$$

(5.4)

$$= \frac{1}{\tau}\int_0^\tau (u^2 - 2u)\, dt$$

which in turn implies

$$0 > \frac{1}{\tau}\int_0^\tau (u^2 - 2u + 1)\, dt = \frac{1}{\tau}\int_0^\tau (u-1)^2\, dt. \tag{5.5}$$

This inequality is false and thus a and b are "contained" in $\mathcal{S}$(OPC). Any additional elements of $\mathcal{S}$(OPC) must be elements of $\mathcal{S}$(NCOPC). But c has cost $J = 0$ and it can be shown that d has cost $J > -1$. Thus $\mathcal{S}$(OPC) corresponds to a and b and $\mathcal{S}$(SSOPC) = $\mathcal{S}$(OPC).

The above results are summarized in the first line of Table 1. It is easy to show that $\mathcal{S}$(SRMOSS), $\mathcal{S}$(WRMOSS), $\mathcal{S}$(SRMSSOPC) and $\mathcal{S}$(WRMSSOPC) all correspond to a and b. The element d is a "time-dependent" strong relative minimum of OPC.

Example 5.2. $k = j = 0$, $n = l = 1$, $X = Y = R$, $U = [-2, 2] \subset R$, $T > 0$, $f = -x + u + 1$, $\tilde{f} = x(u+1)(u-1)^2$, $g_0 = y$. For $(u(\cdot), x(\cdot), \tau) \in \mathcal{S}$(SS), $x = u + 1$ and $y = (u+1)^2(u-1)^2$. Thus $\mathcal{S}$(OSS) corresponds to $u(t) \equiv 1$, $x(t) \equiv 2$, $\tau \in (0, T]$ (labeled a) and $u(t) \equiv -1$, $x(t) \equiv 0$, $\tau \in (0, T]$ (labeled b). Consideration of (3.5) shows that $\mathcal{S}$(NCOSS) corresponds to a, b and $u(t) \equiv 0$, $x(t) \equiv 1$, $\tau \in (0, T]$ (labeled c).

Theorem 3.1 leads to the characterization of $\mathcal{S}$(NCOPC). From (3.2-2), $\tilde{p} = \alpha_0$ and inspection of (3.2-3) and (3.2-4) shows that $\alpha_0 = 0$ is impossible. Thus without loss of generality assume $\alpha_0 = -1$. The maximization of

$$H = p(-x + u + 1) - x(u+1)(u-1)^2 \tag{5.6}$$

with respect to $u \in U$ is complicated somewhat by the fact that the maximizing u may be in the interior or in the boundary of U, depending on x and p. Let L_1, L_2, L_3, L_4 be rays emanating from the origin of the (x, p) plane which do not contain

the origin and have, respectively, slopes: $1.2095\cdots$ (the root of $\frac{16}{27}+\frac{16}{27}\sqrt{1+\frac{3}{4}q}^{3}-\frac{1}{3}q=0$), 7, $2.5097\cdots$ (the root of $-\frac{16}{27}+\frac{16}{27}\sqrt{1+\frac{3}{4}q}^{3}+\frac{1}{3}q=0$), 15. Let A_1, A_2, A_3,

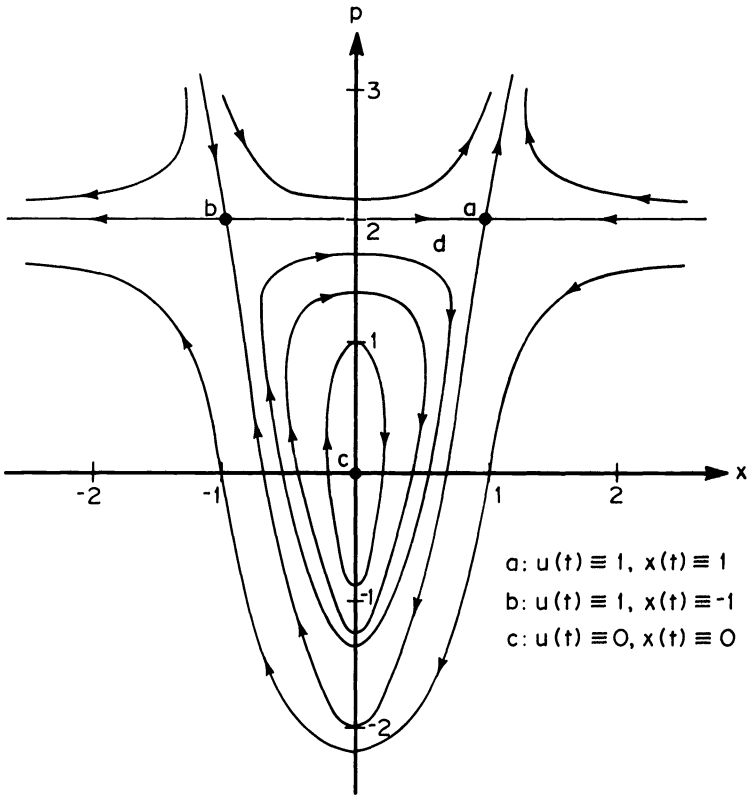


FIG. 3. *(x, p)-phase plane for Example* 5.1

TABLE 1
Characterization of solution sets for examples

Example	$\mathcal{S}$(SSOPC)	$\mathcal{S}$(OSS)	$\mathcal{S}$(NCSSOPC)	$\mathcal{S}$(NCOSS)	$\mathcal{S}$(OPC)	$\mathcal{S}$(NCOPC)	J OSS	J OPC
5.1, 5.7	a, b	a, b	a, b, c	a, b, c	a, b	a, b, c, d	−1	−1
5.2	∅	a, b	b	a, b, c	∅	b	0	*
5.3(i), 5.8(i)	∅	a	a	a	b	a, b, c, d	0	−1
5.3(ii), 5.8(ii)	a	a	a	a	a, d	a, b, c, d	−1	−1
5.4	a, b	a, b	a, b	a, b, c	a, b	a, b	0	0
5.5	∅	a, b	a, b, c	a, b, c	d	a, b, c, d (?)	0	<0
5.6	∅	a, b	a, b	a, b, c	d	a, b, d	0	−1

* Minimum does not exist.

A_4 be the open sectors bounded by these rays (see Fig. 4). Then the maximizing u is given by

$$
\begin{aligned}
u &= 2, && (x,p)\in A_1\cup L_2,\\
&= \tfrac{1}{3}+\tfrac{2}{3}\sqrt{1+\tfrac{3}{4}\tfrac{p}{x}}, && (x,p)\in A_2,\\
&= -2 \text{ or } 1.4651\cdots, && (x,p)\in L_3,\\
&= -2, && (x,p)\in A_3\cup L_4,\\
&= \tfrac{1}{3}-\tfrac{2}{3}\sqrt{1+\tfrac{3}{4}\tfrac{p}{x}}, && (x,p)\in A_4,\\
&= 2 \text{ or } -.5873\cdots, && (x,p)\in L_1,\\
&\in [-2,2], && x=p=0.
\end{aligned} \tag{5.7}
$$

Conditions (2.1-5) and (3.2-3) yield

$$\dot{x} = -x+u+1, \qquad \dot{p} = p+(u+1)(u-1)^2, \tag{5.8}$$

$$x(0)=x(\tau), \qquad p(0)=p(\tau), \qquad \tau\in(0,T]. \tag{5.9}$$

With u given by (5.7), equations (5.8) lead to the characteristic curves shown in the (x,p)-phase plane, Fig. 4. The point $x=p=0$ corresponds to a constant solution if and only if $u(t)\equiv -1$. Points on the ray L_1 below P_1 correspond to a discontinuity in $u(t)$ ($u(t)$ at the discontinuity may be defined to be either 2 or $-.5873\cdots$). Above P_1 solutions of the system (5.7)–(5.8) cannot be continued across L_1 because from both A_1 and A_4 they lead into L_1. On L_3 solutions of (5.7)–(5.8) intersecting above P_3 or below P_2 can be continued across L_3 with a discontinuity in $u(t)$. On L_3 between P_2 and P_3 solutions lead away from L_3, going upward if $u(0)=1.4651\cdots$ and downward if $u(0)=2$. Thus the only solution of (5.7)–(5.8) which satisfies (5.9) is $u(t)\equiv -1$, $x(t)\equiv 0$, $p(t)\equiv 0$. This solution also satisfies (3.2-5) for all $\tau\in(0,T]$ and hence $\mathcal{S}$(NCOPC) corresponds to b in Table 1. It is also clear that $\mathcal{S}$(NCOPC) = $\mathcal{S}$(NCSSOPC).

The following argument shows that $\mathcal{S}$(SSOPC) = $\varnothing$. Suppose to the contrary. Then $\mathcal{S}$(SSOPC) $\subset$ $\mathcal{S}$(NCSSOPC) implies that $\mathcal{S}$(SSOPC) corresponds to b in Table 1. But this contradicts $\mathcal{S}$(OSS) = $\mathcal{S}$(SSOPC) (Remark 2.1). Finally, $\mathcal{S}$(OPC) = $\varnothing$ because there are no "time-dependent" solutions of (5.7)–(5.8).

The above results are summarized in Table 1. Perhaps the most interesting conclusion is that $\mathcal{S}$(NCSSOPC) is a proper subset of $\mathcal{S}$(OSS). Clearly, $\mathcal{S}$(SRMOSS) = $\mathcal{S}$(WRMOSS) correspond to a and b. It is not difficult to show that with weak variations from $u(t)\equiv -1$, $J<0$ can be obtained. To obtain $J<0$ in the neighborhood of $x(t)\equiv 2$ it is necessary to use strong variations from $u(t)\equiv 1$. Thus $\mathcal{S}$(SRMSSOPC) = $\varnothing$ and $\mathcal{S}$(WRMSSOPC) corresponds to a.

Example 5.3. $k=0$, $j=1$, $n>0$, $l=2$, $x=R^n$, $Y=R^2$, $U=R$, $T>0$, $f=Ax+\mathbf{b}u$, $\tilde{f}_1=-\frac{1}{2}(\mathbf{c}'x)^2$, $\tilde{f}_2=\frac{1}{2}u^2$, $g_0=y_1$, $g_{-1}=y_2-1$. A is a real $n\times n$ matrix and $\mathbf{b}, \mathbf{c}\in R^n$. This example is a special case of the problem considered in [6]. If $\mathbf{c}'x$ is interpreted as the output of the linear system $\dot{x}=Ax+\mathbf{b}u$, it corresponds to maximizing the average output power subject to a constraint on the average input power. Assume A is stable (characteristic roots of A have negative real parts), $(A,\mathbf{b})$ is controllable, $(\mathbf{c}', A)$ is observable [22] and let

$$G(s)=\mathbf{c}'(Is-A)^{-1}\mathbf{b} \tag{5.10}$$

denote the system transfer function.

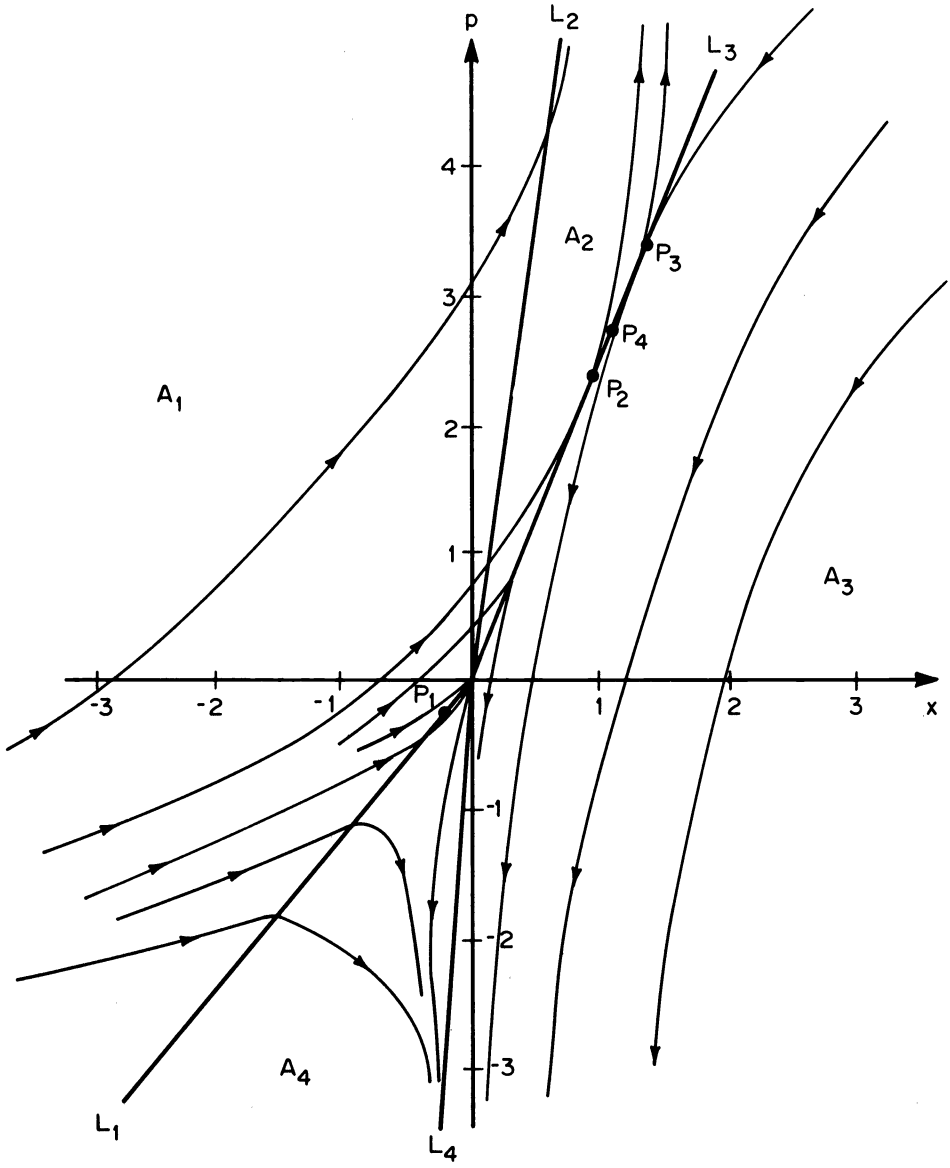


FIG. 4. (x, p)-*phase plane for Example* 5.2

Consider the characterization of $\mathcal{S}$(NCOPC). From (3.2-2), $\tilde{p}' = (\alpha_0, \alpha_{-1})$. First, assume $\alpha_{-1} \neq 0$. Conditions (3.2-1), (2.1-5) and (3.2-3) give

$$u = (\alpha_{-1})^{-1} p' \mathbf{b}, \tag{5.11}$$

$$\dot{x} = Ax + \mathbf{b}u, \qquad \dot{p}' = -p'A + (\alpha_0 \mathbf{c}' x)\mathbf{c}', \tag{5.12}$$

$$x(0) = x(\tau), \qquad p(0) = p(\tau), \qquad \tau \in (0, T]. \tag{5.13}$$

Since A has no characteristic roots with zero real parts, $\alpha_0 = 0$ implies $p(t) \equiv 0$ and thus $u(t) \equiv 0$. But this gives $g_{-1}(y) < 0$ which contradicts (3.2-4). Since $\alpha_0 = 0$ is

impossible, take $\alpha_0 = -1$. The system (5.11)–(5.12) is a linear, constant-coefficient, differential system of order $2n$ which has a periodic solution if and only if the characteristic equation has at least one root with real part zero. A simple calculation shows that there exist characteristic roots $\pm \mathbf{i}\omega$ ($\omega \in R, \omega \geqq 0$, $\mathbf{i} = \sqrt{1}$) if and only if

$$-\alpha_{-1} = G(\mathbf{i}\omega)G(-\mathbf{i}\omega) = |G(\mathbf{i}\omega)|^2, \tag{5.14}$$

an equation which always has a solution for $\alpha_{-1} < 0$ because controllability and observability imply $G(\mathbf{i}\omega) \not\equiv 0$. Since $y_2 = 1$ for $\alpha_{-1} < 0$, there must exist a $u(t)$ satisfying (5.11)–(5.13) of the form

$$u(t) = \begin{cases} 2\cos(\omega t + \theta), & \omega \geqq \dfrac{2\pi}{T}, \\ \pm\sqrt{2}, & \omega = 0, \end{cases} \tag{5.15}$$

where θ is arbitrary. The only remaining condition which must be satisfied if (3.2-5). A rather lengthy but straightforward computation shows that $H_M - \tilde{p}'y = -\omega(d/d\omega)|G(\mathbf{i}\omega)|^2$. Thus ω in (5.14) is a permissible value if and only if

$$\begin{aligned} &\omega = 0 \quad \text{or} \\ &\omega > \frac{2\pi}{T} \quad \text{and} \quad \frac{d}{d\omega}|G(\mathbf{i}\omega)|^2 = 0 \quad \text{or} \\ &\omega = \frac{2\pi}{T} \quad \text{and} \quad \frac{d}{d\omega}|G(\mathbf{i}\omega)| \leqq 0. \end{aligned} \tag{5.16}$$

Now consider $\alpha_{-1} = 0$. The possibility $\alpha_0 = 0$ is excluded because it implies $p(t) \equiv 0$ which violates (3.2-4). Thus take $\alpha_0 = -1$. Then it follows from (3.2-1) that $p'(t)\mathbf{b} \equiv 0$ on $[0, \tau]$ and u is not determined by (3.2-1), i.e., u is a singular control. The condition $p'(t)\mathbf{b} \equiv 0$ can be shown to imply: $u(t) = q\cos(\omega t + \theta)$ where θ is arbitrary, $0 \leqq q \leqq 2$, $\omega \geqq 2\pi/T$, $G(\mathbf{i}\omega) = 0$; or $u(t) \equiv q$ where $-\sqrt{2} \leqq q \leqq \sqrt{2}$, $G(0) = 0$. Thus for $\alpha_{-1} = 0$ the conditions on ω agree with (5.14) and (5.16) (observe that $(d/d\omega)|G(\mathbf{i}\omega)|^2 = 0$ for ω such that $G(\mathbf{i}\omega) = 0$). In the (relatively rare) circumstance that (5.14) and (5.16) permit multiple solutions ($\omega = \omega_1, \cdots, \omega_K$ satisfies (5.16) and $|G(\mathbf{i}\omega_i)|^2 = |G(\mathbf{i}\omega_1)|^2 = -\alpha_{-1}$, $i = 2, \cdots, K$) and $u(t) = \sum_{i=1}^{K} U_i \cos(\omega_i t + \theta_i)$ is periodic with period $\tau \leqq T$ then this $u(t)$ corresponds to a family of solutions of the necessary conditions for OPC provided the U_i are chosen so that $g_{-1}(y) = 0$ (or $g_{-1}(y) \leqq 0$ if $\alpha_{-1} = G(\mathbf{i}\omega_1) = 0$).

Application of Theorem 3.4 shows that $\mathcal{S}$(NCSSOPC) may be obtained by specializing the above results to the case where $x(t)$ and $p(t)$ are constant. Thus for $G(0) \neq 0$: $u(t) \equiv \pm\sqrt{2}$, $x(t) \equiv \mp A^{-1}\mathbf{b}\sqrt{2}$, $\tau \in (0, T]$ corresponds to $\mathcal{S}$(NCSSOPC). For $G(0) = 0$: $\mathcal{S}$(NCSSOPC) corresponds to $u(t) \equiv q$, $x(t) \equiv -A^{-1}\mathbf{b}q$, $\tau \in (0, T]$, $q \in [-\sqrt{2}, \sqrt{2}]$. It is also clear from the form of H and U that $\mathcal{S}$(NCOSS) = $\mathcal{S}$(NCSSOPC).

Simple arguments (see [6]) show that $\mathcal{S}(\text{OPC}) \neq \varnothing$. Since for elements in $\mathcal{S}(\text{NCOPC})$, $g_0(y) = -|G(\mathbf{i}\omega)|^2$ it is clear that $\mathcal{S}(\text{OPC})$ corresponds to those elements in $\mathcal{S}(\text{NCOPC})$ with ω maximizing $|G(\mathbf{i}\omega)|^2$ on $\{0\} \cup [2\pi/T, +\infty)$. The maximum exists and is positive (because $|G(\mathbf{i}\omega)|^2 > 0$ for some ω and $|G(\mathbf{i}\omega)|^2 \to 0$ as $\omega \to +\infty$) and can occur only at a finite number of frequencies (because $|G(\mathbf{i}\omega)|^2$ is rational in ω^2). $\mathcal{S}(\text{SSOPC}) \neq \varnothing$ if and only if $|G(\mathbf{i}\omega)|^2 \leqq G^2(0)$ for all $\omega \geqq 2\pi/T$. $\mathcal{S}(\text{OSS})$ corresponds to $u(t) \equiv \pm\sqrt{2}$, $x(t) \equiv \mp A^{-1}\mathbf{b}\sqrt{2}$, $\tau \in (0, T]$ if $G(0) \neq 0$ and to $u(t) \equiv q$, $x(t) = -A^{-1}\mathbf{b}q$, $\tau \in (0, T]$, $q \in [-\sqrt{2}, \sqrt{2}]$ if $G(0) = 0$.

Since the elements of the solution sets are characterized in terms of $|G(\mathbf{i}\omega)|^2$ it is easy to determine them even though n may be large. Figure 5 gives two cases whose solution sets are summarized in Table 1. With the possible exception of d in Case (ii) it should be obvious what is meant by the designations of the solutions. For d, $u(t) = U_1 + U_2 \cos(\tilde{\omega}t + \theta)$, $\theta \in R$, $\frac{1}{2}U_1^2 + \frac{1}{4}U_2^2 = 1$, $\tau = (\tilde{\omega})^{-1}2\pi$. It is clear that $\mathcal{S}(\text{NCOPC})$ may contain many more elements than $\mathcal{S}(\text{OPC})$. Unfortunately, for most other OPC problems the suboptimal extremals are not so easily determined and rejected as they are in this example.

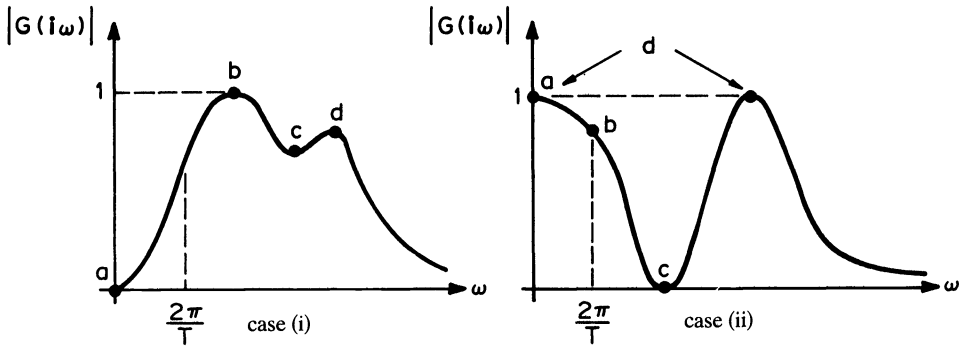


FIG. 5. *Designation of solution sets for Example* 5.3

Example 5.4. $k = j = 0$, $n = l = 1$, $X = Y = U = R$, $T > 0$, $f = -x + (u-1)^2(u+1)^2$, $\tilde{f} = x$, $g_0 = y$. Make the following designations: (a) $u(t) \equiv 1$, $x(t) \equiv 0$, $\tau \in (0, T]$; (b) $u(t) \equiv -1$, $x(t) \equiv 0$, $\tau \in (0, T]$; (c) $u(t) \equiv 0$, $x(t) \equiv 1$, $\tau \in (0, T]$. Then the characterizations for $\mathcal{S}(\text{OSS})$, $\mathcal{S}(\text{NCOSS})$ and $\mathcal{S}(\text{NCSSOPC})$ given in Table 1 can be verified easily. Inspection of (3.2-3) shows that only the allowed solution for $p(t)$ is $p(t) \equiv \alpha_0$. This implies $\mathcal{S}(\text{NCOPC}) = \mathcal{S}(\text{NCSSOPC})$. From (2.1-5) and the form of f it follows that $x(t) \geqq 0$ for all $t \in (0, T]$. This implies $J = y \geqq 0$ and $J = 0$ is only possible if $x(t) \equiv 0$. Thus $\mathcal{S}(\text{OPC}) = \mathcal{S}(\text{SSOPC}) = \mathcal{S}(\text{OSS})$.

Example 5.5. $k = j = 0$, $n = 2$, $l = 1$, $X = R^2$, $Y = U = R$, $T = 2\pi$, $f_1 = x_2$, $f_2 = -x_1 - x_2 + u$, $\tilde{f} = (x_1 - 1)^2(x_1 + 1)^2 - (x_2)^2$, $g_o = y$. Make the following designations: (a) $u(t) \equiv x_1(t) \equiv 1$, $x_2(t) \equiv 0$, $\tau \in (0, T]$; (b) $u(t) \equiv x_1(t) \equiv -1$, $x_2(t) \equiv 0$, $\tau \in (0, T]$; (c) $u(t) \equiv x_1(t) \equiv x_2(t) \equiv 0$, $\tau \in (0, T]$. Then the characterizations for $\mathcal{S}(\text{OSS})$, $\mathcal{S}(\text{NCOSS})$ and $\mathcal{S}(\text{NCSSOPC})$ given in Table 1 can be verified easily. Let $u(t) = 1 + A \cos \omega t$. Then y may be computed easily from (2.1-4) and (2.1-5). For $\omega > \sqrt{2}$ and $A > 0$ sufficiently small the computation shows that $y < 0$. Since the optimal cost for OSS is $J = 0$ this proves that $\mathcal{S}(\text{SSOPC}) = \varnothing$. From standard

existence theorems it follows that $\mathscr{S}(\text{OPC}) \neq \varnothing$. Let the elements of $\mathscr{S}(\text{OPC})$ be designated by d. Since A3 is satisfied it is clear from Theorem 4.1 that a, b, c, d are "included" in $\mathscr{S}(\text{NCOPC})$. It is not known if there are additional elements in $\mathscr{S}(\text{NCOPC})$.

Example 5.6. $k=0, j=1, n=l=2, X=Y=R^2, U=\{u: u_3 \geqq u_2^2\} \subset R^3, T=3\pi$, $f_1=x_2$, $f_2=-x_1-x_2+u_2$, $\tilde{f}_1=(u_1-1)^2(u_1+1)^2-\frac{1}{2}(x_2)^2$, $\tilde{f}_2=\frac{1}{2}u_3$, $g_0=y_1$, $g_{-1}=y_2-1$. In each of the following designations assume that $x_1(t) \equiv u_2(t) \equiv q_2$, $u_3(t) \equiv q_3$, $x_2(t) \equiv 0$, $q_3 \in [0, 2]$, $q_3 \geqq q_2^2$, $\tau \in (0, T]$: (a) $u_1(t) \equiv 1$, (b) $u_1(t) \equiv -1$, (c) $u_1(t) \equiv 0$. The characterizations of $\mathscr{S}(\text{OSS})$, $\mathscr{S}(\text{NCOSS})$ and $\mathscr{S}(\text{NCSSOPC})$ are given in Table 1. To minimize J in OPC it is necessary and sufficient to separately minimize the average of each of the two terms in $\tilde{f}_1(x(t), u(t))$. The first term is minimized by $u_1(t) \equiv \pm 1$ and the second term leads to a minimization problem of the type considered in Example 5.3, because at the minimum $u_3(t) = (u_2(t))^2$ (see also Example 5.8). This problem has a solution of the form: $u_2(t) = x_2(t) = 2\cos(t+\theta)$, $u_3(t) = 4\cos^2(t+\theta)$, $x_1(t) = 2\sin(t+\theta)$ $\tau = 2\pi$, $\theta \in R$. Let the set of all $(u(\cdot), x(\cdot), \tau)$ characterized in the above fashion be denoted by d. Then $\mathscr{S}(\text{OPC})$ corresponds to d. It can be shown that $\mathscr{S}(\text{NCOPC})$ corresponds to a, b and d.

Example 5.7. Same as Example 5.1, except for the following changes: $U=\{u: u_2 \geqq u_1^2\} \subset R^2, f=-x^2+u_1, \tilde{f}=-2x^2+u_2$. This example is essentially the same as Example 5.1. This can be seen by observing that in the characterization of all the solution sets it is required that $u_2(t) = (u_1(t))^2$. Thus the designations in Table 1 hold if: $u_1(t) = u(t)$, $u_2(t) = (u(t))^2$ where $u(t)$ is given as in Example 5.1; $x(t)$ is the same as $x(t)$ in Example 5.1.

Example 5.8. Same as Example 5.3, except for the following changes: $U=\{u: u_2 \geqq u_1^2\} \subset R^2$, $f=Ax+\mathbf{b}u_1$, $\tilde{f}=\frac{1}{2}u_2$. The modifications are similar to those used in Example 5.7. This leads to the designations shown in Table 1.

An immediate application of the examples is the following theorem.

THEOREM 5.1. *Let* A1 {A1 *and* A3} [A1, A2 *and* A3] *be satisfied. Then it is not possible to obtain additional inclusions beyond those which are implied by* (i)–(vi) {(i)–(vii)} [(i)–(viii)] *of Theorem* 4.2.

Proof. Of the 30 nontrivial, pair-wise inclusions involving $\mathscr{S}(\text{SSOPC})$, $\mathscr{S}(\text{OSS})$, $\mathscr{S}(\text{NCSSOPC})$, $\mathscr{S}(\text{NCOSS})$, $\mathscr{S}(\text{OPC})$, $\mathscr{S}(\text{NCOPC})$ which are possible (i)–(vi) {(i)–(vii)} [(i)–(viii)] of Theorem 4.2 imply that 8 {9} [11] are satisfied. Examples 4.1, 5.1, 5.2, 5.3(ii) {5.1, 5.2, 5.3(ii)} [5.5, 5.6] show that with A1 {A1 and A3} [A1, A2 and A3] satisfied the remaining 22 {21} [19] inclusions cannot hold generally.

Now consider the effect of stronger assumptions. Suppose as is the case in many practical problems that A1, A3 and $\mathscr{S}(\text{OSS}) \neq \varnothing$ are satisfied. Additional assumptions which are of interest are (i) OPC is proper ($\mathscr{S}(\text{SSOPC}) = \varnothing$), (ii) OPC is steady-state ($\mathscr{S}(\text{SSOPC}) \neq \varnothing$), (iii) OPC is proper and A2 is satisfied, (iv) OPC is steady-state and A2 is satisfied. For each of these cases Theorem 4.2 yields certain implications which are summarized in Fig. 6. It does not follow from Theorem 5.1 that these are the only implications concerning inclusion which can be drawn. However, the examples do show this. For instance, suppose that (ii) holds. Then Fig. 6 implies 13 nontrivial, pair-wise inclusions; Examples 5.1, 5.3(ii), 5.4 (which satisfy A1, A3, $\mathscr{S}(\text{OSS}) \neq \varnothing$, and (ii)) imply that the remaining

17 pair-wise inclusions cannot hold. All of the results are summarized in the following theorem.

THEOREM 5.2. *Let* A1, A3 *and* $\mathscr{S}(OSS) \neq \varnothing$ *be satisfied. Under the additional hypotheses* (i), (ii), (iii) *or* (*iv*) *the results of Fig.* 6 *are true. In each of the four cases it is not possible to prove additional inclusions exist.*

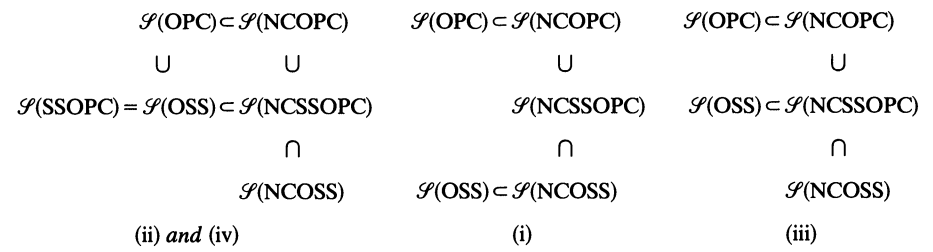


FIG. 6. *Inclusions which are satisfied under* A1, A3, $\mathscr{S}(OSS) \neq \varnothing$ *and*: (i) OPC *is proper*, (ii) OPC *is steady-state*, (iii) OPC *is proper and* A2, *or* (iv) OPC *is steady-state and* A2

6. Tests for optimality. If $\mathscr{S}$(OPC) and $\mathscr{S}$(OSS) are known it is possible to determine immediately whether or not time-dependent control improves performance and, if it does, the amount of the improvement. Since in most practical problems the solutions of OPC are not obtained easily, other paths must be pursued. One such path is suggested by Fig. 6. Under assumptions A1 and A3 it is clear that $\mathscr{S}(OSS) \not\subset \mathscr{S}(NCSSOPC)$ implies that OPC is proper. Thus it can be determined that $\mathscr{S}(SSOPC) = \varnothing$ without obtaining $\mathscr{S}$(OPC). This motivates the class of tests investigated in this section. Triples $\psi = (u(\cdot), x(\cdot), \tau) \in \mathscr{S}(SS)$ are considered and it is supposed that it is possible to determine whether or not $\psi \in \mathscr{S}(A)$ for certain A. The principal concern is if OPC is proper or steady-state, but tests which may help in the search for solutions of OPC are examined too. The tests generalize (to OPC) and supplement tests which have appeared in the literature. An entirely new result is Theorem 6.1 which establishes limits to what can be tested in certain contexts.

To be complete the idea of Remark 4.2 is incorporated into the discussion. The condition given there corresponds to checking $\psi \in \mathscr{S}(NC'OPC)$ where

$$\mathscr{S}(NC'OPC) = \{(\hat{u}(\cdot), \hat{x}(\cdot), \hat{\tau}):\ (\hat{u}(\cdot), \hat{x}(\cdot), \hat{\tau}) \in \mathscr{S}(SS) \text{ and for all } \tau \in [0, T] \text{ there exist } p(\cdot), p, \alpha_{-j}, \cdots, \alpha_k \text{ such that (3.2-1)–(3.2-4) are satisfied for } u(t) \equiv \hat{u}(0), x(t) \equiv \hat{x}(0)\}. \tag{6.1}$$

By tracing the proof of Theorem 3.4, it is easy to see that $\mathscr{S}(NC'OPC) \subset \mathscr{S}(NCSSOPC)$. Moreover, under A3, Theorem 4.1 states that $\mathscr{S}(NCSSOPC) \subset \mathscr{S}(NCOPC)$; since $(u(\cdot), x(\cdot), \hat{\tau}) \in \mathscr{S}(NCSSOPC)$ implies $(u(\cdot), x(\cdot), \tau) \in \mathscr{S}(NCSSOPC)$ for all $\tau \in (0, T]$, this shows that $\psi \in \mathscr{S}(NCSSOPC)$ implies $\psi \in \mathscr{S}(NC'OPC)$. These facts and the content of Remark 4.2 are summarized in

Remark 6.1. $\mathscr{S}(NC'OPC)$ satisfies the following inclusions: $\mathscr{S}(SSOPC) \subset \mathscr{S}(NC'OPC) \subset \mathscr{S}(NCSSOPC)$. If A3 is satisfied, $\mathscr{S}(NC'OPC) = \mathscr{S}(NCSSOPC)$.

From this and the results of § 4, it is clear that the following tests are valid.

Test T1. The existence of ψ, $\psi \in \mathscr{S}(SS)$, $\psi \in \mathscr{S}(OPC)$, implies OPC is steady-state.

Test T2. The existence of ψ, $\psi \in \mathcal{S}(\text{OSS})$, $\psi \notin \mathcal{S}(\text{OPC})$, implies OPC is proper.

Test T3. The existence of ψ, $\psi \in \mathcal{S}(\text{OSS})$, $\psi \notin \mathcal{S}(\text{NC'OPC})$, implies OPC is proper.

Test T4. The existence of ψ, $\psi \in \mathcal{S}(\text{OSS})$, $\psi \notin \mathcal{S}(\text{NCSSOPC})$, implies OPC is proper.

Tests T1 and T2 arise directly from the definitions of proper and steady-state. Since T1 requires the determination of an element of $\mathcal{S}(\text{SSOPC})$, it is the most difficult test to apply in practice. Usually, it involves inequalities which make use of particular structures in the problem data as in Example 5.1. Test T2 is easier to apply since it only requires exhibiting an admissible time-dependent triple $(\hat{u}(\cdot), \hat{x}(\cdot), \hat{\tau})$ which has lower cost than any element of $\mathcal{S}(\text{OSS})$. See Example 5.5. General tests which implement T2 have been based on sinusoidal perturbations from an optimum steady-state solution [5], [12] and relaxed controls (see [1] and T8 of the next section). From Remark 6.1 it is seen that T2, T3 and T4 are successively weaker tests. Under A3 Remark 6.1 shows that T3 and T4 are equivalent; however, when A3 is not satisfied T3 may be a stronger test than T4 (Remark 4.2). Test T4 is stronger than tests of a similar type which have appeared previously [1], [13] in that it applies to a very general OPC problem and does not require $f_x(x(0), u(0))$, $(x(\cdot), u(\cdot), \tau) \in \mathcal{S}(\text{OSS})$, to be nonsingular. The following theorem shows that T1, T2, T3 and T4 are not vacuous and that there exist no other tests in a reasonable class of tests.

THEOREM 6.1. *Suppose OPC satisfies no special assumptions* {A3} [A2] (A2 *and* A3). *Then tests* T1, T2, T3 *and* T4 {T1, T2 *and* $T3 = T4$} [$T1$, $T2$ *and* T3] (T1 *and* T2) *are not vacuous* (*always negative*) *or pairwise equivalent* (*one test positive always implies the other test positive*). *Let* $\psi \in \mathcal{S}(\text{SS})$. *In the class of tests which employ an evaluation of* all *five conditions*, $\psi \in \mathcal{S}(A)$ *or* $\psi \notin \mathcal{S}(A)$ *for* $A =$ *OSS, NCSSOPC, NCOSS, OPC, NC'OPC, there exist no tests other than* T1, T2, T3 *and* T4 {T1, T2 *and* T3 = T4} [T1, T2 *and* T3] (T1 *and* T2) *which can show that OPC is proper or steady-state.*

Proof. First suppose that OPC satisfies no special assumption. Attach to $\psi \in \mathcal{S}(\text{SS})$ the designation $h(\psi)$ where $h(\psi) = (h_{\text{OSS}}, h_{\text{NCSSOPC}}, h_{\text{NCOSS}}, h_{\text{OPC}}, h_{\text{NC'OPC}})$ is a five digit binary number such that $h_A = 1$ if $\psi \in \mathcal{S}(A)$ and $h_A = 0$ if $\psi \notin \mathcal{S}(A)$. From Theorem 4.2 and Remark 6.1 it follows that 24 of the 32 possible values of $h(\psi)$ are excluded. The remaining eight with examples taken from Table 1 (where $\mathcal{S}(\text{NC'OPC}) = \mathcal{S}(\text{NCSSOPC})$) and § 4 are: (1, 1, 1, 1, 1)—Example 5.4 with $\psi \sim$ a or b; (1, 0, 1, 0, 0)—Example 5.2 with $\psi \sim$ a; (1, 1, 1, 0, 0)—Example 4.2 with $\psi \sim u(t) \equiv x(t) \equiv -1$; (1, 1, 1, 0, 1)—Example 5.5 with $\psi \sim$ a or b; (0, 1, 1, 0, 1)—Example 5.5 with $\psi \sim$ c and Example 5.7 with $\psi \sim$ c; (0, 1, 1, 0, 0)—Example 4.1 with $\psi \sim u(t) \equiv x(t) \equiv 1$ and Example 4.2 with $\psi \sim u(t) \approx x(t) \equiv 1$; (0, 0, 1, 0, 0)—Example 5.4 with $\psi \sim$ c and Example 5.6 with $\psi \sim$ c; (0, 0, 0, 0, 0)—Example 5.4 with $\psi \sim u(t) \equiv \frac{1}{2}$, $x(t) \equiv \frac{9}{16}$ and Example 5.6 with $\psi \sim u(t) \equiv x_1(t) \equiv \frac{1}{2}$, $x_2(t) \equiv 0$. The first result of the theorem follows because: (1, 1, 1, 1, 1) implies T1 positive; (1, 0, 1, 0, 0) implies T2, T3, T4 positive; (1, 1, 1, 0, 0) implies T2, T3 positive; (1, 1, 1, 0, 1) implies T2 positive. For each of the four remaining values of $h(\psi)$ there are examples of OPC which are both proper and steady-state. This is a consequence of Table 1, Example 4.2 being proper (see Remark 4.2) and Example 4.1 being steady-state (to show this

requires an investigation of the solutions of (3.2) and an application of an existence theorem to OPC). Thus there are no additional tests for proper or steady-state. Now consider A3. Since $\mathcal{S}(\text{NC'OPC}) = \mathcal{S}(\text{NCSSOPC})$, $h(\psi) = (1, 1, 1, 0, 0)$ and $h(\psi) = (0, 1, 1, 0, 0)$ are impossible. The remaining examples apply as before. Under A2 Theorem 4.2 gives $\mathcal{S}(\text{OSS}) \subset \mathcal{S}(\text{NCSSOPC})$ and this eliminates $h(\psi) = (1, 0, 1, 0, 0)$. All of the above stated examples except Example 5.2 satisfy A2 and thus the results for A2 are obtained. When A2 and A3 both hold, the argument is essentially a combination of the previous two arguments.

The preceding results and Fig. 2 suggest how a search for solutions of OPC might proceed. Since the determination of elements of $\mathcal{S}(\text{NCOPC})$ requires the solution of the difficult two-point-boundary-value problem (3.2), it is worthwhile to see what can be learned by trying triples $\psi = (u(\cdot), x(\cdot), \tau) \in \mathcal{S}(\text{SS})$. If there is some reason to believe that OPC is proper, it is useful to have tests which indicate how to begin a search for time-dependent controls. For $\psi \in \mathcal{S}(\text{SS})$ conditions which may be checked (listed in order of increasing difficulty) include: $\psi \in \mathcal{S}(\text{NCOSS})$, the system (3.5); $\psi \in \mathcal{S}(\text{WNCSSOPC})$, for some $\varepsilon > 0$ the system (4.4), (3.5-2)–(3.5-4); $\psi \in \mathcal{S}(\text{NCSSOPC})$, the system (3.5-1)', (3.5-2)–(3.5-4); $\psi \in \mathcal{S}(\text{NC'OPC})$, the system (3.2-1)–(3.2-4) for all $\tau \in [0, T]$. The test $\psi \in \mathcal{S}(\text{NCOSS})$ has little value, except perhaps to narrow the search. If elements $\psi \in \mathcal{S}(\text{OSS})$ are known, T3 and T4 may be applied. While there may be fewer elements ψ that satisfy $\psi \in \mathcal{S}(\text{OSS})$, $\psi \notin \mathcal{S}(\text{WNCSSOPC})$ than T4, this test provides somewhat greater information than T4. In particular, reference to Fig. 2 shows $\psi \in \mathcal{S}(\text{WRMOSS})$ and $\psi \notin \mathcal{S}(\text{WRMSSOPC})$. This gives

Test T5. The existence of ψ, $\psi \in \mathcal{S}(\text{OSS})$, $\psi \notin \mathcal{S}(\text{WNCSSOPC})$, implies OPC is locally proper [5], i.e., OPS is proper *and* for all $\varepsilon > 0$ there exists a time-dependent admissible triple $(\hat{u}(\cdot), \hat{x}(\cdot), \hat{\tau})$ with $\|\hat{u}(t) - u(0)\|, \|\hat{x}(t) - x(0)\| < \varepsilon$ for all $t \in [0, T]$ which has lower cost than ψ.

Thus if T5 is positive the search for better time-dependent controls may begin with a guarantee of success in the neighborhood of $(u(0), x(0))$. If T4 is positive $\psi \in \mathcal{S}(\text{SRMOSS})$ and $\psi \notin \mathcal{S}(\text{SRMSSOPC})$. Thus there exist time-dependent admissible triples $(\hat{u}(\cdot), \hat{x}(\cdot), \hat{\tau})$ with $\hat{x}(t)$ in the neighborhood of $x(0)$ which reduce the cost, but large variations, $\hat{u}(t) - u(0)$, may be necessary. If $\psi \in \mathcal{S}(\text{OSS})$ and $\psi \in \mathcal{S}(\text{NC'OPC})$ ($\psi \in \mathcal{S}(\text{NCSSOPC})$ under A3), ψ is a likely candidate for $\mathcal{S}(\text{SSOPC})$. Since $\psi \in \mathcal{S}(\text{NCOPC})$, Theorem 3.1 can reject ψ only if other (time-dependent) solutions of (3.2) are found which have lower cost. However, since it is not known that $\psi \in \mathcal{S}(\text{WRMSSOPC})$, a search for better time-dependent controls might prove successful in the neighborhood of $(u(0), x(0))$.

If it is not possible to determine elements of $\mathcal{S}(\text{OSS})$ much less can be said. Figure 2 suggests several conditions for optimality including $\psi \in \mathcal{S}(\text{NCSSOPC}) \cap \mathcal{S}(\text{SRMOSS})$ and $\psi \in \mathcal{S}(\text{WNCSSOPC}) \cap \mathcal{S}(\text{WRMOSS})$. Checking $\psi \in \mathcal{S}(\text{SRMOSS})$ and $\psi \in \mathcal{S}(\text{WRMOSS})$ may be difficult. Since $\mathcal{S}(\text{NCSSOPC}) \subset \mathcal{S}(\text{WNCSSOPC}) \subset \mathcal{S}(\text{NCOSS})$ necessary conditions for elements of $\mathcal{S}(\text{SRMOSS})$ and $\mathcal{S}(\text{WRMOSS})$ are of value only if they are stronger than (3.5). Obvious candidates for such conditions are second order necessary conditions [9], [15]. Adjoining second order necessary conditions for OSS to the condition $\psi \in \mathcal{S}(\text{NCSSOPC})$ can produce a stronger necessary condition for

elements of $\mathcal{S}$(SSOPC) than $\psi \in \mathcal{S}$(NCSSOPC). This happens in Example 5.1 where elements of $\mathcal{S}$(NCSSOPC) corresponding to c are eliminated.

Finally, it should be observed that the following simple tests, evident from Fig. 2, may be useful.

Test T6. The existence of ψ, $\psi \in \mathcal{S}$(SS), $\psi \notin \mathcal{S}$(NCSSOPC), implies that for all $\varepsilon > 0$ there exists an admissible triple $(\hat{u}(\cdot), \hat{x}(\cdot), \hat{\tau})$, possibly in $\mathcal{S}$(SS), with $\|x(t) - x(0)\| < \varepsilon$ for all $t \in [0, T]$ which has lower cost than ψ.

Test T7. The existence of ψ, $\psi \in \mathcal{S}$(SS), $\psi \notin \mathcal{S}$(WNCSSOPC), implies that for all $\varepsilon > 0$ there exists an admissible triple $(\hat{u}(\cdot), \hat{x}(\cdot), \hat{\tau})$, possibly in $\mathcal{S}$(SS), with $\|\hat{u}(t) - u(0)\|$, $\|\hat{x}(t) - x(0)\| < \varepsilon$ for all $t \in [0, T]$ which has lower cost than ψ.

Remark 6.2. The importance of the assumption A2 is clear. If A2 is satisfied T4 and T5 are vacuous. Moreover, $\psi \in \mathcal{S}$(SRMOSS) and $\psi \in \mathcal{S}$(OSS) are stronger necessary conditions for $\psi \in \mathcal{S}$(SSOPC) than $\psi \in \mathcal{S}$(NCSSOPC). Under A4 $\psi \in \mathcal{S}$(WRMOSS) is a stronger necessary condition than $\psi \in \mathcal{S}$(WNCSSOPC). Tests T1, T2, T3, T6 and T7 remain useful.

7. Relaxed steady-state optima. The replacement of an original optimal control problem by a relaxed optimal control problem is a well established technique in the application of existence theory [4], [21]. In the treatment of optimal periodic control problems it has been recognized [1], [3], [11], [13] that the replacement has an additional function. Steady-state analysis of the relaxed problem, which is relatively easy to carry out, may shed light on the dynamic behavior of the original problem. This path is pursued here; a principal objective is to extend the results of [1].

To introduce the relaxed problem let

$$w = (\rho^1, \cdots, \rho^{l+n+1}, \mu^1, \cdots, \mu^{l+n+1}) \in W \tag{7.1}$$

where

$$W = \left\{ w: \sum_{i=1}^{l+n+1} \rho^i = 1 \text{ and } \rho^i \geqq 0,\ \mu^i \in U \text{ for } i = 1, \cdots, l+n+1 \right\} \subset R^{(l+n+1)(m+1)}. \tag{7.2}$$

Define $f^r: X \times W \to R^n$ and $\tilde{f}^r: X \times W \to R^l$ by

$$f^r(x, w) = \sum_{i=1}^{l+n+1} \rho^i f(x, \mu^i), \tag{7.3}$$

$$\tilde{f}^r(x, w) = \sum_{i=1}^{l+n+1} \rho^i \tilde{f}(x, \mu^i) \tag{7.4}$$

and let

$$y = \frac{1}{\tau} \int_0^\tau \tilde{f}^r(x(t), w(t))\, dt \in Y, \tag{2.1-4}^r$$

$$\dot{x}(t) = f^r(x(t), w(t)) \quad \text{almost all } t \in [0, T], \quad x(0) = x(\tau), \tag{2.1-5}^r$$

$$w(\cdot) \in \mathcal{W} = \{w(\cdot): w(\cdot) \text{ measurable and essentially bounded on } [0, T] \text{ and } w(t) \in W \text{ for all } t \in [0, T]\}. \tag{2.1-6}^r$$

The system (2.1-1), (2.1-2), (2.1-3), (2.1-4)^r, (2.1-5)^r, (2.1-6)^r, (2.1-7), (2.1-8), which is denoted by (2.1)^r, constitutes the *relaxed* OPC problem. The same substitutions apply with obvious modifications elsewhere, e.g., in the statement of the *relaxed* OSS problem, (2.2)^r. Solution sets for the relaxed problem are defined as before and are denoted by $\mathscr{S}^r(\cdot)$. By the Carathéodory theorem [20], $\hat{f}^r(x, W) = \operatorname{co} \hat{f}(x, U)$. This result and an obvious modification lead to the following conclusions.

Remark 7.1. The relaxed OPC problem satisfies A2 and A4.

Suppose that OPC satisfies A2. Then $\hat{f}^r(x, W) = \hat{f}(x, U)$ and it is possible to show that for every solution of (2.1)^r with cost J there is a solution of (2.1) with cost J. Thus the relaxed problem has no interest when OPC satisfies A2.

DEFINITION 7.1. The sequence $\{(u^q(\cdot), x^q(\cdot), \tau)\}$ is an *approximate solution* of (2.1) with *period* τ and *cost* J if: (i) for all $q > 0$ $(u^q(\cdot), x^q(\cdot), \tau) \in \mathscr{U} \times \mathscr{X} \times (0, T]$, $\dot{x}^q(t) = f(x^q(t), u^q(t))$ for almost all $t \in [0, T]$ and $y^q = (1/\tau)\int_0^\tau \tilde{f}(x^q(t), u^q(t))\, dy \in Y$; (ii) for all $\varepsilon > 0$ there exists an integer $Q(\varepsilon)$ such that for $q > Q(\varepsilon)$

$$g_i(y^q, x^q(0)) \leqq \varepsilon \quad \text{for } i = -j, \cdots, -1, \quad |g_0(y^q, x^q(0)) - J| < \varepsilon,$$

$$|g_i(y^q, x^q(0))| < \varepsilon \quad \text{for } i = 1, \cdots, k \quad \text{and} \quad \|x^q(0) - x^q(\tau)\| < \varepsilon.$$

By suitably adapting well known results [4] the following theorem can be proved.

THEOREM 7.1. *Let* $(w(\cdot), x(\cdot), \tau)$ *satisfy* (2.1)^r. *Then there is an approximate solution of* (2.1) *with period* τ *and cost* J.

Since for every $(u(\cdot), x(\cdot), \tau)$ which satisfies (2.1) there exists a $w(\cdot)$ such that $(w(\cdot), x(\cdot), \tau)$ satisfies (2.1)^r, inf J over (2.1)^r is not greater than inf J over (2.1). The system (2.1)^r is of interest because it may have a solution whose cost is less than can be achieved in (2.1). In such a case the corresponding approximate solution of (2.1) has particular importance. These observations also apply when only steady-state solutions are considered. There may exist elements $\psi^r \in \mathscr{S}^r(\mathrm{SS})$ which have lower cost than the cost of any element $\psi \in \mathscr{S}(\mathrm{SS})$. Elements of $\mathscr{S}^r(\mathrm{SS})$ are relatively easy to determine and lead to approximate solutions of (2.1) which have a particularly simple form: $(w(\cdot), x(\cdot), \tau) \in \mathscr{S}^r(\mathrm{SS})$ implies $(u^q(\cdot), x^q(\cdot), \tau)$ can be constructed as a "chattering" solution [4] in which $x^q(\cdot)$ is approximately constant $(x^q(t) \to x(0)$ for all $t \in [0, T])$ and $u^q(t)$ takes on the value $\mu^i(0)$ on a subset of measure $\rho^i(0)T$. As suggested in § 6 this motivates an additional test for proper. Before stating the test it is necessary to extend the definition of proper to allow for approximate solutions.

DEFINITION 7.2. OPC is *approximately proper* if OSS has a minimum cost J_{OSS}, and there exists an approximate solution of (2.1) with cost J such that $J < J_{\mathrm{OSS}}$.

Test T8. Suppose there exist $\psi \in \mathscr{S}(\mathrm{OSS})$ and $\psi^r \in \mathscr{S}^r(\mathrm{SS})$ such that ψ^r has lower cost than ψ. Then OPC is approximately proper.

The validity of the test is obvious from Theorem 7.1. It can be seen from Example 5.2 that the test is not vacuous (take $\rho^1(t) \equiv \frac{2}{3}$, $\rho^2(t) \equiv \frac{1}{3}$, $\mu^1(t) \equiv 1$, $\mu^2(t) \equiv -2$, $x(t) \equiv 1$ which gives $J = -3$). In fact, it is easy to find examples (in Example 5.2 replace $X = R$ by $X = \{x : x < 1.8\}$) where there exist no ψ such that

T4 is positive and yet T8 is positive. The relationships between T8 and the tests T4, T5, T6 and T7 is clarified by the following theorem.

THEOREM 7.2. *Suppose* $(u(\cdot), x(\cdot), \tau) = \psi \in \mathcal{S}$(SS) *and* $\psi \notin \mathcal{S}$(NCSSOPC) ($\psi \in \mathcal{S}$(SS) *and* $\psi \notin \mathcal{S}$(WNCSSOPC)). *Then there exists* $(w(\cdot), x^r(\cdot), \tau) = \psi^r \in \mathcal{S}^r$(SS) *with lower cost than* ψ. *Furthermore, for any* $\varepsilon > 0$ *it is possible to choose* ψ^r *so that* $\|x^r(0) - x(0)\| < \varepsilon$ ($\|x^r(0) - x(0)\| < \varepsilon$ *and* $\|\mu^i(0) - u(0)\| < \varepsilon$ *for* $i = 1, \cdots, l+n+1$).

Proof. Consider $w(\cdot)$ such that $\rho^1(t) \equiv 1$, $\mu^1(t) \equiv u(0)$. Then $\psi^r = (w(\cdot), x(\cdot), \tau) \in \mathcal{S}^r$(SS). Suppose $\psi^r \in \mathcal{S}^r$(NCSSOPC). Then there exist p, $\tilde{p}$, $\sigma_{-j}, \cdots, \alpha_k$ which satisfy the conditions of Theorem 3.4 with notation appropriately modified to account for the relaxed problem. Since $H^r(x, w, p, \tilde{p}) \geqq H^r(x, v, p, \tilde{p})$ for all $v \in W$ the same inequality holds for all $v = (1, 0, \cdots, \mu^1, 0, \cdots, 0)$ such that $\mu^1 \in U$. This implies that for the same $p, \tilde{p}$, $\alpha_{-j}, \cdots, \alpha_k$, ψ satisfies the conditions of Theorem 3.4, i.e., $\psi \in \mathcal{S}$(NCSSOPC). This is a contradiction and hence $\psi^r \notin \mathcal{S}^r$(NCSSOPC). Now suppose $\psi^r \in \mathcal{S}^r$(OSS). Then because of Remark 7.1 and Theorem 4.2 $\psi^r \in \mathcal{S}^r$(NCSSOPC). Thus by contradiction $\psi^r \notin \mathcal{S}^r$(OSS) and there must exist an element of $\mathcal{S}^r$(SS) with lower cost than ψ^r. The argument still applies if X is replaced by an arbitrarily small neighborhood of $x(0)$. Thus the part of the theorem corresponding to $\psi \notin \mathcal{S}$(NCSSOPC) is proved. For $\psi \in \mathcal{S}$(WNCSSOPC) the argument is the same except U is replaced by $U \cap \{\hat{u} : \|\hat{u} - u\| \leqq \varepsilon\}$ with $\varepsilon > 0$ sufficiently small and parts of Theorem 4.3 are used.

Applying Theorem 7.2 with $\psi \in \mathcal{S}$(OSS) shows that if T4 or T5 are positive there exists a $\psi^r \in \mathcal{S}^r$(SS) such that T8 is positive. Additionally, if OPC is proper then OPC is approximately proper. These facts and the comment before Theorem 7.2 are combined in the following conclusion.

Remark 7.2. T8 is a stronger test for OPC approximately proper than either T4 or T5.

To put this remark in perspective it should be observed that T8 has a weaker consequence than T4 or T5. Specifically, there are examples which show that "OPC is approximately proper" does not imply "OPC is proper."

Example 7.1. $j = 0$, $k = 2$, $n = 1$, $U = [-1, 1] \subset R$, $T > 0$, $f = -x + u$, $\tilde{f}_1 = x^2$, $\tilde{f}^2 = -u^2$, $g_0 = y_2$, $g_1 = y_1$. It is clear that (2.1) is satisfied if and only if $x(t) \equiv u(t) \equiv 0$ and $J = 0$. Thus OPC is steady-state. But $\rho^1(t) \equiv \frac{1}{2}$, $\rho^2(t) \equiv \frac{1}{2}$, $\mu^1(t) = 1$, $\mu^2(t) \equiv -1$, $x(t) \equiv 0$ satisfies $(2.1)^r$ with $J = -1$. Thus Theorem 7.2 implies OPC is approximately proper.

Similarly, Theorem 7.2 establishes a connection between relaxed steady-state solutions and tests T6 and T7. When T6 and T7 are positive there exists a $\psi^r \in \mathcal{S}^r$(SS) with lower cost than ψ. Moreover, ψ^r can be chosen to that the "chattering" approximate solution of (2.1) corresponding to ψ^r satisfies the same closeness requirements as do the regular solutions whose existence is guaranteed by T6 and T7. If it can be determined that $\psi \notin \mathcal{S}$(SRMOSS) (for T6) or $\psi \notin \mathcal{S}$(WRMOSS) (for T7) there is no need to resort to the relaxed problem and approximate solutions; it is clear that there are elements of $\mathcal{S}$(SS) which reduce the costs according to the requirements of T6 or T7. However, relaxed steady-state solutions may produce larger reductions in cost than the regular steady-state solutions.

The main practical value of Theorem 7.2 is that it provides a constructive approach for seeking controls which improve performance when any of the tests T4–T7 is positive. Bailey and Horn [1] make the same observation but with respect to T4 only. Their method of proof is more direct but requires $\psi \in \mathcal{S}(\text{OSS})$ *and* $f_x(x(0), u(0))$ nonsingular. The key to the proof presented here is part (viii) of Theorem 4.2 which is a direct consequence of Theorem 3.3.

Remark 7.2 makes it clear that the solution of the relaxed OSS problem deserves special attention. This is the conclusion of Bailey and Horn. Their sufficient condition I (equivalent to T8 under certain restrictions) is stronger than their sufficient condition II (equivalent to T4 under certain restrictions). Because of Remark 7.1, Remark 6.2 applies to the relaxed OPC problem. Hence, there is a hierarchy of necessary conditions which can be applied to the solution of the relaxed OSS problem: $\mathcal{S}^r(\text{OSS}) \subset \mathcal{S}^r(\text{SRMOSS}) \subset \mathcal{S}^r(\text{NCSSOPC}) \subset \mathcal{S}^r(\text{NCOSS})$. If it is not possible to obtain elements of $\mathcal{S}^r(\text{SRMOSS})$ it may be useful (see below) to combine second order necessary conditions for the relaxed OSS problem with $\psi^r \in \mathcal{S}^r(\text{NCSSOPC})$. Also notice that T4 and T5 are useless when applied to the relaxed OPC problem.

Example 5.2 illustrates some of the points which have been made in the preceding paragraphs. The solution of the relaxed OSS problem is given by ψ^r: $\rho^1 = .3896\cdots$, $\rho^2 = .6103, \cdots$, $\mu^1 = -2$, $\mu^2 = 1.5$, $x = 1.1363\cdots$, $J = -3.5511\cdots$. ψ^r is also the (unique) solution of the relaxed OPC problem. This can be deduced from the application of Theorem 3.1 which yields an (x, p)-phase plane which is the same as Fig. 4 except: on L_1 there is a solution which moves from P_1 toward the origin, on L_2 there are solutions which move from P_2 and P_3 toward P_4, P_4 is an equilibrium solution. Thus $\mathcal{S}^r(\text{NCOPC})$ has two elements corresponding to $x(t) \equiv p(t) \equiv 0$, $J = 0$ and $x(t) \equiv 1.1363\cdots$, $p(t) \equiv 2.8518\cdots$, $J = -3.5511\cdots$. Because the relaxed OPC problem has a solution (an existence theorem can be applied) the second extremal must be optimal. OPC does not have a solution but all chattering solutions corresponding to ψ^r satisfy (2.1-2)–(2.1-8) exactly and as $q \to \infty$ the cost approaches $-3.5511\cdots$. The elements of $\mathcal{S}(\text{OSS})$ labeled "a" are in $\mathcal{S}(\text{WNCCSSOPC})$ but not $\mathcal{S}(\text{NCSSOPC})$. Thus T4 and T6 are positive but T5 and T7 are negative. This is consistent with elements "a" in $\mathcal{S}(\text{WRMSSOPC})$ but not $\mathcal{S}(\text{SRMSSOPC})$. For element "b" T4–T7 are all negative but T8 is positive. $\mathcal{S}^r(\text{NCSSOPC})$ has two elements corresponding to $x = 0$ and $x = 1.1363\cdots$. The first element does not satisfy second order necessary conditions for the relaxed OSS problem.

Other examples illustrate that the relaxed OPC problem need not be steady-state. For instance, Example 5.8(i), which can be shown to be equivalent to the relaxed version of 5.3(i), is proper.

Appendix A. Necessary conditions for a general optimal control problem. Consider the following notation and assumptions: μ and ν are positive integers; $\hat{t} \in R$ is positive; $U \subset R^m$ is an arbitrary set; $\hat{X}, \hat{X}^1, \hat{X}^2 \subset R^{\hat{n}}$ are open sets; for $i = -\mu, \cdots, \nu$ the functions $\theta_i: \hat{X}^1 \times \hat{X}^2 \times (0, \hat{t}) \to R$ are continuously differentiable; the function $\hat{f}: \hat{X} \times U \to R^{\hat{n}}$ is continuous and for each $u \in U$ is continuously differentiable in $\hat{x}$. Let $\hat{f}_{\hat{x}}(\hat{x}, u)$ denote the Jacobian matrix of $\hat{f}(\hat{x}, u)$ with respect

to $\hat{x}$; let $\theta_{i\hat{x}^1}(\hat{x}^1, \hat{x}^2, \tau)$, $\theta_{i\hat{x}^2}(\hat{x}^1, \hat{x}^2, \tau)$ and $\theta_{i\tau}(\hat{x}^1, \hat{x}^2, \tau)$ denote respectively the Jacobian matrices of $\theta_i(\hat{x}^1, \hat{x}^2, \tau)$ with respect to $\hat{x}^1$, $\hat{x}^2$ and τ.

General optimal control problem (GOC). Find $u(\cdot)$, $\hat{x}(\cdot)$ and τ which minimize J subject to

(A.1-1) $J = \theta_0(\hat{x}(0), \hat{x}(\tau), \tau)$,

(A.1-2) $\theta_i(\hat{x}(0), \hat{x}(\tau), \tau) \leqq 0, \qquad i = -\mu, \cdots, -1,$

(A.1-3) $\theta_i(\hat{x}(0), \hat{x}(\tau), \tau) = 0, \qquad i = 1, \cdots, \nu,$

(A.1-4) $\dot{\hat{x}}(t) = \hat{f}(\hat{x}(t), u(t))$ almost all $t \in [0, \hat{t}]$,

(A.1-5) $u(\cdot) \in \mathcal{U} = \{u(\cdot)$: measurable and essentially bounded on $[0, \hat{t}]$, $u(t) \in U$ for all $t \in [0, \hat{t}]\}$,

(A.1-6) $\hat{x}(\cdot) \in \hat{\mathcal{X}} = \{\hat{x}(\cdot)$: $\hat{x}(\cdot)$ absolutely continuous on $[0, \hat{t}]$, $\hat{x}(t) \in \hat{X}$ for all $t \in [0, \hat{t}]$, $\hat{x}(0) \in \hat{X}^1$, $\hat{x}(\tau) \in \hat{X}^2\}$,

(A.1-7) $\tau \in (0, \hat{t})$.

THEOREM A.1 (necessary conditions for GOC). *Let*

$$\hat{H}(\hat{x}, u, \hat{p}) = \hat{p}'\hat{f}(\hat{x}, u) \tag{A.2}$$

where $\hat{p} \in R^{\hat{n}}$. *Let* $(u(\cdot), \hat{x}(\cdot), \tau)$ *solve GOC. Then there exist an absolutely continuous function* $\hat{p}(\cdot)$: $[0, \tau] \to R^{\hat{n}}$ *and real numbers* $\alpha_{-\mu}, \cdots, \alpha_\nu$ *such that the following conditions are satisfied*:

(A.3-1) $\max_{v \in U} H(\hat{x}(t), v, \hat{p}(t)) = H(\hat{x}(t), \hat{u}(t), \hat{p}(t))$ *almost all* $t \in [0, \tau]$,

(A.3-2)
$$\hat{p}'(0) = -\sum_{i=-\mu}^{\nu} \alpha_i \theta_{i\hat{x}^1}(\hat{x}(0), \hat{x}(\tau), \tau),$$
$$\hat{p}'(\tau) = \sum_{i=-\mu}^{\nu} \alpha_i \theta_{i\hat{x}^2}(\hat{x}(0), \hat{x}(\tau), \tau),$$

(A.3-3)
$$\dot{\hat{p}}'(t) = -\hat{p}'(t)\hat{f}_{\hat{x}}(\hat{x}(t), u(t)) \quad \text{almost all } t \in [0, \tau],$$
$$\alpha_i \leqq 0, \qquad i = -\mu, \cdots, 0,$$

(A.3-4)
$$\alpha_i \theta_i(\hat{x}(0), \hat{x}(\tau), \tau) = 0, \qquad i = -\mu, \cdots, -1,$$
$$(\alpha_{-\mu}, \cdots, \alpha_\nu) \neq 0.$$

If $\hat{f}(\hat{x}(t), u(t))$ *is continuous at* $t = \tau$ *the following additional condition is satisfied*:

$$\max_{v \in U} H(\hat{x}(\tau), v, \hat{p}(\tau)) = -\sum_{i=-\mu}^{\nu} \alpha_i \theta_{i\tau}(\hat{x}(0), \hat{x}(\tau), \tau). \tag{A.3-5}$$

Proof. With minor changes in notation the conditions are taken from § 7 of [16], assuming that: τ_1 is fixed, $t_1 = \tau_1 = 0$, the θ_i do not depend on τ_3 and $z(\tau_3)$. The regularity condition (7.3) of [16] is not required. This can be seen by changing the proof in [16] to follow the pattern used in [17].

Appendix B. Necessary conditions for a finite-dimensional optimization problem. Consider the following notation and assumptions: μ and ν are nonnegative integers, $\hat{U} \subset R^m$ and $\hat{X} \subset R^{\hat{n}}$ are open sets, $U \subset \hat{U}$ is an arbitrary set, for $i = -\mu, \cdots, \nu$ the functions $\theta_i: \hat{X} \times \hat{U} \to R$ are continuously differentiable. Let $\theta_{i\hat{x}}(\hat{x}, u)$ and $\theta_{iu}(\hat{x}, u)$ denote respectively the Jacobian matrices of $\theta_i(\hat{x}, u)$ with respect to $\hat{x}$ and u.

Finite-dimensional optimization problem (FDO). Find u and $\hat{x}$ which minimize J subject to

$$J = \theta_0(\hat{x}, u), \tag{B.1-1}$$

$$\theta_i(\hat{x}, u) \leqq 0, \qquad i = -\mu, \cdots, -1, \tag{B.1-2}$$

$$\theta_i(\hat{x}, u) = 0, \qquad i = 1, \cdots, \nu, \tag{B.1-3}$$

$$u \in U, \tag{B.1-4}$$

$$\hat{x} \in \hat{X}. \tag{B.1-5}$$

THEOREM B.1 (necessary conditions for FDO). *Let $C(u, U)$ be a conical approximation to U at $u \in U$. Let $(u, \hat{x})$ solve FDO. Then there exist real numbers $\alpha_{-\mu}, \cdots, \alpha_\nu$ such that the following conditions are satisfied*:

$$\sum_{i=-\mu}^{\nu} \alpha_i \theta_{iu}(\hat{x}, u)\delta u \leqq 0 \quad \text{for all } \delta u \in \operatorname{cl} C(u, U), \tag{B.2-1}$$

$$\sum_{i=-\mu}^{\nu} \alpha_i \theta_{i\hat{x}}(\hat{x}, u) = 0, \tag{B.2-2}$$

$$\alpha_i \leqq 0, \qquad i = -\mu, \cdots, 0,$$

$$\alpha_i \theta_i(\hat{x}, u) = 0, \qquad i = -\mu, \cdots, -1, \tag{B.2-3}$$

$$(\alpha_{-\mu}, \cdots, \alpha_\nu) \neq 0.$$

Proof. Apply Theorem 2.3.12 of [7] letting: the equality constraint correspond to (B.1-3) and $\theta_i(\hat{x}, u) = v_{-i}$, $i = -\mu, \cdots, -1$; $z = (u, \hat{x}, v) \in R^{m+\hat{n}+\mu}$; $\Omega = U \times \hat{X} \times V$ where $V = \{v: v_i \leqq 0, i = 1, \cdots, \mu\}$; $(\alpha_0, \alpha_1, \cdots, \alpha_\nu, \alpha_{-1}, \cdots, \alpha_{-\mu})$ correspond to ψ.

THEOREM B.2 (maximum principle for FDO). *Assume that for all $\hat{x} \in \hat{X}$ the set $\{(\theta_{-\mu}(\hat{x}, u), \cdots, \theta_\nu(\hat{x}, u)): u \in U\}$ is convex. Weaken the differentiability requirements on the θ_i to the following: for $i = -\mu, \cdots, \nu$ the functions θ_i are continuous and for each $u \in U$ continuously differentiable in $\hat{x}$. Then the conditions in Theorem* B.1 *apply with* (B.2-1) *replaced by*

$$\sum_{i=-\mu}^{\nu} \alpha_i \theta_i(\hat{x}, u) = \max_{v \in U} \sum_{i=-\mu}^{\nu} \alpha_i \theta_i(\hat{x}, v). \tag{B.2-1}'$$

Proof. See Theorem 4.6 of [19] and take note of the comment on p. 221. Alternatively, the approach taken in § 4.2 of [7] may be adapted.

Acknowledgment. The author expresses his thanks to George D. Ianculescu whose excitement in periodic control motivated his own interest in the subject.

REFERENCES

[1] J. E. BAILEY AND F. J. M. HORN, *Comparisons between two sufficient conditions for improvement of an optimal steady-state process by periodic operation*, J. Optimization Theory Appl., 17 (1971), pp. 378–384.

[2] J. E. BAILEY, *Necessary conditions for optimality in a general class of non-linear mixed boundary value control problems*, Internat. J. Control, 16 (1972), pp. 311–320.

[3] ———, *Periodic operation of chemical reactors: a review*, Chem. Engrg. Comm., 1 (1973), pp. 111–124.

[4] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.

[5] S. BITTANTI, G. FRONZA AND G. GUARDABASSI, *Periodic control: a frequency domain approach*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 33–38.

[6] ———, *Periodic optimization of linear systems under control power constraints*, Automatica—J. IFAC, 9 (1973), pp. 269–271.

[7] M. CANON, C. CULLUM AND E. POLAK, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill, New York, 1970.

[8] K. S. CHANG, *Necessary and sufficient conditions for optimality*, Periodic Optimization, R. MARZOLLO, ed., vol. 1, Springer-Verlag, New York, 1972, pp. 183–217.

[9] A. FIACCO AND G. MCCORMICK, *Nonlinear Programming*, John Wiley, New York, 1968.

[10] E. G. GILBERT, *Vehicle cruise: improved fuel economy by periodic control*, Automatica—J. IFAC, 12 (1976), pp. 159–166.

[11] G. GUARDABASSI, A. LOCATELLI AND S. RINALDI, *Status of periodic optimization of dynamical systems*, J. Optimization Theory Appl., 14 (1974), pp. 1–20.

[12] G. GUARDABASSI AND N. SCHIAVONI, *Boundary optimal constant control versus periodic operation*, preprint Part 1C-IFAC, 6th Triennial World Congress, Internat. Fed. Automatic Control, Boston, 1975.

[13] F. J. M. HORN AND R. C. LIN, *Periodic processes: a variational approach*, Indust. and Eng. Chem. Proc. Design Dev., 6 (1967), pp. 21–30.

[14] M. MATSUBARA, Y. NICHIMURA AND N. TAKAHASHI, *Optimal periodic control of lumped parameter systems*, J. Optimization Theory Appl., 13 (1974), pp. 13–31.

[15] E. J. MESSERLI AND E. POLAK, *On second order necessary conditions for optimality*, this Journal, 7 (1969), pp. 272–291.

[16] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. II: Applications*, this Journal, 5 (1967), pp. 90–137.

[17] ———, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.

[18] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[19] B. N. PSHENICHNYI, *Necessary Conditions for an Extremum*, M. Dekker, New York, 1971.

[20] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

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

[22] W. A. WOLOVICH, *Linear Multivariable Systems*, Springer-Verlag, New York, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

A "CONJUGATE" INTERIOR PENALTY METHOD FOR CERTAIN CONVEX PROGRAMS*

HISASHI MINE, KATSUHISA OHNO AND MASAO FUKUSHIMA†

Abstract. This paper deals with convex programming problems satisfying certain growth conditions. The "conjugate" interior penalty method proposed in this paper utilizes conjugate convex functions and is based on Fenchel's duality theorem in convex analysis. Conjugate interior penalty functions behave quite mildly and hence avoid the ill-conditioning of ordinary interior penalty methods. Convergence of the method is proved, and the relationship between ordinary and conjugate interior penalty methods is shown.

1. Introduction. Numerous classes of penalty functions have been proposed for solving constrained minimization problems, and the effectiveness of those methods has been verified in the literature. One of the familiar classes is the class of interior penalty methods, which convert a constrained problem into a sequence of unconstrained problems. The convergence properties of these methods have been investigated by Fiacco and McCormick [4]. However, it is inevitable that the interior penalty functions become ill-conditioned near the boundary of the constraint region as the iteration proceeds [7], [8]. It should be noted that this difficulty is encountered even when other penalty methods (e.g., exterior or mixed interior–exterior [4], [7]) are employed.

In this paper, restricting our attention to convex programs, we present a new class of sequential unconstrained optimization methods which we call conjugate interior penalty methods. Under appropriate assumptions they circumvent the ill-conditioning of ordinary penalty methods. The idea is to dualize ordinary interior penalty methods by use of Fenchel's duality theorem [11, § 31]. Specifically, the conjugate interior penalty method involves sequential unconstrained maximizations of conjugate interior penalty functions which approach infinity nowhere. It is shown that maximizing the conjugate interior penalty functions is dual to minimizing the ordinary interior penalty functions.

The concept of conjugate convex (concave) functions, originated by Fenchel and applied to nonlinear programming variously, e.g., [2], [3], [6], [10], [11], plays a central role in this paper. The material from convex analysis used in this paper can be found in Rockafellar [11].

2. Conjugate penalty functions. Consider the following convex programming problem:

$$\text{(P)} \qquad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to} \\ &g_i(x) \geqq 0, \quad i = 1, \ldots, m, \end{aligned}$$

where f and $-g_i$, $i = 1, \ldots, . m$, are everywhere finite convex functions on R^n.

* Received by the editors May 30, 1975, and in revised form September 16, 1976.
† Department of Applied Mathematics and Physics, Kyoto University, Kyoto 606, Japan.

Define

$$H_0 \triangleq \{x \in R^n;\ g_i(x) \geqq 0,\ i = 1, \ldots, m\},$$

$$H \triangleq \{x \in R^n;\ g_i(x) > 0,\ i = 1, \ldots, m\}.$$

Since the g_i's are everywhere finite concave functions, H_0 is closed and convex and H is open and convex. Furthermore, if H is nonempty, then $H = \operatorname{int} H_0$ and $H_0 = \operatorname{cl} H$.

It is assumed that the following conditions are satisfied in problem (P):

C1: f is co-finite, i.e., the epigraph of f contains no nonvertical halflines;

C2: H_0 is compact and H is nonempty.

Define the class I_m of extended-real-valued functions as follows: $G \in I_m$ if

(i) G is a continuous concave function with dom $G = R^m_+$, where R^m_+ is the (strictly) positive orthant in R^m;

(ii) G is nondecreasing, i.e., for $\xi, \eta \in R^m$, $\xi \leqq \eta$ implies $G(\xi) \leqq G(\eta)$.

Note that (i) implies that $G(\xi)$ tends to $-\infty$ if ξ approaches the boundary of R^m_+.

An interior penalty method for solving problem (P) is defined for each $G \in I_m$. In what follows, let $G \in I_m$ be given. Define a function h on R^n by

$$h(x) \triangleq G(g_1(x), \ldots, g_m(x)).$$

Then by the concavity of g_i together with the properties of G, h is a concave function with dom $h = H$. Furthermore, $h(x)$ tends to $-\infty$ as x approaches the boundary of H.

Let

$$u_t \triangleq \inf_{x \in R^n} U_t(x) \tag{1}$$

and

$$S_t \triangleq \{x;\ U_t(x) = u_t\}$$

for all $t \geqq 0$, where

$$U_t \triangleq \begin{cases} f - \gamma(\cdot\,|H_0) & \text{if } t = 0, \\ f - th & \text{if } t > 0, \end{cases}$$

and $\gamma(\cdot\,|H_0)$ is the *concave* indicator function of H_0 defined by

$$\gamma(x|H_0) = \begin{cases} 0 & \text{if } x \in H_0 \\ -\infty & \text{if } x \notin H_0. \end{cases}$$

Clearly, the U_t are convex functions with dom $U_t = H_0$ if $t = 0$, dom $U_t = H$ if $t > 0$, and the S_t are convex subsets of dom U_t. Minimizing U_0 over R^n is equivalent to solving problem (P). The functions U_t with a parameter $t > 0$ are ordinary interior penalty functions for problem (P). As is well known, $\{u_t\}$ and S_t converge to u_0 and S_0, respectively, as t decreases to zero [4], [5], [7], [8].

Now we introduce a family of problems dual to (1) by means of

$$v_t \triangleq \sup_{y \in R^n} V_t(y) \tag{2}$$

and

$$T_t \triangleq \{y;\ V_t(y) = v_t\}$$

for all $t \geqq 0$, where

$$V_t \triangleq \begin{cases} \gamma^*(\cdot\,|H_0) - f^* & \text{if } t = 0, \\ h^*t - f^* & \text{if } t > 0, \end{cases}$$

and * denotes conjugacy [11, pp. 104 and 308]. Right scalar multiplication of h^* is defined as $(h^*t)(y) \triangleq th^*(t^{-1}y)$, $t > 0$. Note that $h^*t = (th)^*$ [11, Thm. 16.1]. Notice our double usage of *, that is, convex conjugates for convex functions and concave conjugates for concave functions. However, this should cause no difficulty, since the distinction is always clear from the context. The functions V_t, $t > 0$, are called the *conjugate interior penalty functions* for problem (P).

LEMMA. *Under conditions* C1 *and* C2, *the* V_t *are everywhere finite and concave for all* $t \geqq 0$. *Furthermore,*

$$V_0(y) = \lim_{t \downarrow 0} V_t(y) \quad \textit{for every } y \in R^n.$$

Proof. By C1, f^* is everywhere finite and convex [11, Cor. 13.3.1]. Since $\operatorname{dom} \gamma(\cdot\,|H_0) = H_0$ and $\operatorname{dom} th = H$ which are nonempty and bounded by C2, $\gamma^*(\cdot\,|H_0)$ and h^*t are everywhere finite and concave [11, Cor. 13.3.1]. Hence, the first part of the lemma follows. The latter half follows from the fact that

$$\begin{aligned} \gamma^*(y|H_0) &= \gamma^*(y|H) = \gamma^*(y|\operatorname{dom} h) \\ &= (h^*0^+)(y) \quad [11, \text{Thm. } 13.3] \\ &= \lim_{t \downarrow 0} (h^*t)(y) \quad [11, \text{Cor. } 8.5.2] \end{aligned}$$

for every $y \in R^n$. This completes the proof. □

The lemma says that as t decreases to zero the conjugate penalty functions $\{V_t\}$ converge pointwise to V_0 which is finite everywhere.

The following theorem demonstrates the relationship between the minima of problems (1) and the maxima of problems (2).

THEOREM 1. *Assume that conditions* C1 *and* C2 *are satisfied. Then* S_t *and* T_t *are nonempty and compact, and*

$$-\infty < u_t = v_t < +\infty$$

for every $t \geqq 0$. *Furthermore, the following* (i), (ii), (iii) *are equivalent for each* $t > 0$, *and* (i), (ii'), (iii') *are equivalent for* $t = 0$:

(i) $x \in S_t$ *and* $y \in T_t$;

(ii) $x \in \partial(h^*t)(y) \cap \partial f^*(y)$; (ii') $x \in \partial\gamma^*(y|H_0) \cap \partial f^*(y)$;

(iii) $y \in \partial f(x) \cap \partial(th)(x)$; (iii') $y \in \partial f(x) \cap \partial\gamma(x|H_0)$.

Proof. Since f is everywhere finite, f^* is co-finite [11, Cor. 13.3.1]. From this and C1, both U_t and V_t are co-finite and, in particular, have no directions of

recession for all $t \geqq 0$. Hence, S_t and T_t are nonempty and compact [11, Thm. 27.1]. The assertion about optimal values follows from Fenchel's duality theorem, for by C2 and the lemma, conditions (a) and (b) in [11, Thm. 31.1] are satisfied. Therefore,

$$-\infty < u_t = \min U_t = \max V_t = v_t < +\infty$$

for all $t \geqq 0$. Finally, we shall prove the equivalences for $t > 0$. A necessary and sufficient condition for (i) to hold is

$$y \in \partial f(x) \quad \text{and} \quad x \in \partial(h^* t)(y) \qquad [11, p. 333].$$

This is equivalent to (ii) and to (iii) by [11, Thm. 23.5]. The equivalences for $t = 0$ follow analogously. This completes the proof. □.

By virtue of Theorem 1, for each t the minimum of U_t can be obtained in terms of the maximum of V_t, and vice versa. Therefore, the two sequences of minimization problems (1) and maximization problems (2) are essentially equivalent, since they are convertible to each other without loss of equality.

The following theorem describes a convergence property enjoyed by maxima of the conjugate interior penalty functions.

THEOREM 2. *Assume that conditions* C1 *and* C2 *are satisfied. Then*

$$v_0 = \lim_{t \downarrow 0} v_t$$

and

$$0 = \lim_{t \downarrow 0} \sup_{z \in T_t} \inf_{y \in T_0} \|z - y\|.$$

Proof. We shall prove the last equality; then the first equality follows from the lemma and the continuity of V_0. Let us assume that there exist a decreasing null sequence $\{t_k\}$ of positive numbers and an $\varepsilon > 0$ such that

$$\exists\, z_k \in T_{t_k} \quad \text{and} \quad \inf_{y \in T_0} \|z_k - y\| > \varepsilon \quad \text{for all } k.$$

Let T^ε be the boundary of the set $T_0 + \varepsilon B$, where B is the unit sphere in R^n. Since T_0 is compact, T^ε is also compact. Choosing y_0 arbitrarily in T_0, let $\tilde{z}_k$ be a point where the line segment joining y_0 and z_k intersects T^ε. Then $\{\tilde{z}_k\}$ has a convergent subsequence by the compactness of T^ε. We assume without loss of generality that $\tilde{z}_k$ converges to $\tilde{z}$. For all k, by the definition

$$V_{t_k}(z_k) \geqq V_{t_k}(y_0)$$

from which we have

$$V_{t_k}(\tilde{z}_k) \geqq V_{t_k}(y_0)$$

by the concavity of V_{t_k}. Taking the limit as $k \to \infty$, we have

$$V_0(\tilde{z}) \geqq V_0(y_0),$$

because, by the lemma and [11, Thm. 10.7], $V_t(y)$ is jointly continuous in y and t. But $\tilde{z} \notin T_0$, which is a contradiction. □

Theorem 2 implies that the point-to-set map T_t is upper semicontinuous (u.s.c.) at $t=0$ [1, p. 109]. In particular, if T_0 is a singleton, say $\{y_0\}$, then every sequence $\{y_k \in T_{t_k}\}$ converges to y_0. Note that the first part of Theorem 2 could also be deduced from the fact that $u_0 = \lim_{t \downarrow 0} u_t$ and $u_t = v_t$ for all $t \geqq 0$.

3. Discussion. The conjugate penalty functions $\{V_t\}$ converge to V_0, which is everywhere finite but in general not everywhere differentiable. It can be shown, however, that the V_t are actually everywhere differentiable for all $t>0$, provided f and $-h$ are strictly convex on their effective domains $\operatorname{dom} f = R^n$ and $\operatorname{dom} h = \operatorname{int}(\operatorname{dom} h)$ [11, Thm. 26.3]. In such cases, first derivative methods may be used in each unconstrained maximization of V_t. Furthermore, for each $t>0$ and $y \in T_t$, necessarily $\nabla f^*(y) = \nabla(h^* t)(y)$. Consequently, from (ii) in Theorem 1, $x \in S_t$ may be written as $x = \nabla f^*(y) = \nabla(h^* t)(y)$. More generally, if either f^* or h^* is differentiable, (ii) in Theorem 1 reduces to either $x = \nabla f^*(y)$ or $x = \nabla(h^* t)(y)$. In such cases the conversion of y into x may be considerably simplified. We note here that $\nabla(h^* t)(y) = \nabla h^*(t^{-1} y)$. When V_t is nodifferentiable, some suitable method for maximizing nondifferentiable functions should be employed (cf. [2]).

In general, it may not be so easy to evaluate the conjugate penalty functions, because the class of functions for which the conjugate has a simple closed form is limited. However, when the functions possess certain structure, the difficulty might be relaxed somewhat be means of various dual operations [11, § 16]. For instance, if h is separable, i.e.,

$$\begin{aligned} U_t(x) &= f(x) - t \sum_{i=1}^{m} G_i(g_i(x)) \\ &= f(x) - t \sum_{i=1}^{m} h_i(x) \quad \text{for } t>0, \end{aligned}$$

where $G_i \in I_1$, $i = 1, \ldots, m$, then V_t may be written as

$$\begin{aligned} V_t(y) &= \Big(t \sum_{i=1}^{m} h_i\Big)^*(y) - f^*(y) \\ &= t \sup\Big\{ \sum_{i=1}^{m} h_i^*(y^i) \Big| \sum_{i=1}^{m} y^i = y/t \Big\} - f^*(y) \quad [11, \text{Thm. } 16.4]. \end{aligned}$$

We should mention that the evaluation of V_t above is sometimes expensive in practical computation, because one needs to solve constrained subproblems. Hòwever, the difficulty can be eliminated provided one is willing to increase the dimentionality of variables. In fact, we have

$$\sup_y V_t(y) = \sup_{y^1, \ldots, y^m} \Big\{ t \sum_{i=1}^{m} h_i^*(y^i) - f^*\Big(t \sum_{i=1}^{m} y^i\Big) \Big\},$$

where the maximization is completely unconstrained in R^{mn}.

To ensure the finiteness of all V_t, especially of V_0, rather strong conditions C1 and C2 have been imposed on problem (P). Those conditions may be relaxed somewhat, while preserving the favorable properties of V_t. One way of doing this is to use the asymptotic properties of problem functions, e.g., the recession

function of f and the recession cone of H_0 (cf. [11, § 8]). However, the everywhere finiteness of V_t may be lost there. Such an approach is developed in [9].[1]

Finally, the rate of convergence of the conjugate penalty method is considered. For ordinary penalty methods, the convergence rate analysis has been well investigated [7]. We assume here that f and f^* are differentiable and that ∇f and ∇f^* are Lipschitz continuous on some neighborhoods of x_0 and y_0, respectively, where $\{x_0\} = S_0$ and $\{y_0\} = T_0$. Then there exist positive scalars M_1 and M_2 such that for every $t > 0$ sufficiently small

$$\|y_t - y_0\| = \|\nabla f(x_t) - \nabla f(x_0)\| \leqq M_1 \|x_t - x_0\|$$

and

$$\|x_t - x_0\| = \|\nabla f^*(y_t) - \nabla f^*(y_0)\| \leqq M_2 \|y_t - y_0\|,$$

where $x_t \in S_t$ and $y_t \in T_t$. Hence, we have

$$\frac{1}{M_1} \leqq \frac{\|x_t - x_0\|}{\|y_t - y_0\|} \leqq M_2$$

for all $t > 0$ sufficiently small. Consequently, we may conclude that $\{y_t\}$ converges to y_0 as fast as $\{x_t\}$ does. For instance, if U_t is the logarithmic interior penalty function, we have

$$\|y_t - y_0\| = O(t).$$

4. Examples.

Example 1. Consider the following one-dimensional problem:

$$\begin{aligned} &\text{minimize} && (x-2)^2 \\ &\text{subject to} && 1-x \geqq 0, \\ & && 1+x \geqq 0. \end{aligned}$$

The logarithmic interior penalty function is of the form

$$\begin{aligned} U_t(x) &= (x-2)^2 - t \log (1-x^2) && \text{if } -1 < x < 1, \\ &= +\infty && \text{otherwise.} \end{aligned}$$

The conjugate penalty function becomes

$$V_t(y) = t - (t^2 + y^2)^{1/2} + t \log \frac{t + (t^2 + y^2)^{1/2}}{2t} - 2y - \frac{y^2}{4}$$

for all $y \in R^n$. Fig. 1 and Fig. 2 illustrate U_t and V_t, respectively, for some values of t. It is seen that V_t is more moderate than U_t for small t. Obviously the optimal solution of the problem is $\bar{x} = 1$ and $f(\bar{x}) = 1$.

[1] One of the referees suggested this. At that time, two of the authors had already completed the work [9].

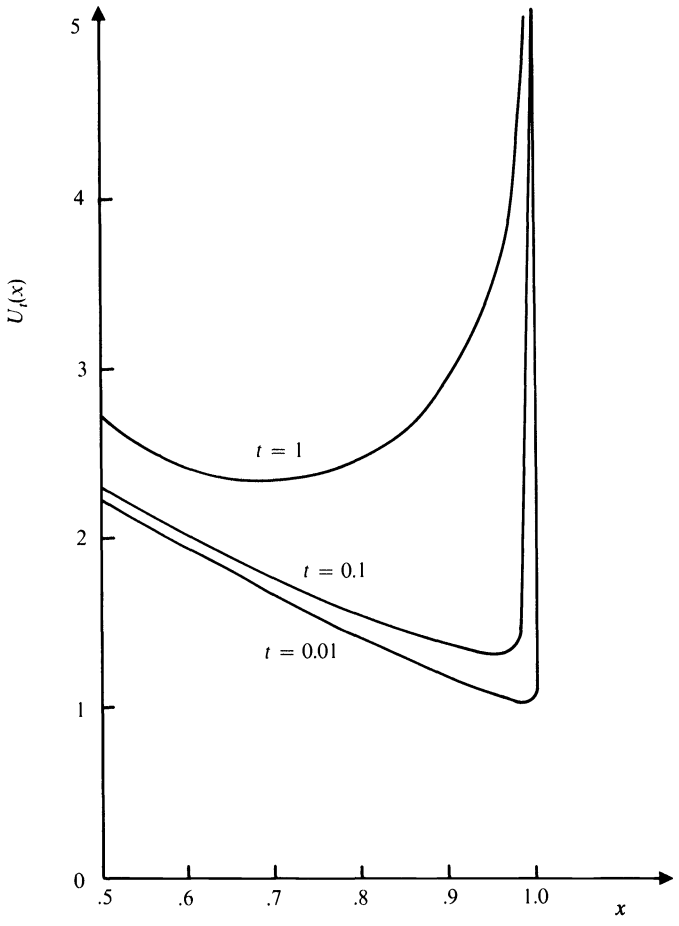


FIG. 1

Example 2. Consider the problem:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}x_1^2+x_2^2-x_1x_2-7x_1-7x_2 \\ &\text{subject to} && 25-4x_1^2-x_2^2 \geqq 0. \end{aligned}$$

As in Example 1, we employ the logarithmic interior penalty function. Then, by the calculation, the conjugate penalty function is

$$\begin{aligned} V_t(y) = t - p_t(y) + t \log \frac{t+p_t(y)}{50t} \\ -y_1^2-\tfrac{1}{2}y_2^2-y_1y_2-21y_1-14y_2-122.5, \end{aligned}$$

where $y = (y_1, y_2)$ and

$$p_t(y) = [t^2+25(y_1^2/4+y_2^2)]^{1/2}.$$

The optimal solution of the problem is $\bar{x} = (2, 3)$ and $f(\bar{x}) = -30$.

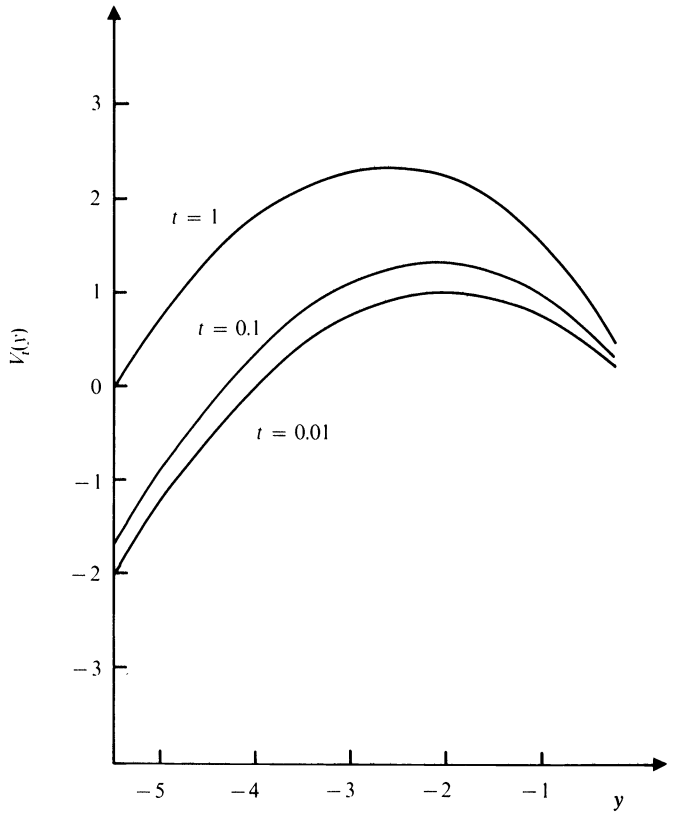


FIG. 2

The computational results for Examples 1 and 2 are presented in Tables 1 and 2, respectively. In each example, the initial point is the origin for $t = 1$, and the subsequent unconstrained maximizations of V_t are initiated from the previous terminating points. The termination criterion is $\|y^{k+1} - y^k\| < 10^{-5}$. The values of x_t are given by $x_t = \nabla f^*(y_t)$.

TABLE 1

t	No. of iterations	y_t	v_t	$x_t = \nabla f^*(y_t)$
1	5	−2.62222	2.36255	0.688892
10^{-1}	3	−2.09326	1.33503	0.953368
10^{-2}	3	−2.00993	1.05610	0.995037
10^{-3}	2	−2.00100	1.00791	0.999500
10^{-4}	2	−2.00010	1.00102	0.999950
10^{-5}	2	−2.00001	1.00013	0.999995
10^{-6}	1	−2.00000	1.00001	0.999999

initial point $y = 0$.

TABLE 2

t	No. of iterations	y_t		v_t	$x_t = \nabla f^*(y_t)$	
1	13	−8.01275	−3.06639	−29.6688	1.90812	2.92087
10^{-1}	9	−8.00139	−3.00667	−29.7388	1.99055	2.99194
10^{-2}	7	−8.00014	−3.00067	−29.9509	1.99904	2.99919
10^{-3}	5	−8.00002	−3.00007	−29.9928	1.99990	2.99991
10^{-4}	3	−8.00000	−3.00001	−29.9990	1.99998	2.99998
10^{-5}	1	−8.00000	−3.00001	−29.9999	1.99998	2.99998

initial point $y = (0, 0)$.

In both examples, V_t is twice differentiable everywhere for all $t > 0$. We have used the *pure* Newton's method for the maximization of each V_t. On the other hand, we would not be able to use the method in each unconstrained minimization of the ordinary interior penalty function, since we should always go outside the feasible region when t is small. In conjugate penalty methods, therefore, we can avoid the considerable effort of determining step sizes to maintain feasibility, as is required by ordinary interior penalty methods.

Finally, in conjugate penalty methods, we need not determine an initial point in the interior of the feasible region, as is required by ordinary interior penalty methods.

Acknowledgment. The authors wish to express their appreciation to one of the referees for his careful review and helpful suggestions.

REFERENCES

[1] C. BERGE, *Topological Spaces*, Oliver & Boyd, Edinburgh and London, 1963.

[2] D. P. BERTSEKAS AND S. K. MITTER, *A descent numerical method for optimization problems with nondifferentiable cost functionals*, this Journal, 11 (1973), pp. 637-652.

[3] J. E. FALK, *Lagrange multipliers and nonconvex programs*, this Journal, 7 (1969), pp. 534–545.

[4] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[5] F. J. GOULD, *A class of inside-out algorithms for general programs*, Management Sci., 16 (1970), pp. 350–356.

[6] B. W. KORT AND D. P. BERTSEKAS, *Combined primal-dual and penalty methods for convex programming*, this Journal, 14 (1976), pp. 268–294.

[7] F. A. LOOTSMA, *A survey of methods for solving constrained minimization problems via unconstrained minimization*, Numerical Methods for Non-Linear Optimization, F. A. Lootsma, ed., Academic Press, London, 1972, pp. 313–347.

[8] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[9] H. MINE AND M. FUKUSHIMA, *Application of Fenchel's duality theorem to penalty methods in convex programming*, Working Paper, Dept. of Appl. Math. and Physics, Kyoto Univ., Kyoto, Japan, 1976.

[10] R. T. ROCKAFELLAR, *Duality in nonlinear programming*, Mathematics of the Decision Sciences, Part 1, G. B. Dantzig and A. F. Veinott, Jr., eds., American Mathematical Society, Providence, RI, 1968, pp. 401–422.

[11] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

THE OPTIMAL CONTROL OF A STOCHASTIC SYSTEM*

ROBERT J. ELLIOTT†

Abstract. The optimal control of a stochastic system with both complete and partial observations is considered. In the completely observable case, because the cost function is, in the terminology of Meyer, a "semimartingale spéciale," a dynamic programming condition for the optimal control is obtained in terms of a certain Hamiltonian. The partially observable case is then discussed from first principles, and it is shown that, almost surely, the optimum control should minimize the conditional expectation of a certain Hamiltonian, with respect to an optimum measure and the observed σ-field.

1. Introduction. In a notable paper [4] M. H. A. Davis and P. P. Varaiya obtained dynamic programming conditions for the optimal control of a stochastic dynamical system using martingale methods. The solutions to the dynamical equations are interpreted by the Girsanov measure transformation method and no complicated existence conditions for stochastic or parabolic equations are required. (A review of previous results in stochastic control is given in the introduction to [4].) Using a different approach, Haussmann [9] applies a basic result of Neustadt [13] on extremals to obtain general necessary conditions for the optimal control of a stochastic system (which may be subject to certain stochastic constraints).

The work below is presented as a simplification of the works of Davis and Varaiya and Haussmann, (though we do not consider state space constraints). After describing the dynamics and cost in § 2 the completely observable case is discussed first in §§ 3 and 4. By observing that the cost function is a "semimartingale spéciale" (see Meyer [12]), an explicit dynamic programming condition is obtained immediately. Unlike [4], we do not have to restrict ourselves to "value decreasing controls" (which are automatically optimal if the cost is terminal), and the absolute continuity of the increasing process in the Doob–Meyer decomposition of the cost is immediate, because special semimartingales have a unique decomposition.

For the partially observable case we again work from first principles, and so do not in particular need Neustadt's results. However, we do need certain very delicate estimates for the L^P norms of the Radon–Nikodym derivatives introduced in the Girsanov measure transformation; our proofs here are adapted from Haussmann [9] and Beneš [1]. Our final result says that the optimal partially observable control should, almost surely, at any time and position minimize the conditional expectation of a certain Hamiltonian, where the conditional expectation is taken with respect to the partially observed σ-field using the optimal measure. This optimality principle is in terms of just one process, and so is an improvement on [4] where corresponding to each control there is a different Hamiltonian. The work in §§ 3 to 8 is based on the hypothesis that there is an optimal control which attains the minimum cost. In § 9 it is shown how this condition can be removed; in fact, although there may not be an optimal control there is always an idealized optimal measure.

* Received by the editors November 4, 1975, and in revised form November 30, 1976.
† Department of Pure Mathematics, University of Hull, Hull, HU5 2DW England.

2. Dynamics. Consider a system whose evolution is described by a stochastic functional differential equation of the form

$$dx_t = f(t, x, u)\,dt + \sigma(t, x)\,dB_t. \tag{2.1}$$

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. Denote a member of $\mathscr{C}$ by x and the value of x at time t by x_t. The drift term f can depend at time t on the past $\{x_s : s \leqq t\}$ of the process; in the Markov case f will depend only on x_t. We shall consider a solution of (2.1) which has an initial value $x_0 \in R^m$ at time 0.

The control u is chosen from a set U, which is a Borel subset of R^l. U is given the Borel σ-field $\mathscr{U}$.

$\mathscr{F}_t$ is the σ-field on $\mathscr{C}$ generated by $\{x_s : x \in \mathscr{C}, s \leqq t\}$.

DEFINITION 2.1. At this stage we suppose the $m \times m$ matrix $\sigma = (\sigma_{ij})$ satisfies

(i) for $1 \leqq i, j \leqq m$, $\sigma_{ij}(t, \cdot)\colon \mathscr{C} \to R$ is $\mathscr{F}_t$ measurable, and $\sigma_{ij}(\cdot, x)\colon [0, 1] \to R$ is Lebesgue measurable for each $x \in \mathscr{C}$.

(ii) $\sigma(t, x)$ is nonsingular.

(iii) each σ_{ij} satisfies a uniform Lipschitz condition in x, where $x \in \mathscr{C}$ is given the uniform norm $\|x\|_s = \sup_{0 \leqq t \leqq s} |x(t)|$.

Suppose an m-dimensional Brownian motion B_t is given and that B_t is defined on a probability space $(\Omega, \mathscr{A}, \mu)$. Then from properties given in Definition 2.1 the equation

$$dx_t = \sigma(t, x)\,dB_t,$$
$$x(0) = x_0 \in R^m,$$

has a unique solution x_t, and there is an induced measure P on its sample space $(\Omega, \mathscr{F}_1)$ given by

$$P(A) = \mu\{\omega : X(\omega) \in A\}; \qquad A \in \mathscr{F}_1.$$

In the partially observable case we suppose that $x_t \in R^m$ is written in terms of two components

$$x_t = (y_t, z_t)$$

where $y_t \in R^n$ and $z_t \in R^{m-n}$. Correspondingly f and σ will be written

$$f = (f_1, f_2), \qquad \sigma = \begin{pmatrix} \sigma_1 \\ \sigma_2 \end{pmatrix},$$

where f_1 (resp. f_2) is an n (resp. $m-n$) dimensional vector function and $\sigma_1(t, x)$ (resp. $\sigma_2(t, x)$) is an $n \times n$ (resp. $(m-n) \times m$) matrix function. The variables y represent the (noisy) observations that are made of the whole process and

$$dy_t = f_1(t, x, u)\,dt + \sigma_1(t, x)\,dB_t.$$

$\mathscr{Y}_t$, the observation σ-field of $\mathscr{F}_t$ is defined by

$$\mathscr{Y}_t = \sigma\{y_s : s \leqq t\}.$$

DEFINITION 2.2. Write Φ for the set of functions $\phi\colon [0, 1] \times \mathscr{C} \to R^m$ which

satisfy

(i) for each $t \in [0, 1]$, $\phi(t, \cdot)$ is $\mathscr{F}_t$ measurable,

(ii) for each $x \in \mathscr{C}$, $\phi(\cdot, x)$ is Lebesgue measurable,

(iii) $|\sigma^{-1}(t, x)\phi(t, x)| \leqq M(1 + \|x\|_t)$.

In the partially observable case we also require

(iv) there is a constant $k_0 < \infty$ such that $\sum \int_0^1 \sigma_{ij}^2 \, dt < k_0$ a.s. P.

Write $a_t = \sigma(t, x)\sigma'(t, x)$, and for $\phi \in \Phi$ define

$$\xi(\phi) = \int_0^1 \phi_t a_t^{-1} \, dx_t - \tfrac{1}{2} \int_0^1 \phi_t a_t^{-1} \phi_t \, dt.$$

A measure P_ϕ is defined on $(\mathscr{C}, \mathscr{F}_1)$ by writing

$$P_\phi A = \int_A \exp(\xi(\phi)) \, dP, \qquad A \in \mathscr{F}_1,$$

and, because $\sigma^{-1}\phi$ has linear growth (see Lemma 0 of [1]), Girsanov's theorem [7] states the following:

THEOREM 2.3.

(i) *P_ϕ is a probability measure on $(\mathscr{C}, \mathscr{F}_1)$.*

(ii) *P_ϕ is mutually absolutely continuous with respect to P.*

(iii) *$\{w_t; t \in [0, 1]\}$ is a Brownian motion under μ_ϕ where*

$$\begin{aligned} dw_t &= dB_t - \sigma^{-1}(t, x)\phi(t, x) \, dt \\ &= \sigma^{-1}(t, x)(dx_t - \phi(t, x) \, dt), \end{aligned}$$

and μ_ϕ is defined on Ω by

$$\mu_\phi(A) = P_\phi(x(A)).$$

DEFINITION 2.4. In the completely observable case an *admissible feedback control over* $[s, t] \in [0, 1]$ is a measurable function

$$u: [s, t] \times \mathscr{C} \to \mathscr{U}$$

such that

(i) for each τ, $s \leqq \tau \leqq t$, $u(\tau, \cdot)$ is $\mathscr{F}_\tau$ measurable.

(ii) for each $x \in \mathscr{C}$, $u(\cdot, x)$ is Lebesgue measurable.

A *partially observable admissible feedback control over* $[s, t]$ is defined similarly, except condition (i) above is replaced by:

(i) (P) for each τ, $s \leqq \tau \leqq t$, $u(\tau, \cdot)$ is $\mathscr{Y}_\tau$ measurable and $E|u(\tau, \cdot)| < \infty$.

Write $\mathscr{M}_s^t$ (resp. $\mathscr{N}_s^t$) for the set of such completely (resp. partially) observable controls. Set $\mathscr{M} = \mathscr{M}_0^1$ and $\mathscr{N} = \mathscr{N}_0^1$, and note that, for example,

$$\mathscr{N} \subset \mathscr{M}.$$

DEFINITION 2.5. We suppose the *drift function* f satisfies:

(i) $f: [0, 1] \times \mathscr{C} \times U \to R^m$ is jointly measurable; and

(ii) for every $u \in \mathscr{M}$,

$$f^u(t, x) = f(t, x, u(t, x)) \in \Phi,$$

(iii) for each $(t, x) \in [0, 1] \times \mathscr{C}$, $f(t, x, \cdot)$ is continuous on U.

Writing P_{f^u} as P_u and μ_{f^u} as μ_u Theorem 2.3 can be re-phrased as follows:

THEOREM 2.6. *Suppose B_t is an m-dimensional Brownian motion on $(\Omega, \mathcal{A}, \mu)$ and, as above, x_t is the solution of $dx_t = \sigma(t, x)\, dB_t$, $x(0) = x_0 \in R^m$. Then under the measure μ_u on Ω, w_t^u is a Brownian motion, where*

$$dw_t^u = \sigma^{-1}(t, x)(dx_t - f^u(t, x)\, dt).$$

That is, from Lemma 6 of [7],

$$dx_t = f(t, x, u(t, x))\, dt + \sigma(t, x)\, dw_t^u,$$

and

$$x(0) = x_0.$$

This result enables us to interpret solutions of the dynamical equations (2.1) under weak conditions on f.

Write

$$\xi_s^t(f_u) = \int_s^t \{\sigma^{-1}(\tau, x) f^u(\tau, x)\}'\, dB_\tau - \frac{1}{2}\int_s^t |\sigma^{-1}(\tau, x) f^u(\tau, x)|^2\, d\tau$$

and

$$\rho_s^t(u) = \exp(\xi_s^t(f^u)).$$

Then (see [1] and [4]) the linear growth condition on $|\sigma^{-1} f^u|$, ensures that

$$E[\rho_s^t(u) | \mathcal{F}_s] = 1 \quad \text{a.s. } P_0.$$

COST 2.7. We suppose the cost associated with the process is of the form

$$g(x(1)) + \int_0^1 c(t, x, u)\, dt,$$

where

(i) g and c are real valued,

(ii) $0 \leqq |g| \leqq k$ and $0 \leqq |c| \leqq k$ for some constant k,

(iii) g is $\mathcal{F}_1$ measurable and for each $u \in \mathcal{M}$ and $t \in [0, 1]$, $c^u(t, \cdot) = c(t, \cdot, u(t, \cdot))$ is $\mathcal{F}_t$ measurable and for each $x \in \mathcal{C}$, $c^u(\cdot, x)$ is Lebesgue measurable,

(iv) for each $(t, x) \in [0, 1] \times \mathcal{C}$, $c(t, x, \cdot)$ is continuous on U.

If E_u denotes the expectation with respect to the measure P_u, then the expected value of the cost corresponding to control $u \in \mathcal{M}$ is

$$J(u) = E_u\left\{g(x(1)) + \int_0^1 c^u(t, x)\, dt\right\}.$$

The optimal control problem is to determine how $u \in \mathcal{M}$ (resp. $u \in \mathcal{N}$) should be chosen so that $J(u)$ is minimized.

3. Completely observable principle of optimality. In the following two sections we consider completely observable systems, that is, at time t the controller has all the information in $\mathcal{F}_t$ and he uses completely observable controls. We quote below a result, the principle of optimality, from the paper of Davis and Varaiya [4].

Suppose $u, v \in \mathcal{M}$; then we can define a control $u \in \mathcal{M}$ by putting

$$w(s,x) = \begin{cases} u(s,x), & 0 \leqq s \leqq t, \\ v(s,x), & t \leqq s \leqq 1. \end{cases}$$

If a control u is used on $[0, t]$ and a control v is used on $(t, 1]$, giving rise to an admissible control $w \in \mathcal{M}$ as above, then the expected remaining cost at time t, given the information $\mathcal{F}_t$, is

$$\psi_{uv}(t) = E_w\left[g(x(1)) + \int_t^1 c_s^v \, ds | \mathcal{F}_t \right]$$

where $c_s^v = c(s, x, v(s,x))$.

Now $\rho_0^t(u)$ is $\mathcal{F}_t$ measurable so

$$\begin{aligned} \psi_{uv}(t) &= \frac{E[\rho_0^t(u)\rho_t^1(v)(\int_t^1 c_s^v \, ds + g(x(1)))|\mathcal{F}_t]}{\rho_0^t(u)} \\ &= E\left[\rho_t^1(v)\left(\int_t^1 c_s^v \, ds + g(x(1)) \right) \middle| \mathcal{F}_t \right] \\ &= E_v\left[\int_t^1 c_s^v \, ds + g(x(1)) | \mathcal{F}_t \right] \quad \text{say.} \end{aligned}$$

Therefore, $\psi_{uv}(t) = \psi_v(t)$ is independent of the control used up to time t. Now $L^1(\mathcal{C}, P_0)$ is a complete lattice under the partial ordering.

$$\psi_1 < \psi_2 \quad \text{if and only if} \quad \psi_1(x) \leqq \psi_2(x) \quad \text{a.s. } P_0,$$

and the set $\{\psi_v(t): v \in \mathcal{M}\}$ is bounded below so the following infimum exists in L^1 for each t:

$$W(t) = \bigwedge_{v \in \mathcal{M}} \psi_v(t).$$

$W(t)$ is, therefore, an $\mathcal{F}_t$ measurable function representing the minimum cost that can be incurred from time t onwards, given the situation at time t.

We now quote the principle of optimality from Theorem 3.1 of [4]:

THEOREM 3.1. *For any $u \in \mathcal{M}$,*

$$W(t) \leqq E_u\left[\int_t^{t+h} c_s^u \, ds | \mathcal{F}_t \right] + E_u[W(t+h)|\mathcal{F}_t].$$

$u^* \in \mathcal{M}$ is optimal if and only if

$$W(t) = E_{u^*}\left[\int_t^{t+h} c_s^{u^*} \, ds | \mathcal{F}_t \right] + E_{u^*}[W(t+h)|\mathcal{F}_t].$$

Immediately we can state the following:

COROLLARY 3.2. (i) $u^* \in \mathcal{M}$ *is optimal if and only if*

$$\int_0^t c_s^{u^*}\, ds + W(t) \quad \text{is a } (\mathcal{F}_t, P_{u^*}) \text{ martingale.}$$

(ii) *In general, for* $u \in \mathcal{M}$,

$$\int_0^t c_s^{u}\, ds + W(t) \quad \text{is a } (\mathcal{F}_t, P_{u}) \text{ submartingale.}$$

Proof.

$$\int_0^t c_s^{u^*}\, ds \quad \left(\text{resp.} \int_0^t c_s^{u}\, ds\right)$$

is $\mathcal{F}_t$ measurable and the results are obtained by adding these quantities to the inequalities of Theorem 3.1.

4. Completely observable minimum principle. Using martingale representation results and the unique decomposition of "semimartingales spéciales" [12] we now exhibit in a very explicit way a Hamiltonian for the control system and an optimality condition in terms of the Hamiltonian, just as in the deterministic case.

From [12] we quote the following definition:

DEFINITION 4.1. An $\mathcal{F}_t$ adapted process S_t is called a *special semimartingale* if it has a representation $S_t = S_0 + M_t + A_t$ where M_t is a local martingale and A_t is a predictable process of locally integrable variation.

Furthermore, it is shown in [12] that the decomposition of a special semimartingale is unique.

LEMMA 4.2. *Suppose* S_t *is a submartingale of class D with a representation*

$$S_t = S_0 + M_t + A_t,$$

where M_t *is a local martingale and* A_t *is a predictable process of locally integrable variation.*

Then M_t *is a martingale and* A_t *is an increasing predictable process.*

Proof. From the Doob-Meyer decomposition result for submartingales of class D (see [11])

$$S_t = S_0 + M_t' + A_t',$$

where M_t' is a martingale and A_t' is an increasing predictable process. By the uniqueness of the decomposition for special semimartingales, therefore, we have

$$M_t = M_t'$$

and

$$A_t = A_t'.$$

THEOREM 4.3. (a) $u^* \in \mathcal{M}$ *is an optimal control if and only if there is a predictable process* $g^*: [0, 1] \times \Omega \to R^m$ *such that* $\int_0^t (g_s^*)^2\, ds < \infty$ *a.s. and*

$$\int_0^t c_s^{u^*}\, ds + W(t) = \int_0^t g^*\, dw^* + W(0)$$

where w^* is a certain Brownian motion on Ω.

(b) *Suppose $u^* \in \mathcal{M}$ is an optimal control.*

Writing

$$f^*(t, x) = f(t, x, u^*(t, x)),$$

$$c^*(t, x) = c(t, x, u^*(t, x))$$

and

$$f^u(t, x) = f(t, x, u(t, x)),$$

$$c^u(t, x) = c(t, x, u(t, x)),$$

we have

$$(g^* \cdot \sigma^{-1} f^* + c^*) = \min_{u \in U} (g^* \cdot \sigma^{-1} f^u + c^u) \quad a.s.$$

That is, the optimum control value $u^*(t, x)$ is almost surely, (Lebesgue $\times \mu$), obtained by minimizing the Hamiltonian

$$g^*(t, x) \cdot \sigma^{-1}(t, x) f^u(t, x) + c^u(t, x).$$

Proof. (a) From Corollary 3.2, $u^* \in \mathcal{M}$ is optimal if and only if

$$\int_0^t c_s^* \, ds + W(t) \quad \text{is a } (\mathcal{F}_t, P_{u^*}) \text{ martingale.}$$

However, by construction

$$B_t = \int_0^t \sigma^{-1} \, dx_t$$

so if $\mathcal{A}_t$ is the σ-field on Ω generated by $\{B_s : s \leqq t\}$ we have that $\mathcal{A}_t = x^{-1}(\mathcal{F}_t)$. Therefore, in terms of the original measure space Ω we can say that

$$\int_0^t c_s^* \, ds + W(t) \quad \text{is an } (\mathcal{A}_t, \mu_{u^*}) \text{ martingale.}$$

However, on $(\mathcal{A}_t, \mu_{u^*})$, w_t^* is a Brownian motion, where

$$dw_t^* = \sigma^{-1}(t, x)(dx_t - f^*(t, x) \, dt).$$

Consequently, by the martingale representation theorem there is a predictable process $g^*: [0, 1] \times \Omega \to R$ such that

$$E_{u^*}\left[\int_0^t (g_s^*)^2 \, ds\right] < \infty$$

and

$$\int_0^t c_s^* \, ds + W(t) = \int_0^t g^* \, dw^* + W(0) \quad \text{a.s.}$$

(b) Consider now a different control $u(t, x) \in \mathcal{M}$ giving rise to a measure P_u on $\mathcal{C}$ and a measure μ_u on Ω. On $(\mathcal{F}_t, P_u)$ we have from Corollary 3.2 that $\int_0^t c_s^u\, ds + W(t)$ is a submartingale. It is, therefore, a submartingale on $(\mathcal{A}_t, \mu_u)$ and so has a unique Doob–Meyer decomposition (see [11, Chap. VII, Thm. 29). That is

$$\int_0^t c_s^u\, ds + W(t) = J^* + M_t^u + A_t^u,$$

where M_t^u is an $(\mathcal{A}_t, \mu_u)$ martingale and A_t^u is a unique predictable increasing process.

However, from the representation in (a),

$$\begin{aligned}\int_0^t c_s^u\, ds + W(t) &= J^* + \int_0^t g^*\, dw^* + \int_0^t (c_s^u - c_s^*)\, ds \\ &= J^* + \int_0^t g^*(\sigma^{-1}\, dx_s - \sigma^{-1} f^u\, ds) \\ &\quad + \int_0^t g^*(\sigma^{-1} f^u - \sigma^{-1} f^*)\, ds + \int_0^t (c_s^u - c_s^*)\, ds \quad \text{a.s. } \mu_u.\end{aligned}$$

Now, by Theorem 2.2, w_t^u is a Brownian motion on $(\mathcal{A}_t, \mu_u)$ where

$$dw_t^u = \sigma^{-1}\, dx - \sigma^{-1} f^u\, dt.$$

Therefore,

$$\int_0^t g^* \sigma^{-1}(dx_s - f_s^u\, ds) = \int_0^t g^*\, dw_t^u$$

is a stochastic integral of this Brownian motion. Because

$$E_{u^*}\left[\int_0^1 (g_s^*)^2\, ds\right] < \infty,$$

$$\int_0^1 (g_s^*)^2\, ds < \infty \quad \text{a.s.}$$

and so $\int_0^t g_s^*\, dw_s^u$ is a local martingale on $(\mathcal{A}_t, \mu_u)$.

Also, by construction

$$\int_0^t ((g^* \sigma^{-1} f^u + c^u) - (g^* \sigma^{-1} f^* + c^*))\, ds$$

is a predictable process. Therefore, applying Lemma 4.2 to the submartingale $\int_0^t c_s^u\, ds + W(t)$, in the Doob–Meyer decomposition we must have

$$M_t^u = \int_0^t g^*\, dw^u \quad \text{a.s. } \mu_u$$

and

$$A_t^u = \int_0^t (g^* \sigma^{-1} f^u + c^u) - (g^* \sigma^{-1} f^* + c^*)\, ds \quad \text{a.s. } \mu_u.$$

Because A_t^u is monotonic increasing we have, almost surely with respect to Lebesgue measure, that

$$g^*\sigma^{-1}f^*+c^*=\min_u (g^*\sigma^{-1}f^u+c^u).$$

Therefore, as in the deterministic case, $g^*\sigma^{-1}f+c$ is the Hamiltonian and the optimum control value is the one that minimizes the Hamiltonian.

Remark 4.4. From the paper of Davis [3] we know that an optimal admissible control u^* exists if for each $(t, x, p)\in[0, 1]\times R^m\times R^m$ there is a $u_0\in U$ such that

$$p\cdot f(t, x, u_0)+c(t, x, u_0)=\inf_{u\in U}(p\cdot f(t, x, u)+c(t, x, u)).$$

This is the case if f and c are continuous in u and U is compact.

5. Partially observable principle of optimality. We now consider a partially observable system, so that at time t the controller has only the information $\mathcal{Y}_t$ and only partially observable controls from $\mathcal{N}$ are used. Suppose control $u\in\mathcal{N}_0^t$ is used to time t and control $v\in\mathcal{N}_t^1$ is used from time t to time 1. Then as in § 3 a control $w\in\mathcal{N}$ can be constructed by concatenation. If v is used from time t to 1, the expected remaining cost at time t, given the information in $\mathcal{Y}_t$ and given that control u has been used to time t is:

$$\tilde{\psi}_{uv}(t)=E_w\Big[g(x(1))+\int_t^1 c_s^v\,ds\Big|\mathcal{Y}_t\Big],$$

where $c_s^v=c(s, x, v(s, x))$. By Loève [10, § 24.4] this is

$$=\frac{E[\rho_0^t(u)\rho_t^1(v)(g(x(1))+\int_t^1 c_s^v\,ds)|\mathcal{Y}_t]}{E[\rho_0^t(u)\rho_t^1(v)|\mathcal{Y}_t]}$$

and, because $E[\rho_t^1(v)|\mathcal{F}_t]=1$, the denominator is $E[\rho_0^t(u)|\mathcal{Y}_t]$. Now, although $\rho_0^t(u)$ is $\mathcal{F}_t$ measurable it is not necessarily $\mathcal{Y}_t$ measurable, so we note that, unlike the completely observable case of § 3, $\tilde{\psi}_{uv}(t)$ is not independent of u. Following Davis and Varaiya [4] write

$$f_{uv}(t)=E\Big[\rho_0^t(u)\rho_t^1(v)\Big(g(x(1))+\int_t^1 c_s^v\,ds\Big)\Big|\mathcal{Y}_t\Big].$$

Again using the partial ordering defined in § 3, $L^1(\mathcal{C}, P_0)$ is a complete lattice, so the infimum

$$V(u, t)=\bigwedge_{v\in\mathcal{N}_t^1} f_{uv}(t)\quad\text{exists.}$$

$E[\rho_0^t(u)|\mathcal{Y}_t]$ does not depend on v so the partially observable expected remaining value function is defined to be

$$W_u(t)=\bigwedge_{v\in\mathcal{N}_t^1} E_{uv}\Big[g(x(1))+\int_t^1 c_s^v\,ds\Big|\mathcal{Y}_t\Big]$$
$$=V(u, t)/E[\rho_0^t(u)|\mathcal{Y}_t].$$

Note that the $\mathcal{Y}_t$-measurable function $W_u(t)$ does depend on the control u used up to time t. Theorem 3.1 of [4] then states the following principle of optimality for the functions $W_u(t)$:

THEOREM 5.1. (i) *$u^* \in \mathcal{N}$ is optimal if and only if for each $t \in [0, 1]$ and $h > 0$,*

$$W_{u^*}(t) = E_{u^*}\left[\int_t^{t+h} c_s^{u^*}\, ds \middle| \mathcal{Y}_t\right] + E_{u^*}[W_{u^*}(t+h)|\mathcal{Y}_t].$$

(ii) *In general, for $u \in \mathcal{N}$,*

$$W_u(t) \leqq E_u\left[\int_t^{t+h} c_s^u\, ds \middle| \mathcal{Y}_t\right] + E_u[W_u(t+h)|\mathcal{Y}_t].$$

For the time being we make the following supposition:

HYPOTHESIS 5.2. There is an optimal control $u^* \in \mathcal{N}$.

We indicate in § 9 how this condition is removed. An immediate consequence of Theorem 5.1(i) is the following result:

COROLLARY 5.3. *Write*

$$\tilde{W}(t) = E_{u^*}\left[g(x(1)) + \int_t^1 c_s^{u^*}\, ds \middle| \mathcal{F}_t\right]$$

so that, because u^* is optimal,

$$W_{u^*}(t) = E_{u^*}[\tilde{W}(t)|\mathcal{Y}_t].$$

Note that $\tilde{W}(t)$ involves only the values of u^* between t and 1. Then

(i) *$u^* \in \mathcal{N}_0^t$ is optimal if and only if $E_{u^*}[N_t^*|\mathcal{Y}_t]$ is a $(\mathcal{C}, \mathcal{Y}_t, P_{u^*})$ martingale, where*

$$N_t^* = \int_0^t c_s^{u^*}\, ds + \tilde{W}(t).$$

(ii) *For general $u \in \mathcal{N}$ and $h \geqq 0$,*

$$E_{u^*}[N_t^u|\mathcal{Y}_t] \leqq E_{u^*}[E_u[N_{t+h}^u|\mathcal{F}_t]\mathcal{Y}_t],$$

where

$$N_t^u = \int_0^t c_s^u\, ds + \tilde{W}(t).$$

Proof. (i) If u^* is optimal, then because $\int_0^t c_s^{u^*}\, ds$ is $\mathcal{F}_t$ measurable,

$$N_t^* = E_{u^*}\left[g(x(1)) + \int_0^1 c_s^{u^*}\, ds \middle| \mathcal{F}_t\right],$$

so N_t^* is a $(\mathcal{C}, \mathcal{F}_t, P_{u^*})$ martingale. It is easily seen that

$$\begin{aligned} E_{u^*}[E_{u^*}[N_{t+h}^*|\mathcal{Y}_{t+h}]\mathcal{Y}_t] &= E_{u^*}[N_t^*|\mathcal{Y}_t] \\ &= E_{u^*}\left[\int_0^t c_s^{u^*}\, ds \middle| \mathcal{Y}_t\right] + W_{u^*}(t), \end{aligned}$$

and the converse follows from Theorem 5.1(i).

(ii) Suppose that control u^* is used to time t, control u is used from time t to

$t+h$ and control u^* is used from time $t+h$ to 1. Then, because u^* is optimal,

$$E_{u^*}[\tilde{W}(t)|\mathcal{Y}_t] = W_{u^*}(t)$$
$$\leqq E_{u^*}\left[E_u\left[\int_t^{t+h} c_s^u\, ds + \tilde{W}(t+h)|\mathcal{F}_t\right]\mathcal{Y}_t\right].$$

In effect, on the right hand side of the above inequality we are considering the nonoptimal control $v \in \mathcal{N}_t^1$, where

$$v(s,x) = \begin{cases} u(s,x), & t < s \leqq t+h, \\ u^*(s,x), & t+h < s \leqq 1. \end{cases}$$

The result follows by adding $E_u[\int_0^t c_s^u\, ds|\mathcal{Y}_t]$ to each side.

6. Uniform boundedness. Suppose $u^* \in \mathcal{N}$ is an optimal control. Then because $N_t^* = \int_0^t c_s^{u^*}\, ds + \tilde{W}(t)$ is a square integrable $\mathcal{F}_t$ martingale there is a predictable process g^* such that $\int_0^1 E^*(g_s^*)^2\, ds < \infty$ and

$$N_t^* = J^* + \int_0^t g^*\, dw^*.$$

Here E^* denotes the expectation with respect to the optimal measure $P_{u^*} = P^*$, J^* is the constant

$$\tilde{W}(0) = W_{u^*}(0)$$

and w^* is the Brownian motion on $(\Omega, \mathcal{A}_t, \mu_{u^*})$ defined by

$$dw_t^* = \sigma^{-1}(t,x)(dx_t - f^{u^*}(t,x)\, dt).$$

The following result is adapted from [9].

LEMMA 6.1. *For any* $u \in \mathcal{M}$,

$$\|x\|_1^2 \leqq K_1\left(1 + \sup_{0 \leqq t \leqq 1}\left|\int_0^t \sigma\, dw^u\right|^2\right),$$

where w^u *is the Brownian motion on* $(\Omega, \mathcal{A}_t, \mu_u)$ *defined by* $dw_t^u = \sigma^{-1}(t,x)(dx_t - f^u(t,x)\, dt)$.

Further, for any $q \geqq 1$,

$$E_u\|x\|_1^q \leqq K(q) < \infty.$$

Proof. Under μ_u,

$$x_t = x_0 + \int_0^t f_s^u\, ds + \int_0^t \sigma(s,x)\, dw_s^u.$$

Now

$$|\sigma^{-1} f_s^u| \leqq M(1 + \|x\|_s)$$

and

$$\sum \int_0^1 \sigma_{ij}^2\, ds < k_0 \quad \text{a.s.}$$

so $|\int_0^t f_s^u\,ds|^2 \leqq k_0 M^2 \int_0^t (1+\|x\|_s)^2\,ds$ and the first result follows by Gronwall's inequality.

For $q>2$,

$$E_u\|x\|_1^q \leqq K_2\Big(1+E_u \sup_t \Big|\int_0^t \sigma\,dw^u\Big|^q\Big)$$

and

$$E_u \sup_t \Big|\int_0^t \sigma\,dw^u\Big| \leqq K_3 \sum_{ij}\Big(\int_0^1 \sigma_{ij}^2\,ds\Big)^{q/2},$$

by § 9 of [2]. The result for $q \geqq 1$ follows by Hölder's inequality.

From [9] we quote the following result. The proof is adapted from Lemma 1 of Beneš [1] by a time-change argument.

LEMMA 6.2. *Suppose w is an n-dimensional Brownian motion on (Ω, μ) and $\sigma=(\sigma_{ij})$: $[0,1]\times\Omega\to R^n\times R^n$ is a matrix process such that*

$$\sum \int_0^1 \sigma_{ij}^2\,dt < k_0 \quad a.s.$$

for some finite constant k_0. Then

$$E\exp\Big\{\lambda \sup_{0\leqq t\leqq 1}\Big|\int_0^t \sigma\,dw\Big|^2\Big\}<\infty$$

if

$$\lambda < (4k_0 n)^{-1}.$$

The proofs of the following delicate estimates are adapted from Lemma 1 of Beneš [1] and Lemma 2.1 of Haussman [9].

LEMMA 6.3. *For u, $u^*\in\mathcal{M}$, and $p>1$ near enough 1, $\rho_0^1(u)\rho_0^1(u^*)^{-1}\in L^p(\mathcal{C}, P^*)$ where P^* is the measure associated with u^*.*

Proof. Using the notation of § 2, we have

$$\begin{aligned}
&E^*(\rho_0^1(u)\rho_0^1(u^*)^{-1})^p \\
&\quad = E\exp\Big(p\int_0^1 \sigma^{-1}f^u - (p-1)\int_0^1 \sigma^{-1}f^{u^*} - \frac{p}{2}\int_0^1 (\sigma^{-1}f^u)^2 \\
&\qquad\qquad + \frac{(p-1)}{2}\int_0^1 (\sigma^{-1}f^{u^*})^2\Big) \\
&\quad = E\exp\Big(\xi_0^1(p\sigma^{-1}f^u - (p-1)\sigma^{-1}f^{u^*}) + \frac{p^2}{2}\int_0^1 (\sigma^{-1}f^u)^2 \\
&\qquad + \frac{(p-1)^2}{2}\int_0^1 (\sigma^{-1}f^{u^*})^2 \\
&\qquad - p(p-1)\int_0^1 (\sigma^{-1}f^u)(\sigma^{-1}f^{u^*}) - \frac{p}{2}\int_0^1 (\sigma^{-1}f^u)^2 \\
&\qquad + \frac{(p-1)}{2}\int_0^1 (\sigma^{-1}f^{u^*})^2 \\
&\quad \leqq E\exp\Big(\xi_0^1(p\sigma^{-1}f^u - (p-1)\sigma^{-1}f^{u^*}) \\
&\qquad \times \exp\big(p(p-1)K_4(1+\|x\|_1^2)\big).
\end{aligned} \tag{6.1}$$

Now under the measure $\bar{\mu}$ defined by $d\bar{\mu}/d\mu = \exp(\xi_0^1(p\sigma^{-1}f^u - (p-1)\sigma^{-1}f^{u^*})$,

$$\int_0^t \sigma_s\, d\bar{w}_s = x(t) - \int_0^t (pf^u - (p-1)f^{u^*})\, ds,$$

where $\bar{w}$ is a Brownian motion. Therefore,

$$|x(t)|^2 \leqq 2\left|\int_0^t \sigma_s\, d\bar{w}_s\right|^2 + 2p^2K_5 \int_0^t (1+\|x\|_s^2)\, ds$$

so by Gronwall's inequality

$$|x(t)|^2 \leqq \left(p^2K_5 + \sup_{0\leqq t\leqq 1}\left|\int_0^t \sigma\, dw\right|^2\right) e^{2p_2K_5}$$

and the expectation (6.1) is of the form

$$\exp K_6(p)\bar{E}\exp\left(K_7(p)p(p-1)\sup_t\left|\int_0^t \sigma_s\, d\bar{w}_s\right|^2\right)$$

where

$$K_7(p) = 2K_4 \exp 2p^2K_5.$$

From Lemma 6.2 this is finite for $p>1$ and p near enough 1.

A result established in almost the same way is the following. Suppose $u(h)$ is a perturbation of the optimal control u^* as in § 5. That is, for some $u \in \mathcal{N}$,

$$u(h)(s,x) = \begin{cases} u^*(s,x) & 0\leqq s\leqq t \quad \text{and} \quad t+h<s\leqq 1, \\ u(s,x) & t<s\leqq t+h. \end{cases}$$

LEMMA 6.4. *For any $p\geqq 1$ there is an $h_p>0$ such that $(\rho_0^1(u(h))\rho_0^1(u^*)^{-1})$ is bounded in $L^p(\mathscr{C}, P^*)$ for $h\leqq h_p$.*

Proof. Because $f^u(h) \equiv f^{u^*}$ except when $t<s\leqq t+h$ a calculation similar to Lemma 6.3 shows that

$$\begin{aligned} &E^*(\rho_0^1(u(h))\rho_0^1(u^*)^{-1})^p \\ &\quad = E^*(\rho_t^{t+h}(u(h))\rho_t^{t+h}(u^*)^{-1})^p \\ &\quad = E\exp(\xi_0^1(pf^{u(h)} - (p-1)f^{u^*})) \\ &\qquad \times \exp\left(hp(p-1)K_7(p)\sup_t\left|\int_0^t \sigma_s\, d\bar{w}_s\right|^2\right). \end{aligned}$$

For any p, choosing h small enough the result follows by Lemma 6.2.

COROLLARY 6.5. *For any $p\geqq 1$ and any $A\in\mathscr{Y}_t$,*

$$E^*(I_A(\rho_t^{t+h}(u(h))\rho_t^{t+h}(u^*)^{-1} - 1)^p) \to 0 \quad \text{as } h\to 0.$$

Proof. Write

$$\phi_t(h) = \rho_t^{t+h}(u(h))\rho_t^{t+h}(u^*)^{-1}.$$

Then

$$E^*(I_A\phi_t(h)^p) \leqq P^*(A)^{1/p'}(E^*\phi_t(h)^{pq'})^{1/q'}.$$

For small enough h $\phi_t(h)^p$ is uniformly integrable for any p. $\phi_t(h)$ converges almost surely to 1 so the result follows.

LEMMA 6.6. *Suppose g^* is the process introduced at the beginning of this section and $u \in \mathcal{N}$. Then*

$$\int_0^1 E_u(g_s^*)\, ds < \infty.$$

Proof. First consider any q, $1 \leqq q < \infty$. By § 9 of [2]

$$E^*\left[\int_0^1 (g_s^*)^2\, ds\right]^q \leqq \text{const. } E^* \sup_{0 \leqq t \leqq 1} \left|\int_0^t g_s^*\, dw_s\right|^{2q}$$

$$= \text{const. } E^* \sup_t \left|E^*\left(g(x(1)) + \int_0^1 c_s^{u^*}\, ds \middle| \mathcal{F}_t\right) - J^*\right|^{2q}$$

$$\leqq \text{const. } (4k)^{2q},$$

where k is the constant of Cost 2.7. Then

$$\int_0^1 E_u(g_s^*)^2\, ds = \int_0^1 E^* \rho_0^1(u) \rho_0^1(u^*)^{-1} (g_s^*)^2\, ds$$

$$\leqq \{E^*(\rho_0^1(u)\rho_0^1(u^*)^{-1})^p\}^{1/p} \left\{E^*\left(\int_0^1 (g_s^*)^2\, ds\right)^q\right\}^{1/q}$$

$$< \infty,$$

for $p > 1$ and p near enough 1 by Lemma 6.3.

7. Differentiability.

Remark 7.1. Many of the technical details below could be avoided if we assumed all processes (f, g^*, u etc.) were right continuous. Difficulties arise because we obtain certain results for almost all $t \in [0, 1]$. However, the corresponding null set in [0, 1] will depend on the control used. (This problem is not resolved in Theorem 4.2 of [4].) Using the metrizability of U and the continuity of f and c in u we show there is a single null set of [0, 1], outside which our results hold for all controls $u \in \mathcal{N}$.

Because the trajectories y are almost surely continuous, for any rational r, $0 \leqq r \leqq 1$, $\mathcal{Y}_r$ is countably generated by the sets

$$\{A_{ir}\}, \quad i = 1, 2, \cdots, \quad \text{say.}$$

Write $\mathcal{G}_t$ for the measurable functions $\{u\}$ from $(\mathcal{C}, \mathcal{Y}_t)$ to $U \subset R^l$ such that $E|u| < \infty$. Note that if $u \in \mathcal{N}$ then $u(t, x) \in \mathcal{G}_t$. Approximating in each coordinate by finite linear combinations with rational coefficients of the characteristic functions of the sets A_{ir} we see, as in Halmos [8, p. 177], that there is a countable subset $\mathcal{H}_r = \{u_{jr}\} \subset \mathcal{G}_r$ such that given $u \in \mathcal{Y}_r$ and $\varepsilon > 0$ there is a u_{jr} such that $E|u_{jr} - u| < \varepsilon$.

Further, for any $t \in [0, 1)$,

$$\mathcal{H}_t = \bigcup_{r \leqq t} \mathcal{H}_r$$

is a countable dense subset of $\mathcal{G}_t$. Note that if $u_{jr} \in \mathcal{H}_r$, then, as a function constant in time, $u_{jr} \in \mathcal{N}_t^{t+h}$ for any $t \geqq r$ and $h > 0$.

For $r \leqq t \leqq s \leqq t+h \leqq 1$ write

$$\phi(s, x, u_{jr}) = g_s^* \sigma^{-1}(f(s, x, u_{jr}) - f(s, x, u^*(s, x)) + (c(s, x, u_{jr}) - c(s, x, u^*(s, x))).$$

Now for each i, j, r the indefinite integral

$$\int_0^t \int_{A_{ir}} \phi(s, x, u_{jr})\, dP^*\, ds$$

has a derivative equal to $\int_{A_{ir}} \phi(t, x, u_{jr})\, dP^*$ for almost all $t \in [0, 1]$. There is, therefore, a set $T_1 \subset [0, 1]$ of zero measure such that for $t \notin T_1$

$$\lim_{h \to 0} \frac{1}{h} \int_t^{t+h} \int_{A_{ir}} \phi(s, x, u_{jr})\, dP^*\, ds = \int_{A_{ir}} \phi(t, x, u_{jr})\, dP^*$$

for all i, j, r.

Furthermore, there is a set of zero measure $T_2 \subset [0, 1]$ such that for $t \notin T_2$,

$$\lim_{h \to 0} \frac{1}{h} \int_t^{t+h} E^*(g_s^*)^2\, ds$$

exists and equals $E^*(g_t^*)^2 < \infty$, as $E^*(g_s^*)^2$ is integrable on $[0, 1]$. Write $T = T_1 \cup T_2$, so T is of zero measure.

LEMMA 7.2. *For all $t \notin T$, all $r \leqq t$ and all i, j,*

$$\lim_{h \to 0} \frac{1}{h} \int_t^{t+h} \int_{A_{ir}} \phi(s, x, u_{jr}) \rho_0^t(u^*) \rho_t^{t+h}(u_{jr})\, dP\, ds = \int_{A_{ir}} \phi(t, x, u_{jr})\, dP^*.$$

Proof. Because $\phi(s, x, u_{jr})$ is $\mathcal{F}_s$ measurable for $s \geqq r$ and

$$E[\rho_0^1(u^*) | \mathcal{F}_{t+h}] = \rho_0^{t+h}(u^*) E[\rho_{t+h}^1(u^*) | \mathcal{F}_{t+h}] = \rho_0^{t+h}(u^*) \quad \text{a.s.},$$

we know that if $t \neq T$,

$$\lim_{h \to 0} \frac{1}{h} \int_t^{t+h} \int_{A_{ir}} \phi(s, x, u_{jr}) \rho_0^t(u^*) \rho_t^{t+h}(u^*)\, dP\, ds = \int_{A_{ir}} \phi(t, x, u_{jr})\, dP^*.$$

Now

$$\left| \frac{1}{h} \int_t^{t+h} \int_{A_{ir}} \phi(s, x, u_{jr}) \rho_0^t(u^*) (\rho_t^{t+h}(u_{ir}) - \rho_t^{t+h}(u^*))\, dP\, ds) \right|$$

$$\leqq \left(\int_{A_{ir}} |\rho_t^{t+h}(u_{jr}) \rho_t^{t+h}(u^*)^{-1} - 1|^p\, dP^* \right)^{1/p}$$

$$\cdot \left(\int_{A_{ir}} \left| \frac{1}{h} \int_t^{t+h} \phi(s, x, u_{jr})\, ds \right|^q dP^* \right)^{1/q}$$

for some $p > 2$ and small enough h. From Corollary 6.5,

$$\left(\int_{A_{jr}} |\rho_t^{t+h}(u_{ir}) \rho_t^{t+h}(u^*)^{-1} - 1|^p\, dP^* \right)^{1/p} \to 0 \quad \text{as } h \to 0.$$

We must show the second term above is bounded as $h \to 0$. Write u for u_{jr}; then

$$\left(\int_{A_{ir}} \left|\frac{1}{h}\int_t^{t+h} \phi(s,x,u)\,ds\right|^q dP^*\right)^{1/q}$$

$$\leqq \left(\frac{1}{h}\int_t^{t+h}\int_{A_{ir}} |g^*\sigma^{-1}(f^u - f^{u^*})|^q\,ds\,dP^*\right)^{1/q}$$
$$+\left(\frac{1}{h}\int_t^{t+h}\int_{A_{ir}} |c^u - c^{u^*}|^q\,ds\,dP^*\right)^{1/q}.$$

Because $|c| \leqq k$ the last term above is less than $2k$ for all h. Now $p>2$ implies $q<2$, so there is a $p_1>1$ such that $p_1 q = 2$. With $1/p_1 + 1/q_1 = 1$ the first term above is less than

$$\left(\frac{1}{h}\int_t^{t+h} E^*(g_s^*)^2\,ds\right)^{1/qp_1}\left(\int_0^1 E^*|\sigma^{-1}(f^u - f^{u^*})|^{qq_1}\,ds\right)^{1/qq_1}.$$

For $t \notin T$ both terms are bounded as $h \to 0$. The result follows.

8. Partially observable minimum principle.

LEMMA 8.1. *For any $t \in [0,1]$, any $A \in \mathcal{Y}_t$ and any $u \in \mathcal{N}$,*

$$\int_t^{t+h}\int_A \phi_s^u \rho_0^t(u^*)\rho_t^{t+h}(u)\,dP\,ds \geqq 0,$$

where, as above, $\phi_s^u = g_s^*\sigma^{-1}(f_s^u - f_s^{u^*}) + (c_s^u - c_s^{u^*})$.

Proof. We have represented

$$N_t^* = \int_0^t c_s^{u^*}\,ds + \tilde{W}(t)$$

as

$$J^* + \int_0^t g_s^*\,dw_s^*$$

for a predictable process g^* such that

$$\int_0^1 E^*(g_s^*)^2\,ds < \infty.$$

As in § 2, w^* is the Brownian motion on (Ω, μ_{u^*}) given by

$$dw^* = \sigma^{-1}(t,x)(dx_t - f^{u^*}(t,x)\,dt).$$

Now

$$N_t^u = \int_0^t c_s^u\,ds + \tilde{W}(t)$$

$$= J^* + \int_0^t g_s^*\,dw_s^u + \int_0^t \phi_s^u\,ds$$

where w^u is the Brownian motion on (Ω, μ_u) given by

$$dw^u = \sigma^{-1}(t,x)(dx_t - f^u(t,x)\,dt).$$

Because

$$\int_0^1 E_u(g_s^*)^2\,ds<\infty,$$

from Lemma 6.6, $\int_0^t g_s^*\,dw_s^u$ is a square integrable $\mathcal{F}_t$ martingale, so

$$E_u[N_{t+h}^u|\mathcal{F}_t]=N_t^u+E_u\left[\int_t^{t+h}\phi_s^u\,ds|\mathcal{F}_t\right].$$

Substituting in part (ii) of Corollary 5.3 we see that

$$E_{u^*}\left[E_u\left[\int_t^{t+h}\phi_s^u\,ds|\mathcal{F}_t\right]\mathcal{Y}_t\right]\geqq 0.$$

Now

$$E_u\left[\int_t^{t+h}\phi_s^u\,ds|\mathcal{F}_t\right]=\frac{E[\rho_0^t(u)\rho_t^{t+h}(u)\rho_{t+h}^1\int_t^{t+h}\phi_s^u\,ds|\mathcal{F}_t]}{E[\rho_0^t(u)\rho_t^{t+h}(u)|\mathcal{F}_t]}.$$

Because $\rho_0^t(u)$ is $\mathcal{F}_t$ measurable this is

$$=E\left[\rho_t^{t+h}(u)\int_t^{t+h}\phi_s^u\,ds|\mathcal{F}_t\right].$$

Therefore,

$$E_{u^*}\left[E_u\left[\int_t^{t+h}\phi_s^u\,ds|\mathcal{F}_t\right]\mathcal{Y}_t\right]=\frac{E[\rho_0^t(u^*)E[\rho_t^{t+h}(u)\int_t^{t+h}\phi_s^u\,ds|\mathcal{F}_t]\mathcal{Y}_t]}{E[\rho_0^t(u^*)|\mathcal{Y}_t]}$$

$$\geqq 0.$$

So almost surely

$$E\left[\rho_0^t(u^*)\rho_t^{t+h}(u)\int_t^{t+h}\phi_s^u\,ds|\mathcal{Y}_t\right]\geqq 0.$$

Therefore, for any $A\in\mathcal{Y}_t$,

$$\int_t^{t+h}\int_A\rho_0^t(u^*)\rho_t^{t+h}(u)\phi_s^u\,dP\,ds\geqq 0.$$

LEMMA 8.2. *For any $t\notin T$, and rational number $r\leqq t$ and any i, j,*

$$\int_{A_{ir}}\phi_t^u\,dP^*\geqq 0,\quad \textit{where } u=u_{jr}\in\mathcal{H}_r.$$

Proof. We have noted that for $r\leqq t$, u_{jr} can be considered as an element of $\mathcal{N}_t^{t+h}$. A perturbation $u\in\mathcal{N}$ of the optimal control u^* is defined by setting

$$u(s,x)=\begin{cases}u^*(s,x), & 0\leqq s\leqq t,\quad\text{and}\quad t+h<s\leqq 1,\\ u_{jr}, & t<s\leqq t+h.\end{cases}$$

From Lemma 8.1 we have that

$$\int_t^{t+h}\int_{A_{ir}}\rho_0^t(u^*)\rho_t^{t+h}(u)\phi_s^u\,dP\,ds\geqq 0.$$

Dividing by h and letting $h \to 0$, because $t \notin T$ we have from Lemma 7.2 that

$$\int_{A_{ir}} \phi_t^u \, dP^* \geqq 0.$$

LEMMA 8.3. *For $t \notin T$, $A \in \mathcal{Y}_t$ and $u \in \mathcal{N}$,*

$$\int_A \phi_t^u \, dP^* \geqq 0.$$

Proof. Suppose $u \in \mathcal{N}$, so that $u(t, x) \in \mathcal{G}_t$. We can, therefore, find a sequence $u_{jr} \in \mathcal{H}_t$ such that $\lim_{j \to \infty} E|u_{jr} - u| = 0$. Consequently, there is a subsequence $\{u(k, r)\} \subset \{u_{jr}\}$ such that $\lim_k u(k, r) = u(t, x)$ a.s. and so, because ϕ_t^u is continuous in u, $\lim \phi_t^{u(k,r)} = \phi_t^u$ a.s. By the bounded convergence theorem we can conclude that

$$\int_{A_{ir}} \phi_t^u \, dP^* \geqq 0$$

for all $t \notin T$ and all A_{ir}. Finally, by the monotone class theorem

$$\int_A \phi_t^u \, dP^* \geqq 0 \tag{8.1}$$

for all $t \notin T$ and all $A \in \mathcal{Y}_t$.

We, therefore, conclude with our main result.

THEOREM 8.4. *Suppose $u^* \in \mathcal{N}$ is an optimal control and $u \in \mathcal{N}$. Then there is a set of zero measure $T \subset [0, 1]$ such that if $t \notin T$,*

$$E^*[g^* \sigma^{-1}(f_t^u - f_t^{u^*}) + (c_t^u - c_t^{u^*}) | \mathcal{Y}_t] \geqq 0 \quad a.s.$$

Proof. This inequality is just a restatement of (8.1) above.

Remarks 8.5. Therefore, the optimal partially observable control is the one that minimizes the conditional expectation of the Hamiltonian ϕ_t^u with respect to the optimum measure P^* and $\mathcal{Y}_t$. The expectation can be with respect to the original measure P if the Radon–Nikodym derivative is introduced. The left hand side above then becomes

$$E[\rho_0^t(u^*) \phi_t^u | \mathcal{Y}_t] \geqq 0 \quad \text{a.s.}$$

9. Suboptimal controls. Our results so far depend on the existence of an optimal control $u^* \in \mathcal{M}$, (resp. $\mathcal{N}$). The existence of u^* enables us to represent the optimal cost as a stochastic integral with respect to μ_{u^*}. We now investigate how this condition can be removed.

The optimal control problem we consider is the same as that described in § 2. However, we make the cost completely into a terminal cost by introducing the new state variable x_{m+1} and a new independent Brownian motion w_{m+1} on a probability space (Ω', μ'). x_{m+1} satisfies the stochastic equation

$$dx_{m+1} = c(t, x, u) \, dt + dw_{m+1},$$

$$x(0) = 0.$$

The $(m+1)$ dimensional process (x, x_{m+1}) is defined on the product space $(\Omega^+, \mu^+) = (\Omega' \times \Omega, \mu' \times \mu)$. In fact $(m+1)$ dimensional processes defined on this augmented space Ω^+ will be indicated with a $+$. Therefore, writing $x^+ = (x, x_{m+1})$ we have

$$dx^+ = f_u^+ \, dt + \sigma^+ \, dw^+$$

where

$$f_u^+ = (f(t, x, u), c(t, x, u)),$$

$\sigma^+ = \begin{pmatrix} \sigma & 0 \\ 0 & 1 \end{pmatrix}$ and $w^+ = (w, w_{m+1})$. $\mathscr{C}^+$ denotes the space of continuous functions from $[0, 1]$ to R^{m+1} and $\mathscr{F}_t^+$ the σ-field generated to time t. The cost can be written $g^+(x(1)) = x_{m+1}(1) + g(x(1))$, and P^+ is the probability measure induced on $(\mathscr{C}^+, \mathscr{F}_1^+)$ by μ^+.

Corresponding to a control $u \in \mathscr{M}$ (resp. $\mathscr{N}$) a measure $P_u^+ \ll P^+$ is defined on Ω^+ by putting $dP_u^+/dP^+ = \exp \xi_0^1(f_u^+)$. If E_u^+ denotes expectation with respect to P_u^+ the expected cost corresponding to u is

$$E_u^+[g^+(x(1))] = E_u^+\left[\int_0^1 c(t, x, u(t, x)) \, dt + w_{m+1}^u(1) + g(x(1))\right]$$

$$= E_u\left[\int_0^1 c_t^u \, dt + g(x(1))\right],$$

because c and g are independent of x_{m+1} and $w_{m+1}^u(t)$ is a Brownian motion under μ_u^+, where μ_u^+ is the measure on Ω^+ induced by P_u^+.

Completely observable case. The expected remaining cost from time t given $\mathscr{F}_t^+$ and using completely observable controls is defined as

$$W^+(t) = \bigwedge_{v \in \mathscr{M}_t^1} E_v^+[g^+(x(1)) | \mathscr{F}_t^+].$$

As before, $W^+(t)$ is independent of the control used up to time t. Note that if there is an optimal control $u^* \in \mathscr{N}$,

$$W^+(t) = W(t) + \int_0^t c_s^{u^*} \, ds.$$

However, because of the nature of the optimum measure (see Lemma 9.2 below), this representation is not valid in general.

For $W^+(t)$ the principle of optimality becomes:

THEOREM 9.1. (i) *$u^* \in \mathscr{M}$ is optimal if and only if $W^+(t)$ is a martingale under $\mu_{u^*}^+$.*

(ii) *In general, for $u \in \mathscr{M}$ $W^+(t)$ is a submartingale under μ_u^+.*

Now $W^+(t)$ exists whether or not there is an optimal control $u^* \in \mathscr{M}$. For the completely observable case we shall show first that under quite weak conditions, even though there may not be an optimal control, there is a measure P^* on $\mathscr{C}^+$ under which $W^+(t)$ is a martingale. The proof is adapted from Lemma 5.1 of [4], and we first quote Lemma 3 of [3]:

LEMMA 9.2. *Suppose* Φ^+ *is the analogue of the set of functions introduced in Definition* 2.2, *so that if* $\gamma \in \Phi^+$, $\gamma: [0,1] \times \mathscr{C}^+ \to R^n$, *then, because* γ *has linear growth,* $E[\exp \xi_0^1(\gamma)] = 1$.

Write

$$\mathscr{D} = \{\exp \xi_0^1(\gamma): \gamma \in \Phi^+\}.$$

Then $\mathscr{D}$ *is a weakly compact subset of*

$$L^1(\mathscr{C}^+, \mathscr{F}^+, P^+).$$

Note that if $\phi \in \Phi$ (see Definition 2.2) $(\phi, c) \in \Phi^+$.

THEOREM 9.3. *There is a function* $H \in \Phi^+$ *such that* $(W^+(t), \mathscr{F}_t, P^*)$ *is a martingale. Here* P^* *is defined by* $dP^*/dP^+ = \exp \xi_0^1(H)$ *and* E^* *denotes expectation with respect to* P^*.

Proof. Consider a sequence $\{u_n\} \subset \mathscr{M}$ such that $\psi_{u_n}^+(0) = E_{u_n}^+[g^+(x(1))]$ decreases to $W^+(0) = J^*$. Now $f_{u_n}^+ \in \Phi^+$ so $\rho_0^1(u_n) \in \mathscr{D}$. There is, therefore, a subsequence also denoted by $\{u_n\}$, and $H \in \Phi^+$ such that

$$\rho_0^1(u_n) \to \rho^*$$

weakly in $L^1(P^+)$ where $\rho^* = \exp[\xi_0^1(H)]$. Conditional expectations are continuous mappings so

$$\rho_0^t(u_n) = E^+[\rho_0^1(u_n) | \mathscr{F}_t^+]$$

converges to

$$E^+[\rho^* | \mathscr{F}_t^+] = \exp[\xi_0^t(H)].$$

Writing

$$\rho_* = E^+[\rho^* | \mathscr{F}_{t+h}^+],$$
$$\rho_n = \rho_0^{t+h}(u_n)$$

and

$$\psi_n^+(t) = E_{u_n}^+[g^+(x(1)) | \mathscr{F}_t^+],$$

we have for any set $F \in \mathscr{F}_t^+$,

$$\begin{aligned}
\int_F (W^+(t+h) - W^+(t))\, dP^* &= \int_F (W^+(t+h) - W^+(t))\rho_*\, dP^+ \\
&= \int_F (\rho_* - \rho_n)(W^+(t+h) - W^+(t))\, dP^+ \\
&\quad + \int_F \rho_n(\psi_n^+(t) - W^+(t))\, dP^+ \\
&\quad + \int_F \rho_n(W^+(t+h) - \psi_n^+(t+h))\, dP^+ \\
&\quad + \int_F \rho_n(\psi_n^+(t+h) - \psi_n^+(t))\, dP^+.
\end{aligned}$$

Now the last term is zero because ψ_n^+ is an $\mathscr{F}_t^+$ martingale under $P_{u_n}^+$. Because $W^+(t) \leqq \psi_n^+(t)$ the third term is nonpositive.

Now $W^+(t)$ is a submartingale under each u_n, so

$$W^+(t) \leqq E_{u_n}^+[W^+(t+h)|\mathscr{F}_t^+].$$

Consider $\varepsilon > 0$; then there is n' such that $\psi_{u_n}^+(0) \leqq W^+(0) + \varepsilon$ for $n \geqq n'$. Therefore

$$E_{u_n}^+[\psi_n^+(t)] = \psi_n^+(0) \leqq E_{u_n}^+[W^+(t)] + \varepsilon$$

so as in [4],

$$\int_F \rho_n(\psi_n^+(t) - W^+(t))\, dP \leqq E_{u_n}[\psi_n^+(t) - W^+(t)]$$

$$\leqq \varepsilon \quad \text{if } n \geqq n'.$$

Again, because $(W^+(t+h) - W^+(t))I_F \in L^\infty$ there is n'' such that if $n \geqq n''$

$$\int_F (\rho_* - \rho_n)(W^+(t+h) - W^+(t))\, dP^+ < \varepsilon.$$

Therefore, if $n \geqq \max(n', n'')$,

$$\int_F (W^+(t+h) - W^+(t))\, dP^* \leqq \varepsilon$$

for arbitrary ε, so $(W^+, \mathscr{F}_t^+, P^*)$ is a supermartingale.

However, we know that $(W^+, \mathscr{F}_t^+, P_u^+)$ is a submartingale for any $u \in \mathscr{M}$ so for any $F \in \mathscr{F}_t^+$,

$$\int_F (W^+(t+h) - W^+(t))\rho_0^t(u_n)\, dP^+ \geqq 0.$$

Using weak convergence in $L^1(\mathscr{C}^+, P^+)$ we have that

$$\int_F (W^+(t+h) - W^+(t))\, dP^* \geqq 0$$

so $(W, \mathscr{F}_t, P^*)$ is also a submartingale.

Therefore, although the function $H \in \Phi^+$ may not be of the form f_u^+, the process $(W^+, \mathscr{F}_t^+, P^*)$ is a martingale, where

$$dP^*/dP^+ = \exp \xi_0^1(H).$$

COROLLARY 9.4. *Even though there may not be an optimal control we have the following representations in terms of the function* $H \in \Phi^+$:

Under P^*,

$$W^+(t) = J^* + \int_0^t g^*\, dw^*,$$

where w^* *is the Brownian motion on* (Ω^+, μ^*) *defined by* $dw^* = (\sigma^+)^{-1}(dx^+ - H\, dt)$ *and the measure* μ^* *is the measure on* Ω^+ *induced by* P^*.

Note that in general for this optimal measure we cannot separate out the integral part of the cost, and the integrand g^* is now a predictable $(m+1)$ dimensional stochastic process. The minimum principle has the following form.

THEOREM 9.5. *At any time a position (t, x) the completely observable optimal control should endeavour to minimize the Hamiltonian $g^*(\sigma^+)^{-1}(f^u, c^u)$.*

Proof. For a general control $u \in \mathcal{M}$, $W^+(t)$ is a submartingale and we can obtain its representation under P_u^+ as

$$W^+(t) = J^* + \int_0^t g^* \, dw^u + \int_0^t g^*(\sigma^+)^{-1}((f^u, c^u) - H) \, ds.$$

Again

$$\int_0^t g^*(\sigma^+)^{-1}((f^u, c^u) - H) \, ds$$

is the unique predictable increasing process in the Doob–Meyer decomposition so H is a process such that

$$g^*(\sigma^+)^{-1} H \leqq \inf_u g^*(\sigma^+)^{-1}(f^u, c^u) \quad \text{a.s.}$$

If the Hamiltonian $g^*\sigma^{-1}(f^u, c^u)$ has a minimum for each (t, x), that is if there is a measurable $u^*(t, x)$ such that

(i) $u^*(t, \cdot)$ is $\mathcal{F}_t$ measurable,

(ii) $u^*(\cdot, x)$ is Lebesgue measurable, and

(iii) $g^*(\sigma^+)^{-1}(f^{u^*}, c^{u^*}) \leqq g^*(\sigma^+)^{-1} H$,

then u^* is an optimal control.

(See Davis [3].)

Partially observable case. Let us now turn to the partially observable case. After transferring the cost into a completely terminal cost, the partially observable cost function, given that control $u \in \mathcal{N}$ has been used to time t and given $\mathcal{Y}_t$, is defined to be

$$\bar{W}_u^+(t) = \bigwedge_{v \in \mathcal{N}_t^1} E_{uv}^+[g^+(x(1)) | \mathcal{Y}_t].$$

Consider a sequence $\{u_n\} \subset \mathcal{N}$ such that $\tilde{\psi}_{u_n}^+(0) = E_{u_n}^+[g^+(x(1))]$ decreases to the partially observable minimum cost $\bar{W}^+(0)$. Note that because only controls in $\mathcal{N} \subset \mathcal{M}$ are used $W^+(0) \leqq \bar{W}^+(0)$. Again $f_{u_n}^+ \in \Phi^+$ so $\rho_0^1(u_n) \in \mathcal{D}$. There is, therefore, a subsequence, denoted by $\{u_n\}$, and $H \in \Phi^+$ such that $\rho_0^1(u_n) \to \bar{\rho}_0^1$ weakly in $L^1(P^+)$, where $\bar{\rho} = \exp[\xi_0^1(H)]$. Write $\bar{P}$ for the measure defined by $d\bar{P}/dP = \bar{\rho}_0^1$ and $\bar{E}$ for the expectation with respect to $\bar{P}$. The following analogue of Corollary 5.3(ii) then holds:

LEMMA 9.6. *For any $t \in [0, 1]$, $h \geqq 0$ and $u \in N_t^{t+h}$,*

$$\bar{E}[g^+(x(1)) | \mathcal{Y}_t] \leqq \bar{E}[E_u^+[\bar{E}[g^+(x(1)) | \mathcal{Y}_{t+h}] \mathcal{F}_t] \mathcal{Y}_t].$$

Proof. Suppose the above inequality were not true. Then there would be a

$u \in \mathcal{N}_t^{t+h}$ and a set $A \in \mathcal{Y}_t$ of positive measure such that

$$\int_A g^+(x(1))\bar{\rho}_0^1 \, dP > \int_A g^+(x(1))\bar{\rho}_0^t \rho_t^{t+h}(u)\bar{\rho}_{t+h}^1 \, dP.$$

By considering a modified sequence of $u_n \in \mathcal{N}$, which were all equal to u for $t < s \leqq t+h$ and $x \in A$, we could approximate a quantity strictly smaller than $\bar{W}^+(0)$ using partially observable controls. This contradicts the definition of $\bar{W}^+(0)$, so the result is established.

Remarks 9.7. All the computations of §§ 6, 7 and 8 then go through, with the function $\rho_0^1(u^*)$ replaced by $\bar{\rho}_0^1$. The partially observable minimum principle takes the following form.

THEOREM 9.8. *$\bar{E}[g^+(x(1))|\mathcal{F}_t]$ is a square integrable $\mathcal{F}_t$ martingale, so there is a predictable process $\bar{g}$ such that*

$$\bar{E}[g^+(x(1))|\mathcal{F}_t] = \bar{J} + \int_0^t \bar{g} \, d\bar{w}.$$

Here $\bar{J} = \bar{E}[g^+(x(1))]$, $\bar{w}$ is the Brownian motion on $(\Omega^+, \bar{\mu})$ defined by

$$d\bar{w} = (\sigma^+)^{-1}(dx^+ - H\,dt)$$

and $\bar{\mu}$ is the measure on Ω^+ induced by $\bar{P}$.

At any time t and position x the optimal partially observable control should endeavour to minimize the Hamiltonian

$$\bar{E}[\bar{g}(\sigma^+)^{-1}(f^u, c^u)|\mathcal{Y}_t].$$

Acknowledgment. The author is indebted to Professors Rishel and Varaiya for pointing out an error in an earlier version of this work.

REFERENCES

[1] V. E. BENEŠ, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.
[2] D. L. BURKHOLDER, *Distribution function inequalities for martingales*, Ann. Probability, 1 (1973), pp. 19–42.
[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] C. DELLACHERIE AND P. A. MEYER, *Probabilities et Potential*, 2 eme ed., Hermann, Paris, 1975, Chap. I–IV.
[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*, Theory Probability Appl., 5 (1960), pp. 285–301.
[8] P. R. HALMOS, *Measure Theory*, Van Nostrand, Princeton, NJ, 1950.
[9] U. G. HAUSSMANN, *General necessary conditions for the optimal control of stochastic systems*, Symposium on Stochastic Systems, University of Kentucky, Lexington, 1975.
[10] M. LOÈVE, *Probability Theory*, 3rd ed., Van Nostrand, Princeton, NJ, 1963.
[11] P. A. MEYER, *Probability and Potentials*, Blaisdell, Waltham, MA, 1966.
[12] P. A. MEYER, *Un cours sure les intégrales stochastiques*, Sem. Prob. Univ. Strasbourg 1974–5, Lecture Notes in Math., Vol. 511, Springer-Verlag, Berlin, New York, 1976.
[13] L. W. NEUSTADT, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

CONVERGENCE CONDITIONS FOR A TYPE OF ALGORITHM MODEL*

GERARD G. L. MEYER†

Abstract. This paper is devoted to the study of convergence conditions for the monotone and autonomous iterative algorithm model. Two convergence conditions are presented, and it is shown that (i) they are not comparable and (ii) they contain the known convergence conditions for the model. The presentation of the results is facilitated by the introduction of the concept of extended characteristic set of an iterative procedure.

Introduction. The theory of iterative algorithms has been developed for essentially four purposes: (i) to streamline the analysis of algorithms [1]–[3], [8]–[11]; (ii) to obtain methods for implementing conceptual algorithms [4], [7], [9]; (iii) to synthesize algorithms satisfying a priori given construction constraints; and (iv) to classify algorithms into families having common properties [5], [8].

This paper is devoted to a specific part of the theory of iterative algorithms, namely the study of convergence conditions for the monotone and autonomous iterative algorithm model. Two convergence conditions are presented, and it is shown that (i) they are not comparable and (ii) they contain the known convergence conditions for the model under investigation. The presentation of the results is facilitated by the introduction of the new concept of extended characteristic set. It is proved that the previously known convergence conditions involve the extended characteristic set and that one of the new convergence conditions involves only the characteristic set [6].

The paper's first part contains the model and the definition of its characteristic and extended characteristic sets. The convergence conditions for the model are presented in the second part. The comparison between the new convergence conditions and the classical ones is carried out in the third and last part of the paper.

Algorithm model. The algorithm model under investigation is defined in the most general space compatible with its use. Given a normal topological space $\mathcal{T}$ in which points are closed and a sequentially closed subset T of $\mathcal{T}$, let $A(\cdot)$ be a map from T into all the nonempty subsets of T, and let $c(\cdot)$ be a map from T into the reals E.

1. ALGORITHM. Let z_0 be a given point in T.

Step 0. Set $i = 0$.

Step 1. Pick a point x_i in $A(z_i)$.

Step 2. If $c(x_i) \geqq c(z_i)$, stop; otherwsie, set $z_{i+1} = x_i$, set $i = i + 1$, and go to Step 1.

The algorithm model may generate finite and infinite sequences. Therefore, two types of properties of 1 must be characterized, namely the finite properties of 1 and the asymptotic properties of 1.

* Received by the editors November 4, 1975, and in final revised form November 19, 1976.

† Electrical Engineering Department, Johns Hopkins University, Baltimore, Maryland 21218.

2. DEFINITION. Let F be the set of all last points of all finite sequences generated by 1, and let Q be the set of all cluster points of all infinite sequences generated by 1.

The set F is easily obtained from the maps $A(\cdot)$ and $c(\cdot)$:

$$F=\{z\in T|c(x)\geqq c(z) \quad \text{for at least one } x \text{ in } A(z)\}.$$

On the other hand, the set Q is more elusive. Instead of trying to obtain Q exactly, which may not be possible, one tries to obtain upper bounds for Q.

The sets F and Q are not enough to allow the analysis of 1. One needs a set which in some sense is linked to the asymptotic properties of 1 but which may also be expressed as a function of the maps $A(\cdot)$ and $c(\cdot)$.

3. DEFINITION. The extended characteristic set F_e of 1 is the set of all points z of T such that for each scalar $\delta>0$, there exists at least one point x in $A(z)$, which may depend on z and δ such that

$$c(x)\geqq c(z)-\delta.$$

The complement of a set D with respect to T is denoted by D^c; i.e.,

$$D^c=\{z\in T|z\notin D\}.$$

The set F_e plays an important role in the theory of algorithm models; to remove any misunderstanding, the expression for its complement with respect to T is also given:

$$F_e^c=\{z\in T|c(x)\leqq c(z)-\delta(z), \quad \text{for all } x \text{ in } A(z) \text{ and for some } \delta(z)>0\}.$$

Convergence conditions. The convergence conditions for the model must be stated in a way which allows for variations in their strength. In order to achieve this purpose, a subset D of T is introduced.

4. HYPOTHESIS. If z belongs to D^c, there exist a neighborhood $N(z)$ of z, $\delta(z)>0$, and $\lambda(z)$ such that for all x' in $A(z')$ and for all z' in $N(z)$,

$$\text{(i)} \quad c(x')+\delta(z)\leqq c(z')$$

and

$$\text{(ii)} \quad \lambda(z)\leqq c(z').$$

Note that if the map $c(\cdot)$ is lower semi-continuous on D^c, then given z in D^c there exist $\lambda(z)$ and a neighborhood $N(z)$ of z such that (ii) of Hypothesis 4 is satisfied.

5. HYPOTHESIS. (i) If z belongs to D^c, there exists a neighborhood $N(z)$ of z such that

$$c(x')<c(z)$$

for all x' in $A(z')$ and for all z' in $N(z)$;

(ii) The map $c(\cdot)$ is lower semi-continuous on D^c.

The first set of sufficient conditions (Hypothesis 4) is essentially that given by Polak [9]. Hypothesis 4 generalizes the Polak conditions in that $c(\cdot)$ is not required to be continuous on D^c or to be bounded from below on D^c. The

requirements on $c(\cdot)$, i.e., bounded from below on at least one neighborhood of z for every z in D^c, is therefore a local requirement instead of a global requirement.

6. THEOREM. *If Hypothesis* 4 *is satisfied, then*

$$\text{(i)} \quad D \supseteq F_e$$

and

$$\text{(ii)} \quad D \supseteq Q.$$

Proof. (i) If z is in D^c, there exists a $\delta(z) > 0$ such that

$$c(x) \leqq c(z) - \delta(z)$$

for all x in $A(z)$, and the definition of F_e implies immediately that z is in F_e^c.

(ii) Let z^* be a cluster point of an infinite sequence $\{z_i\}$ generated by 1. There exists K, an infinite subset of the integers, such that the subsequence $\{z_i\}_K$ converges to z^*. Assume that z^* is in D^c and that (i) of 4 is satisfied. Then there exist $N(z^*)$ and $\delta(z^*) > 0$ such that

$$c(x') \leqq c(z') - \delta(z^*)$$

for all x' in $A(z')$ and for all z' in $N(z^*)$. It follows that there exists k such that z_i is in $N(z^*)$ for all $i \geqq k$, i in K; therefore,

$$c(z_{i+1}) \leqq c(z_i) - \delta(z^*)$$

for all $i \geqq k$, i in K. The sequence $\{c(z_i)\}$ is monotonically decreasing; one concludes that the sequence $\{c(z_i)\}$ is unbounded from below.

(iii) If Hypothesis 4 is satisfied and if a point z^* in Q is not in D, then there exists a subsequence $\{z_i\}_K$ converging to z^* such that the subsequence $\{c(z_i)\}_K$ "converges" to $-\infty$. Part (ii) of Hypothesis 4 implies that there exist $\lambda(z^*)$ and $N(z^*)$ such that

$$\lambda(z^*) \leqq c(z')$$

for all z' in $N(z^*)$. It follows that the subsequence $\{c(z_i)\}_K$ is bounded from below, which is a contradiction. One concludes that if Hypothesis 4 is satisfied, then $D \supseteq Q$.

Hypothesis 4 involves the extended characteristic set. If D does not contain F_e, the hypothesis cannot be satisfied. In other words, Hypothesis 4 does not allow upper bounds for Q which are smaller than F_e. This is an important drawback when the set F_e is much larger than the set F. Hypothesis 5 does not involve the set F_e and may therefore be used, whenever applicable, to obtain better upper bounds on Q. Note that if Hypothesis 4 is satisfied with $D = F_e$, then every point z in F_e^c possesses a neighborhood $N(x)$ contained in F_e^c; therefore, F_e must be closed.

7. THEOREM. *If Hypothesis* 5 *is satisfied, then*

$$\text{(i)} \quad D \supseteq F$$

and

$$\text{(ii)} \quad D \supseteq Q.$$

Proof. (i) If z is in D^c, then

$$c(x') < c(z)$$

for all x' in $A(z)$, and the definition of F implies immediately that z is in F^c.

(ii) Let z^* be a cluster point of an infinite sequence $\{z_i\}$ generated by 1. There exists K, an infinite subset of the integers, such that the subsequence $\{z_i\}_K$ converges to z^*. Assume that z^* is in D^c. Part (i) of 5 implies that there exists a neighborhood $N(z^*)$ of z^* such that

$$c(x') < c(z^*)$$

for all x' in $A(z')$ and for all z' in $N(z^*)$. It follows that there exists k such that

$$c(z_{i+1}) < c(z^*)$$

for all $i \geqq k$, i in K. Part (ii) of 5 and the monotonicity of $\{c(z_i)\}$ imply that

$$c(z^*) \leqq c(z_i)$$

for all i. This contradicts the fact just proved, i.e., that

$$c(z_{i+1}) < c(z^*)$$

for all $i \geqq k$, i in K; therefore, z^* cannot be in D^c.

Relations between convergence conditions. This section consists of two parts. The first one shows, with the help of two simple examples, that the conditions of convergence presented in this paper are not comparable: Hypothesis 4 does not imply and is not implied by Hypothesis 5. The second part of this section is devoted to the comparison between Hypotheses 4 and 5 and the classical convergence conditions proposed by Polak, Polyak, and Zangwill. One notes that Hypothesis 4 is a generalization of the known convergence conditions, but that Hypothesis 5 is of a different nature. It is the only convergence condition which may allow one to obtain upper bounds for Q smaller than F_e.

8. *Example.* Let $\mathcal{T} = E$, let $T = \{z \in E | z \geqq 0\}$, and let $A(\cdot)$ and $c(\cdot)$ be defined as follows:

$$A(z) = \begin{cases} [1; 2) & , \quad 0 \leqq z \leqq 5; \\ [z-4; z-3), & 5 < z; \end{cases} \qquad c(z) = \begin{cases} z, & 0 \leqq z < 2; \\ 2, & 2 \leqq z \leqq 4; \\ z, & 4 < z; \end{cases}$$

then $F = [0; 2]$ and $F_e = [0; 4]$. Pick $D = [0; 2]$. Hypothesis 5 is satisfied but Hypothesis 4 is not satisfied.

9. *Example.* Let $\mathcal{T} = E$, let $T = \{z \in E | z \geqq 0\}$, and let $A(\cdot)$ and $c(\cdot)$ be defined as follows:

$$A(z) = \begin{cases} [0.25; 0.5], & z = 10; \\ [z/4; z/2], & z \neq 10; \end{cases} \qquad c(z) = \begin{cases} 1, & z = 10; \\ z, & z \neq 10; \end{cases}$$

then $F = F_e = \{0\}$. Pick $D = \{0\}$. Hypothesis 4 is satisfied but Hypothesis 5 is not satisfied.

The classical conditions of convergence are repeated below to facilitate their comparison with Hypotheses 4 and 5.

10. HYPOTHESIS (Polyak). (i) $A(z)$ contains one and only one point;
(ii) $c(A(z))$ is upper semi-continuous on T;
(iii) $c(\cdot)$ is continuous on T.
11. HYPOTHESIS (Polak). (i) Identical to (i) of Hypothesis 4;
(ii) $c(\cdot)$ is either continuous on D^c or bounded from below on T.
12. HYPOTHESIS (Zangwill). (i) T is sequentially compact;
(ii) $c(\cdot)$ is continuous on T;
(iii) $A(\cdot)$ is closed on D^c; i.e., if $\{z_i\}$ converges to z in D^c, $\{y_i\}$ converges to y in T, and y_i is in $A(z_i)$ for all i, then y belongs to $A(z)$.

Polyak's theorem (Thm. 1, p. 865 of [10]) states that Hypothesis 10 implies that $F \supseteq Q$. One notes that if Hypothesis 10 is satisfied, then $F = F_e$; therefore, Polyak's theorem is a special case of Theorem 6 with $D = F_e$. Polak's theorem (Thm. 10, p. 15 of [9]) states that Hypothesis 11 implies that $D \supseteq Q$; clearly this theorem is a special case of Theorem 6. The case of Zangwill's theorem is slightly more complicated. Zangwill states (Convergence Theorem A, p. 91 of [11]) that if Hypothesis 12 is satisfied and if $D = F$, then $D \supseteq Q$. One can prove that if Hypothesis 12 is satisfied and if $D = F$, then $F = F_e$; therefore, Zangwill's theorem is also a special case of Theorem 6 with $D = F_e$.

Note that if one picks $D = [0; 2]$, then Hypotheses 10, 11, and 12 are not satisfied for Example 8.

Although Hypothesis 4 does not imply and is not implied by Hypothesis 5, it is possible to exhibit a relation between 4 and 5 by strengthening the assumptions on $c(\cdot)$. In particular, if (ii) of Hypothesis 4 is replaced by the assumption that $c(\cdot)$ is continuous, then Hypothesis 4 (modified) implies Hypothesis 5. Thus, in the case of $c(\cdot)$ being continuous, Hypothesis 11 implies Hypothesis 5. Furthermore, since Hypothesis 12 together with the assumption that $D \supseteq F$ implies Hypothesis 11 (§ [9, p. 16]), it follows that Hypothesis 12 with $D \supseteq F$ also implies Hypothesis 5. Finally, Hypothesis 10 also implies Hypothesis 5. One concludes that when $c(\cdot)$ is continuous, Hypothesis 5 is the weakest convergence condition available.

REFERENCES

[1] P. HUARD, *Tentatives de synthese dans les methodes de programmation non lineaire*, Cahiers Centre Etudes Recherche Oper., 16 (1974), pp. 347–368.
[2] ———, *Optimization algorithms and point-to-set mapping*, Math. Programming, 8 (1975), pp. 308–331.
[3] D. G. LUENBERGER, *Introduction to Linear and Non Linear Programming*, Addison-Wesley, Reading, MA, 1973.
[4] G. G. L. MEYER AND E. POLAK, *Abstract models for the synthesis of optimization algorithms*, this Journal, 9 (1971), pp. 547–560.
[5] G. G. L. MEYER, *Algorithm model for penalty functions type iterative procedures*, J. Comput. System Sci., 9 (1974), pp. 20–30.
[6] ———, *A canonical structure for iterative procedures*, J. Math. Anal. Appl., 52 (1975), pp. 120–128.
[7] ———, *A systematic approach to the synthesis of algorithms*, Numer. Math., 24 (1975), pp. 277–289.

[8] E. POLAK, *On the convergence of optimization algorithms*, Rev. Francaise Inf. Rech. Oper., 16 (1969), pp. 17–34.

[9] ———, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[10] B. T. POLYAK, *Gradient Methods for the Minimization of Functionals*, U.S.S.R. Comput. Math. and Math. Phys., 3 (1963), pp. 864–878.

[11] W. I. ZANGWILL, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

IDENTIFIABILITY OF SPATIALLY-VARYING AND CONSTANT PARAMETERS IN DISTRIBUTED SYSTEMS OF PARABOLIC TYPE*

S. KITAMURA† AND S. NAKAGIRI‡

Abstract. For the parameter identification process to minimize the difference between the system output and the model output, this paper discusses the identifiability of spatially-varying and constant parameters of the system described by a linear, 1-dimensional, parabolic partial differential equation. Only the parameters in the system equation (not in the boundary condition) are assumed to be unknown and the identifiability in the deterministic sense is treated. For both cases of distributed and pointwise measurements, several results for the parameter identifiability and nonidentifiability are obtained. As a result, the identifiability conditions depend on the profile of the state of the model for the case of the distributed measurement, while, for the case of the pointwise measurement, such conditions depend on the position of a detector and the form of initial or input functions. The results are represented in terms of a priori known quantities and are easily applied to practical problems.

1. Introduction. Recently, the parameter identification problem of distributed systems has been of great interest. In order to identify (or to estimate) the parameters in distributed systems, under the assumption of a given form of system equations, the method to minimize the difference between the system output and the model output is normally used. In many cases, this formulation leads to an optimization problem and many techniques for this purpose have been proposed [6], [10]. In this process, however, there arises the question of whether the parameters in the mathematical model coincide with those in the real system when the difference between the outputs of both systems vanishes or, more generally, the appropriately defined performance index takes the minimum value. Even for lumped systems, this problem comes into question and some results have been obtained [1, p. 43] [2].

For distributed systems, however, the concept of the transfer function is generally not so effective as in the case of lumped systems, and further the parameters in such systems are often functions of spatial variables. The parameter identifiability problem is, therefore, more important for distributed systems, not only in obtaining the mathematical model of controlled objects but in the experimental determination of physical constants in laboratory. The parameter identifiability problem could be formulated as the one-to-one property of the inverse problem, that is, the one-to-one property of the mapping from the space of system outputs to the space of parameters. However, the uniqueness of such a mapping apparently does not hold. Chavent [3, p. 100] considered this problem for the systems described by elliptic and parabolic partial differential equations and obtained a sufficient condition (unicity in his terminology) only for the case of distributed measurements. He gave also some examples showing the possibility of the appearance of an unbounded solution in the estimation process. Reissenweber [8, p. 76] treated this problem by the sensitivity function for the system with constant parameters. His method, however, requires the parameter value which is a priori unknown for estimators. Seinfeld [9] referred to the relation between the

* Received by the editors April 2, 1976, and in revised form December 14, 1976.

† Department of Instrumentation Engineering, Kobe University, Kobe, Japan. Now at Institut A für Mechanik, Universität Stuttgart, Stuttgart, F. R. Germany.

‡ Department of Applied Mathematics, Faculty of Engineering, Kobe University, Kobe, Japan.

observability and the identifiability but such a relation has not yet been clarified completely.

In this paper, the parameter identifiability problems for the system described by a linear parabolic partial differential equation on the 1-dimensional spatial domain are treated. It is assumed that only the functional and constant parameters in the system equation are unknown and that the measurement is performed without deterministic and stochastic errors. The definition of the parameter identifiability is given in a different form from that defined for lumped systems [1]. Since the identifiability conditions depend so much on the characteristic of measurement systems as the typical form from the mathematical and practical standpoints, two cases of the distributed and pointwise measurements are studied. Under these formulations, several conditions for the parameter identifiability and nonidentifiability are obtained. The conditions are represented in terms of the known quantities only, and are easily interpreted and applied to practical identification process.

2. Statement of the problem. Figure 1 shows a conventional system for the parameter identification with a model. In the following we study the uniqueness of the estimated parameters by the scheme in Fig 1, where the system is a distributed system described by a parabolic partial differential equation.

Let the system be described by

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(a(x)\frac{\partial u}{\partial x}\right)+b(x)u+f(x,t), \qquad x\in(0,1), \quad t>0, \tag{1}$$

where $u=u(x,t)$ is a scalar state variable, $f(x,t)$ a forced input function. Boundary and initial conditions are given as

$$\alpha_0 u(t,0)+(1-\alpha_0)\frac{\partial u}{\partial n}(t,0)=g_0(t), \qquad 0\leqq\alpha_0\leqq 1,$$

$$\alpha_1 u(t,1)+(1-\alpha_1)\frac{\partial u}{\partial n}(t,1)=g_1(t), \qquad 0\leqq\alpha_1\leqq 1, \tag{2}$$

$$u(x,0)=u_0(x) \tag{3}$$

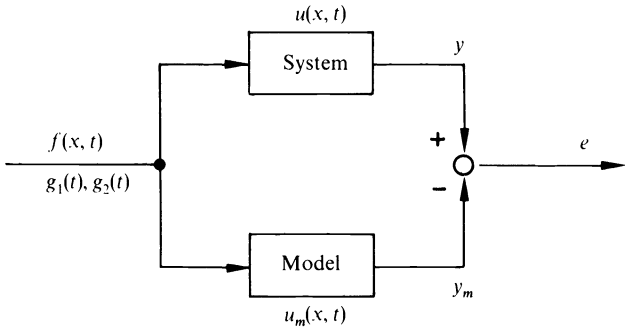


FIG. 1. *Parameter identification by using a model*

where $\partial u/\partial n$ denotes a derivative of u in the outward normal direction of the boundary (0, 1).

The measurement of the system state is made and the output y of the measurement system is written by

$$y(x_p, t) = Cu(x, t), \qquad x_p \in \Omega_p \subseteq [0, 1], \qquad t \geqq 0, \tag{4}$$

where C is an operator with the form depending on the characteristics of the sensor and the equipment followed to the sensor. Ω_p denotes a subspace of $[0, 1]$, where the output variable y is defined. The operator C is a linear bounded, time-invariant operator from $C([0, 1]\times[0, \infty))$ to $C(\Omega_p \times [0, \infty))$.

Throughout this paper we assume that:

(i) the form of equation (1) is a priori known, but parameters $a(x)$ and $b(x)$ are unknown except that $a(x) > 0$, $a(\cdot) \in C^2[0, 1]$ and $b(\cdot) \in C^1[0, 1]$,

(ii) parameters in the boundary condition are a priori known,

(iii) operator C is a priori known, and

(iv) input functions $f(\cdot) \in C^1([0, 1]\times[0, \infty))$ and $g_0(\cdot)$, $g_1(\cdot) \in C^2[0, \infty)$ can be measured, i.e., are known functions.

Under the assumptions (i) and (iv), a unique solution of (1) and (2) exists for a given $u_0(\cdot) \in C[0, 1]$, and this solution satisfies $u(\cdot, t) \in C^2[0, 1]$ for all $t > 0$ and $u(x, \cdot) \in C^1[0, \infty)$ for all $x \in [0, 1]$[5].

From the practical point of view, assumption (i) can be considered as realistic, but assumption (ii), (iii) and (iv) might not be. For identification of such terms as the boundary data, measurement operator, and input functions, we will require different formulations and approaches.

Now the model can be described by

$$\frac{\partial u_m}{\partial t} = \frac{\partial}{\partial x}\left(a_m(x)\frac{\partial u_m}{\partial x}\right) + b_m(x)u_m + f(x, t), \qquad x \in (0, 1), \quad t > 0, \tag{5}$$

$$y_m(x_p, t) = Cu_m(x, t), \qquad x_p \in \Omega_p, \quad t \geqq 0, \tag{6}$$

where $u_m(x, t)$ is the state of the model and the subscript m denotes model quantities. The boundary condition for (5) takes the same form as in (2). The knowledge of the initial condition for (5), however, depends on the form of operator C (see next two sections).

DEFINITION. We shall call an unknown parameter *identifiable* if it can be determined uniquely in all points of its domain by using the input-output relation of the system and the input-output data.

That is, in the following, the parameters $a(x)$ and/or $b(x)$ are said to be identifiable if, for all $x \in [0, 1]$, $a(x) = a_m(x)$ and/or $b(x) = b_m(x)$ follow uniquely from the relation $e(x_p, t) = y(x_p, t) - y_m(x_p, t) = 0$ for all $t \geqq 0$ and all $x_p \in \Omega_p$. The condition $e(x_p, t) = 0$ for all $x_p \in \Omega_p$ and all $t \geqq 0$ means the following: normally, in the identification process in Fig. 1, the parameters in the model are adjusted by some proper algorithm so that the difference e goes to zero in $C(\Omega_p \times [0, \infty))$. The identifiability problem of parameters occurs at the final stage of such an algorithm, that is, at $e = 0$ in $C(\Omega_p \times [0, \infty))$. Note here that the information which is usable to check the identifiability defined above is not the true values of unknown parameters but the measured output and the quantities concerning the model.

Since the identifiability property depends greatly on the form of the operator C, the cases of distributed and pointwise measurements are treated in two sections separately.

3. The case of distributed measurement. In this section, it is assumed that $u(x, t)$ is measured at all points of $x \in [0, 1]$ and $t \geqq 0$; hence, we may set $C = I$, where I denotes the identity operator from $C([0, 1] \times [0, \infty))$ onto itself. We may thus assume, without loss of generality, that the initial function $u_0(x)$ is known. Defining the difference variable $e(x, t) = u(x, t) - u_m(x, t)$, we have the following lemma.

LEMMA 1. *$e(x, t) = 0$ for all $x \in [0, 1]$ and all $t \geqq 0$ holds if and only if*

$$\frac{\partial}{\partial x}\left[(a(x) - a_m(x))\frac{\partial u_m}{\partial x}(x, t)\right] + (b(x) - b_m(x))u_m(x, t) = 0 \tag{7}$$

for all $x \in (0, 1)$ and all $t > 0$.

Proof. See Appendix A.

3.1. Identifiability of $a(x)$. In the following it is assumed that $e(x, t) = 0$ for all $x \in [0, 1]$ and all $t \geqq 0$, and $b(x)$ is known or $b(x) = 0$.

Remark 1. All the results in § 3 are represented in terms of the state of the model u_m. However, u_m may be replaced by the state of the system u, since $e = 0$.

Remark 2. The condition $e = 0$ for all $x_p \in [\Omega_p]$ and all $t \geqq 0$ in §§ 3 and 4 may be weakened by some suitable property of the solution, for example, analyticity.

Let us define

$$\begin{aligned} E(t) &= \left\{x \in [0, 1] \middle| \frac{\partial u_m}{\partial x}(x, t) = 0\right\}, \\ G(t) &= [0, 1] - E(t). \end{aligned} \tag{8}$$

RESULT 1. *Parameter $a(x)$ is identifiable if there exists some $t_1 > 0$ such that*

$$E(t_1) \neq \varnothing \tag{9}$$

and

$$\overline{G(t_1)} = [0, 1] \tag{10}$$

where $\varnothing$ is an empty set and $\bar{G}$ is the closure of G. The condition (10) *especially may be replaced by*

$$\text{meas } E(t_1) = 0 \tag{11}$$

where meas E *denotes the Lebesgue measure of E.*

Proof. By the assumption we obtain from (7)

$$\frac{\partial}{\partial x}\left[q(x)\frac{\partial u_m}{\partial x}\right] = 0 \quad \text{for all } x \in (0, 1) \text{ and all } t > 0 \tag{12}$$

where $q(x) = a(x) - a_m(x)$, and further

$$q(x)\frac{\partial u_m}{\partial x} = c(t)$$

where $c(t)$ is a function of t. By condition (9) there exists a t_1 such that $c(t_1)=0$. From condition (10) the set $\{x \in [0, 1] | q(x)=0\}$ is dense in $[0, 1]$. Here we note $q(x)$ is a continuous function on $[0, 1]$ (assumption in § 2). Thus $q(x)=0$ for all $x \in [0, 1]$, i.e., $a(x)$ is identifiable. Condition (11) implies meas $G(t_1)=1$, and consequently, meas $\overline{G(t_1)}=1$. This means $\overline{G(t_1)}=[0, 1]$. To show this, assume $[0, 1]-\overline{G(t_1)} \neq \varnothing$. Then, there exists an interval J such that $[0, 1]-\overline{G(t_1)} \supset J$. From $[0, 1] \supset \overline{G(t_1)} \cup J$, we obtain $1 \geqq \text{meas}\, \overline{G(t_1)} + \text{meas}\, J$, which implies meas $J=0$. This is a contradiction. Q.E.D.

Now another condition is given for the identifiability of $a(x)$.

RESULT 2. *$a(x)$ is identifiable if*

$$E(t) \neq \varnothing \quad \text{for all } t>0 \tag{13}$$

and

$$\overline{\bigcup_{t>0} G(t)} = [0, 1]. \tag{14}$$

Condition (14) *especially may be replaced by*

$$\text{meas}\left(\bigcap_{t>0} E(t)\right) = 0 \tag{15}$$

Proof. By condition (13) and Lemma 1, we obtain $q(x)(\partial u_m/\partial x)(x, t)=0$ for all $x \in (0, 1)$ and all $t>0$, where $q(x)=a(x)-a_m(x)$. Set $M=\bigcup_{t>0} G(t)$. For any $x \in M$, there exists some $t(x)>0$ such that $x \in G(t)$, i.e., $(\partial u_m/\partial x)(x, t) \neq 0$. Thus, $q(x)=0$ for all $x \in M$, and from condition (14) and the continuity for $q(x)$ it follows that $q(x)=0$ for all $x \in [0, 1]$. Moreover, condition (15) implies condition (14) as in Result 1. Q.E.D.

Note that condition (13) is stricter than (9), while condition (14) weaker than (10). The following Result 3 is a counterpart to Result 2, and Results 4 and 5 are counterparts to Result 1.

RESULT 3. *$a(x)$ is not identifiable if $\bigcup_{t>0} G(t)$ is not dense in $[0, 1]$, especially if $\bigcap_{t>0} E(t)$ includes an interval.*

Proof. We show that $a(x_0) \neq a_m(x_0)$ for some $x_0 \in (0, 1)$ even if $e(x, t)=0$ for all $x \in [0, 1]$ and all $t \geqq 0$ when $\bigcup_{t>0} G(t)$ is not dense in $[0, 1]$. By the first condition, there exists an interval J satisfying $[0, 1]-\overline{\bigcup_{t>0} G(t)} \supset J$. Take x_0 and $\varepsilon>0$ such that $J \supset (x_0-\varepsilon, x_0+\varepsilon)$, and let $r(x)$ be a twice continuously differentiable function in $[0, 1]$ with support in $(x_0-\varepsilon, x_0+\varepsilon)$ and $r(x_0) \neq 0$. Assume here $a(x)=a_m(x)+r(x)$. If $x \in (x_0-\varepsilon, x_0+\varepsilon)$, then $\partial u_m/\partial x=0$ for all $t>0$ since

$$x \in J \subset [0, 1]-\overline{\bigcup_{t>0} G(t)} \subset [0, 1]-\bigcup_{t>0} G(t) = \bigcap_{t>0} E(t)$$

and if $x \notin (x_0-\varepsilon, x_0+\varepsilon)$, then $(a(x)-a_m(x))(\partial u_m/\partial x)=0$ for all $x \in (0, 1)$ and all $t>0$ since $r(x)=0$. Thus, by Lemma 1 $e(x, t)=0$ for all $x \in [0, 1]$ and all $t \geqq 0$, and $a(x)$ is not identifiable. Moreover, if $\bigcap_{t>0} E(t)$ includes an interval, $\bigcup_{t>0} G(t)$ is not dense in $[0, 1]$. Q.E.D.

RESULT 4. *If* $E(t_1)=\varnothing$ *for some* t_1, *then*

$$a(x)-a_m(x)=(a(x_0)-a_m(x_0))\exp\left[-\int_{x_0}^{x}\frac{\frac{\partial^2 u_m}{\partial x^2}(s,t_1)}{\frac{\partial u_m}{\partial x}(s,t_1)}ds\right] \tag{16}$$

for any x *and* $x_0\in[0,1]$.

Proof. By the assumption, (12) holds in this case, i.e.,

$$\frac{\partial u_m}{\partial x}q'(x)+\frac{\partial^2 u_m}{\partial x^2}q(x)=0.$$

Equation (16) is a solution of this differential equation under the condition $(\partial u_m/\partial x)(x,t_1)\neq 0$. Q.E.D.

Remark 3. Result 4 does not necessarily imply the nonidentifiability of parameter $a(x)$. However, if $E(t)=\varnothing$ for all $t\geqq 0$, then $a(x)$ is not identifiable as long as $a(x_0)\neq a_m(x_0)$ for an arbitrary point $x_0\in[0,1]$. This means that a priori knowledge of $a(x)$ is required for the identification of $a(x)$.

RESULT 5. *If* $E(t_1)=\varnothing$ *for some* $t_1>0$ *and if* $u_m(x,t)$ *is represented as* $v_m(x)w_m(t)$, *then* $a(x)$ *is not identifiable* (refer to Result 8 also).

Proof. $E(t_1)=\varnothing$ implies that $w_m(t_1)\neq 0$ and $(\partial v_m/\partial x)(x)\neq 0$ for any $x\in[0,1]$. Let $a(x)=a_m(x)+1/(\partial v_m(x)/\partial x)$, then $a(x)\neq a_m(x)$ for all $x\in[0,1]$, while

$$(a(x)-a_m(x))\frac{\partial v_m(x)}{\partial x}w_m(t)=w_m(t)$$

for all $x\in[0,1]$ and $t\geqq 0$. Thus,

$$\frac{\partial}{\partial x}\left\{(a(x)-a_m(x))\frac{\partial u_m}{\partial x}\right\}=0$$

for all $t>0$ and, from Lemma 1, $e(x,t)=0$ for all $x\in[0,1]$ and all $t\geqq 0$. Thus, $a(x)$ is not identifiable. Q.E.D.

Interpretations. Results 1 and 2 both require, roughly speaking, the existence of at most countably many peaks or valleys in the profile of u_m (or u) for the identifiability of $a(x)$, when e becomes identically zero in Fig. 1. This condition is very easily understood and applied to practical processes. Actually, if the conditions in Result 1 or 2 are not satisfied, we can construct, as below, simple systems which are not identifiable.

Example 1. Let us assume that $b(x)=0$ and $a(x)$ is a priori known to be constant. Take in equations (1)–(3)

$$f(x,t)=\dot{c}(t)x+\dot{d}(t), \qquad \cdot=d/dt,$$
$$u(0,t)=d(t), \qquad u(1,t)=c(t)+d(t),$$
$$u(x,0)=c_0x+d_0,$$

where $c(t)$ and $d(t)$ are continuously differentiable functions with a definite sign, and $c(0)=c_0$, $d(0)=d_0$. Then, the solution of (5) is always given by

$$u_m(x,t)=c(t)x+d(t)$$

whatever positive constant value the parameter $a_m(x)$ takes. Thus, $a(x)$ is not identifiable. Note that $E(t) = \varnothing$ for all $t \geqq 0$ in this example and that Result 5 does not apply here too.

On the other hand, Result 3 shows that, roughly speaking, the parameter $a(x)$ is not identifiable in the spatial interval contained in $\bigcap_{t>0} E(t)$. This becomes rather important when the distributed system is approximated by the lumped system to perform numerical computations. For example, in many cases, the normal finite-difference approximation causes $\partial u_m/\partial x = 0$ to hold in the finite spatial interval even if the true solution of the original distributed system has a zero of $\partial u_m/\partial x$ at only one point. Such an example is given below.

Example 2. Consider the system

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a(x)\frac{\partial u}{\partial x}\right), \qquad a(x) = e^{-x}, \tag{17}$$

$$u(0, t) = 0, \qquad \frac{\partial u}{\partial x}(1, t) = 0,$$
$$u(x, 0) = \text{const.}\ (\neq 0). \tag{18}$$

The model of (5) is used with $b(x) = 0$ and $f(x, t) = 0$, and boundary and initial conditions the same as in (18). The parameter a_m is to be time-invariant during the identification process and is adjusted by using Lyapunov's method so as to make $e = u - u_m$ tend to zero. If $a_m(x, t)$ satisfies

$$K\frac{\partial a_m}{\partial t} + \frac{\partial e}{\partial x}\frac{\partial u_m}{\partial x} = 0; \qquad K\text{: positive constant}, \tag{19}$$

and if the Result 2 holds, then

$$\|e(\cdot, t)\|_{L^2} \text{ and } \|a_m(\cdot, t) - a(\cdot)\|_{L^2} \to 0 \quad \text{as } t \to \infty \tag{20}$$

hold for arbitrary $e(x, 0)$ and $a_m(x, 0)$ (see Appendix B). Figure 2 is a numerical result. Equations (17), (19) and the equation of the model are approximated by the finite-difference method with 20 divisions of [0, 1] and with time-step 0.005. Figure 2 is obtained for $u(x, 0) = 1.5$, $a_m(x, 0) = 0.5$ and $K = 1$. It is clear that $a_m(x, t)$ approaches the true value $a(x)$ except in the domain containing $x = 1$. This is caused by the fact that, due to the finite-difference, $\partial u_m/\partial x = 0$ holds not only at the point $x = 1$ but in the interval $J = (0.95, 1)$, so the parameter $a(x)$ is not identifiable in J according to Result 3. Namely, $a(x)$ in (17) becomes not identifiable in the process of numerical computations although it is mathematically identifiable in the original system.

3.2. Identifiability of $b(x)$. It is assumed that $a(x)$ is known or $a(x) = 0$, and that $e(x, t) = 0$ for all $x \in [0, 1]$ and all $t \geqq 0$. Let us define

$$F(t) = \{x \in [0, 1] | u_m(x, t) = 0\},$$
$$H(t) = [0, 1] - F(t). \tag{21}$$

RESULT 6. *$b(x)$ is identifiable if and only if*

$$\overline{\bigcup_{t>0} H(t)} = [0, 1]. \tag{22}$$

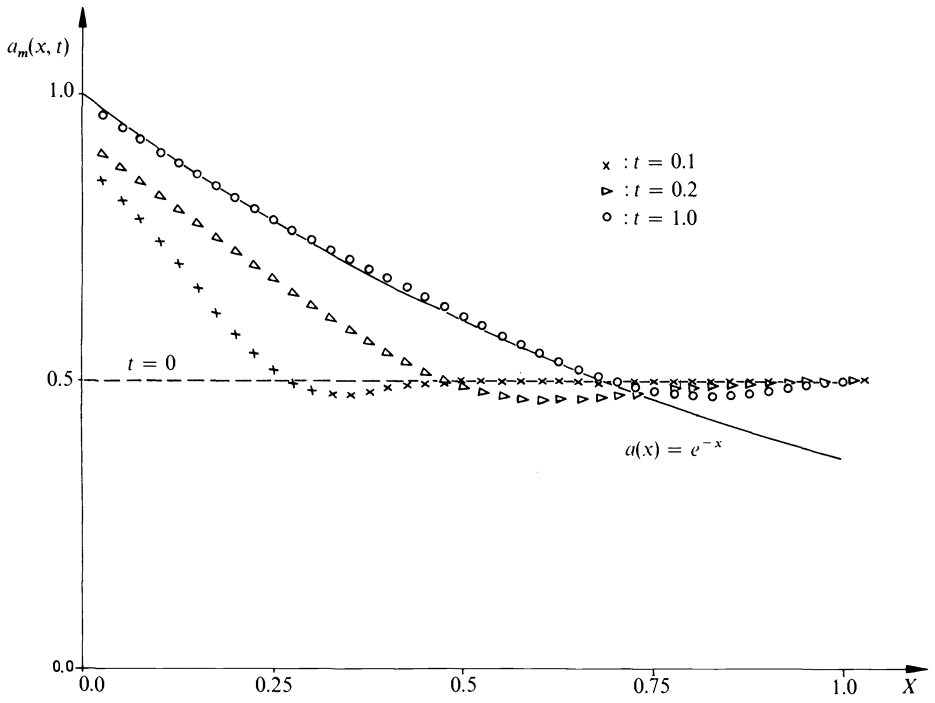


FIG. 2. *Numerical result—identification of* $a(x)$

$b(x)$ is especially identifiable if

$$\text{meas}\left(\bigcap_{t>0} F(t)\right) = 0$$

and not identifiable if $\bigcap_{t>0} F(t)$ *includes an interval.*

Proof. By the assumption and Lemma 1, $e(x, t) = 0$ if and only if $(b(x) - b_m(x))u_m(x, t) = 0$. Sufficiency of the statement follows from condition (22). Necessity follows by proceeding similarly as in the proof of Result 3. The latter statement of the result is self-evident. Q.E.D.

3.3. Identifiability of $a(x)$ and $b(x)$. In this case, the known quantity is $u_m(x, t)$ only, while the unknowns are $a(x)$ and $b(x)$. Fairly restrictive conditions will be required for the identifiability of both parameters.

RESULT 7. *If the functions* $u_m(x, t)$, $(\partial u_m/\partial x)(x, t)$ *and* $(\partial^2 u_m/\partial x^2)(x, t)$ *are linearly independent as functions of* t *on a dense subset in* $[0, 1]$, *then* $a(x)$ *and* $b(x)$ *are simultaneously identifiable.*

Proof. By setting $q_1(x) = a(x) - a_m(x)$ and $q_2(x) = b(x) - b_m(x)$, we obtain from (7)

$$q_1(x)\frac{\partial^2 u_m}{\partial x^2}(x, t) + q_1'(x)\frac{\partial u_m}{\partial x}(x, t) + q_2(x)u_m(x, t) = 0 \tag{23}$$

for all $x \in (0, 1)$ and all $t > 0$. From the assumption of linear independence, $q_1(x) = q_1'(x) = q_2(x) = 0$ on some dense set in $[0, 1]$, and again by continuity, $q_1(x) = q_2(x) = 0$ for all $x \in [0, 1]$. Q.E.D.

For linear independence of the three functions in Result 7, the nonzero input function $f(x, t)$ or $g_i(t)$ $(i = 0, 1)$ will probably be required. The linear independence of functions may be examined by such means like the Wronskian matrix.

For the nonidentifiability of $a(x)$ and $b(x)$, Results 3 and 6 apply. As an extension of Result 5, let us consider the case that a solution u_m (or u) is represented as a product of a function of x by a function of t. Such a case, for example, occurs when the initial function $u_0(x)$ and the input function $f(x, t) = f(x)$ $(g_1(t) = g_2(t) = 0)$ are given by an eigenfunction of Sturm–Liouville's problem as follows:

$$\frac{d}{dx}\left(a_m(x)\frac{d\psi_m}{dx}\right) + b_m(x)\psi_m = 0,$$

$$\alpha_0\psi(0) - (1-\alpha_0)\frac{d\psi_m}{dx}(0) = 0,$$

$$\alpha_1\psi(1) + (1-\alpha_1)\frac{d\psi_m}{dx}(1) = 0.$$

The steady state is a special case of this class, where the steady state $u_m(x, t) = u_{ms}(x)$ is defined by (2) and (5) with

$$\frac{\partial u_m}{\partial t} = 0, \quad f(x, t) = f(x) \quad \text{and} \quad g_i(t) = \text{const.} \quad (i = 0, 1).$$

In that case we have the following.

RESULT 8. *If $u_m(x, t) = v_m(x)w_m(t)$, then $a(x)$ and $b(x)$ are not simultaneously identifiable. This statement holds especially at the steady state.*

Proof. For any function $v_m(x)$ which is twice continuously differentiable, we can select nonzero functions $q_1(x)$ and $q_2(x)$ which satisfy the following equation:

$$\frac{d}{dx}\left(q_1(x)\frac{dv_m}{dx}(x)\right) + q_2(x)v_m(x) = 0 \quad \text{for all } x \in (0, 1).$$

Multiplication by $w_m(t)$ yields

$$\frac{\partial}{\partial x}\left(q_1(x)\frac{\partial u_m}{\partial x}(x, t)\right) + q_2(x)u_m(x, t) = 0$$

for all $x \in (0, 1)$ and all $t > 0$. Since $q_1(x)$ and $q_2(x)$ are nonzero, $a(x)$ and $b(x)$ are not simultaneously identifiable from Lemma 1. Q.E.D.

The above result implies, for example, in the case of the steady state, that it is not sufficient to consider only the difference e for the identification of both $a(x)$ and $b(x)$. However, if we have an a priori knowledge that shows $a(x)$ and $b(x)$ to be constant, the following result is obtained.

RESULT 9. *Let $a(x)$ and $b(x)$ be constant. Parameters a and b are identifiable*

if one of the following conditions is satisfied.

(i) *There exists* $x_0 \in (0, 1)$ *for which* $u_m(x_0, t)$ *and* $(\partial^2 u_m/\partial x^2)(x_0, t)$ *are linearly independent as functions of* t.

(ii) *There exist* $x_1, x_2 \in (0, 1)$ *and* $t_1 > 0$ *such that*

$$u_m(x_1, t_1) = 0, \qquad \frac{\partial^2 u_m}{\partial x^2}(x_1, t_1) \neq 0$$

and

$$u_m(x_2, t_1) \neq 0, \qquad \frac{\partial^2 u_m}{\partial x^2}(x_2, t_1) = 0.$$

Further, in the case of the steady state, parameters a *and* b *are identifiable if there exists* $x_0 \in (0, 1)$ *such that*

$$u_{ms}(x_0) = 0 \quad \textit{and} \quad f(x_0) \neq 0.$$

Proof. The first two results follow from Result 7 and (23). The last statement is proved as follows. The model equation

$$\frac{d}{dx}\left(a_m \frac{du_{ms}}{dx}(x)\right) + b_m u_{ms}(x) + f(x) = 0$$

and (7) yield

$$(b_m q_1 - a_m q_2) u_{ms}(x) - q_1 f(x) = 0.$$

By the assumption, we have directly $q_1 = 0$. Since $f(x_0) \neq 0$, $u_{ms}(x) \not\equiv 0$. Then, we have $u_{ms}(x_1) \neq 0$ at some point x_1, which implies $q_2 = 0$ Q.E.D.

4. The case of pointwise measurement. In this section we consider the case of a pointwise measurement, i.e., the measured output y is represented by the following equation

$$y(t) = Cu(x, t) = \int_0^1 \delta(x - x_p) u(x, t)\, dx = u(x_p, t) \tag{24}$$

where δ is the Dirac function and x_p denotes the position of a detector. Since the state $u(x, t)$ is measured only at one point in the spatial domain, we understand intuitively that it is impossible to determine uniquely $a(x)$ and $b(x)$ as functions of the spatial variable. Actually, we can easily construct an example which is not identifiable. Hence, throughout this section, both $a(x)$ and $b(x)$ are assumed to be constant.

First, we give a lemma concerning the eigenvalue problem. Let us define

$$h(\lambda) = \frac{(\alpha_0 + \alpha_1 - 2\alpha_0\alpha_1)\lambda}{(1 - \alpha_0)(1 - \alpha_1)\lambda^2 - \alpha_0\alpha_1}.$$

LEMMA 2. *Consider the eigenvalue problem*:

$$a\frac{d^2\psi_n(x)}{dx^2}+b\psi_n(x)=-k_n\psi_n(x), \qquad a>0, \quad (a, b: \textit{constants}),$$

$$\alpha_0\psi_n(0)-(1-\alpha_0)\frac{d\psi_n}{dx}(0)=0, \tag{25}$$

$$\alpha_1\psi_n(1)+(1-\alpha_1)\frac{d\psi_n}{dx}(1)=0, \quad (\alpha_0, \alpha_1)\in[0, 1]\times[0, 1].$$

Then we have:

(i) *The case of* $(\alpha_0, \alpha_1)\neq(0, 0), (0, 1)$ *and* $(1, 0)$. *Let* $\{\lambda_n\}_{n=1,2\cdots}$ *be a set of roots of the equation*

$$\tan\lambda=h(\lambda), \qquad \lambda>0,$$

and assume that the λ_n*'s are monotone increasing. Then, the* n*-th eigenvalue* k_n *is given by*

$$k_n=a\lambda_n^2-b, \qquad n=1, 2, \cdots,$$

and the corresponding eigenfunction $\psi_n(x)$ *is given by*

$$\psi_n(x)=\alpha_0\sin\lambda_n x+(1-\alpha_0)\lambda_n\cos\lambda_n x. \tag{26}$$

(ii) *The case of* $(\alpha_0, \alpha_1)=(0, 0)$. *Let* $\lambda_n=(n-1)\pi$. *Then*

$$k_n=a\lambda_n^2-b, \qquad n=1, 2, \cdots,$$

and the corresponding eigenfunction is $\psi_n(x)=\cos(n-1)\pi x$,

(iii) *The case of* $(\alpha_0, \alpha_1)=(0, 1)$ *or* $(1, 0)$. *Let* $\lambda_n=(n-\frac{1}{2})\pi$. *Then*

$$k_n=a\lambda_n^2-b, \qquad n=1, 2, \cdots,$$

and the corresponding eigenfunction is given by (26).

Proof. The proof is omitted.

Note that in each case the eigenfunctions depend only on the boundary data (α_0, α_1), while the eigenvalues depend on a and b and every eigenvalue has multiplicity one. In what follows, it is assumed that the eigenfunctions $\psi_n(x)$ are already normalized. Then, the solution $u(x, t)$ of the system (1)–(3) with $a(x)=a$ and $b(x)=b$ is repesented as

$$\begin{aligned} u(x, t)=&\sum_{n=1}^{\infty}(u_0, \psi_n)\,e^{-k_n t}\psi_n(x)\\ &+\int_0^t\int_0^1\Big(\sum_{n=1}^{\infty}e^{-k_n(t-\tau)}\psi_n(x)\psi_n(y)\Big)f(y, \tau)\,dy\,d\tau\\ &+\int_0^t\Big(\sum_{n=1}^{\infty}(\psi_n(1)-\psi_n'(1))\,e^{-k_n(t-\tau)}\psi_n(x)\Big)g_1(\tau)\,d\tau\\ &-\int_0^t\Big(\sum_{n=1}^{\infty}(\psi_n(0)+\psi_n'(0))\,e^{-k_n(t-\tau)}\psi_n(x)\Big)g_0(\tau)\,d\tau \end{aligned} \tag{27}$$

where

$$(u, \psi_n) = \int_0^1 u(x)\psi_n(x)\, dx$$

and the prime stands for the derivative with respect to x.

The solution $u_m(x, t)$ of the model (see (5)) is obtained from (27) by replacing k_n by k_n^m and $u_0(x)$ by $u_{m0}(x)$, where k_n^m is the nth eigenvalue of the eigenvalue problem (25) with $a = a_m$ and $b = b_m$.

Before giving the first result, we shall consider the following simple example.

Example 3. Consider the system and model defined by

System:	Model:
$\dfrac{\partial u}{\partial t} = a\dfrac{\partial^2 u}{\partial x^2} + bu,$	$\dfrac{\partial u_m}{\partial t} = a_m\dfrac{\partial^2 u_m}{\partial x^2} + b_m u_m,$
$u(0, t) = u(1, t) = 0,$	$u_m(0, t) = u_m(1, t) = 0,$
$u(x, 0) = u_0(x).$	$u_m(x, 0) = u_0(x).$

The two solutions are expanded by the eigenfunctions as

$$u(x, t) = \sum_{n=1}^{\infty} 2(u_0, \sin n\pi x)\, e^{-(an^2\pi^2 - b)t} \sin n\pi x,$$

$$u_m(x, t) = \sum_{n=1}^{\infty} 2(u_0, \sin n\pi x)\, e^{-(a_m n^2\pi^2 - b_m)t} \sin n\pi x.$$

Let, for example, $u_0(x) = \sin 2\pi x + \sin 4\pi x$ and $x_p = \frac{1}{4}$; then both outputs $y(t)$ and $y_m(t)$ are positive, while

$$y(t) - y_m(t) = e^{-(4a\pi^2 - b)t} - e^{-(4a_m\pi^2 - b_m)t} = 0$$

for all $t \geqq 0$ if a_m and b_m satisfy $4a\pi^2 - b = 4a_m\pi^2 - b_m$. Thus, the parameters a and b are not identifiable. However, if we choose x_p such that $\sin 4\pi x_p \neq 0$, then $y(t) - y_m(t) = 0$ for all $t \geqq 0$ yields the identifiability of both parameters.

This example shows that the parameter identifiability depends on the form of the initial function and the measurement point. Actually, as we will discuss later, the identifiability has a close relation to the observability of system states.

Let us now define

$$\begin{aligned}
&P_n = \{x \in [0, 1]: \psi_n(x) \neq 0\}, \qquad P_{n,k} = P_n \cap P_k,\\
&Q_n = \{u(\cdot) \in C[0, 1]: (u, \psi_n) \neq 0\}, \qquad Q_{n,k} = Q_n \cap Q_k,\\
&R = \cup\{P_n \times Q_n : n \in N\},\\
&S = \cup\{P_{n,k} \times Q_{n,k} : n \neq k \text{ and } (n, k) \in N \times N\},
\end{aligned} \tag{28}$$

where $\psi_n(x)$ is the nth eigenfunction in Lemma 2. Note that $S \subset R \subset [0, 1] \times C[0, 1]$ and both S and R are open dense subsets in $[0, 1] \times C[0, 1]$. Under these definitions we obtain the first result.

RESULT 10. *Let $f(x, t) = 0$, $g_0(t) = g_1(t) = 0$ in (1) and (2), and $u_0(x)$ be known. Parameters a and b are identifiable if and only if the pair of the measurement*

point and the initial function, (x_p, u_0), *belongs to the set S. Moreover, if* $(\alpha_0, \alpha_1) \neq (0, 0)$ *and one parameter is known, the other is identifiable if and only if* (x, u_0) *belongs to the set R. If* $(\alpha_0, \alpha_1) = (0, 0)$, *then*

(A) *if a is known, then b is identifiable if and only if* $(x_p, u_0) \in R$, *or*

(B) *if b is known, then a is identifiable if and only if* $(x_p, u_0) \in \cup\{P_n \times Q_n, n \in N - \{1\}\}$.

Proof. First we give

LEMMA 3. *Let* $\{k_n\}_{n=1,2,\cdots}$ *and* $\{k_n^m\}_{n=1,2,\cdots}$ *be strictly monotone increasing sequences tending to infinity and let*

$$\sum_{n=1}^{\infty} C_n(e^{-k_n t} - e^{-k_n^m t}) = 0 \quad \text{for all } t \in [0, \infty).$$

If $C_q \neq 0$ *for some q, then* $k_q = k_q^m$.

Proof of Lemma 3. The proof is omitted.

Under the assumptions, the difference of $u(x_p, t)$ and $u_m(x_p, t)$ is given from (27) as

$$y(t) - y_m(t) = u(x_p, t) - u_m(x_p, t)$$

$$= \sum_{n=1}^{\infty} C_n(e^{-k_n t} - e^{-k_n^m t})$$

where $C_n = (u_0, \psi_n)\psi_n(x_p)$ and k_n^m denotes the nth eigenvalue for the model. The condition $(x_p, u_0) \in S$ is sufficient. Indeed, this condition implies that there exist natural numbers i and j $(i \neq j)$ such that $C_i \neq 0$ and $C_j \neq 0$. Assume that $y(t) - y_m(t) = 0$. Then we have $k_i = k_i^m$ and $k_j = k_j^m$ by Lemma 3; hence, $a\lambda_i^2 - b = a_m\lambda_i^2 - b_m$ and $a\lambda_j^2 - b = a_m\lambda_j^2 - b_m$. Since $\lambda_i \neq \lambda_j$, these two equalities imply $a = a_m$ and $b = b_m$.

Next we shall turn to the proof of the necessity. We show that $(x_p, u_0) \notin S$ implies the nonidentifiability of two parameters a and b at the same time. Since $(x_p, u_0) \notin S$, $C_n = (u_0, \psi_n)\psi_n(x_p) = 0$ except for at most one n. If $C_n = 0$ for all n, we have $y(t) = y_m(t) = 0$ for arbitrary a and b. If $C_q \neq 0$ for some q and $C_n = 0$ for all $n \neq q$, then

$$y(t) = y_m(t) = (u_0, \psi_q)\psi_q(x_p)\, e^{-k_q^m t}$$

for $a_m = a + 1/\lambda_q^2$ and $b_m = b + 1$ if $\lambda_q \neq 0$, and for $a_m = 2a$ if $\lambda_q = 0$. Thus, two parameters a and b are not identifiable at the same time. The latter statement of Result 10 is self-evident. Q.E.D.

Here, let us define a set P_0 by

$$P_0 = \{x \in [0, 1]: \psi_n(x) \neq 0 \text{ for all } n \in N\}. \tag{29}$$

Note that the set $[0, 1] - P_0$ is countable. If an initial function $u_0(x)$ is neither identically zero nor an eigenfunction (multiplied by a constant), then at least two Fourier coefficients of $u_0(x)$, expanded by $\psi_n(x)$, do not vanish. The following is a simplied version of Result 10.

RESULT 11. *Let the assumption in Result* 10 *hold. Let* $x_p \in P_0$ *and* $u_0(x)$ *be neither identically zero nor an eigenfunction (multiplied by a constant). Then parameters a and b are identifiable. Moreover, if* $(\alpha_0, \alpha_1) \neq (0, 0)$ *and one parameter*

is known, then the other is identifiable if $u_0(x)$ *is not identically zero. If* $(\alpha_0, \alpha_1) = (0, 0)$, *then*

(A) *if a is known and* $u_0(x)$ *is not identically zero, then b is identifiable, or*

(B) *if b is known and* $u_0(x)$ *is not constant, then a is identifiable.*

Next, we give two results dealing with the case of $g_i(t) \neq 0$ $(i = 0, 1)$ or $f(x, t) \neq 0$ in (1) and (2).

RESULT 12. *Let* $u_0(x) = 0$ *and* $f(x, t) = 0$ *in* (1)–(3). *Let* $g_0(t)$ *and* $g_1(t)$ *belong to* $L^1[0, \infty)$ *(or to the class of Laplace-transformable functions), and*

(i) $g_0(t) \not\equiv 0$ *and* $g_1(t) \equiv 0$,

(ii) $g_0(t) \equiv 0$ *and* $g_1(t) \not\equiv 0$ *or*

(iii) $g_0(t) \not\equiv 0$ *and* $g_1(t) = \beta g_0(t)$ *where* β *satisfies the inequality* $(\psi_n(1) - \psi_n'(1)) \neq \beta(\psi_n(0) + \psi_n'(0))$ *for all* $n \in N$. *Then, parameters a and b are identifiable if and only if*

$$x_p \in S_1 = \cup\{P_{n,k} : n \neq k \text{ and } (n, k) \in N \times N\}.$$

Moreover, if $(\alpha_0, \alpha_1) \neq (0, 0)$ *and one parameter is known, then the other is identifiable if and only if*

$$x_p \in R_1 = \cup\{P_n : n \in N\}.$$

If $(\alpha_0, \alpha_1) = (0, 0)$, *then*

(A) *if a is known, then b is identifiable if and only if* $x_p \in R_1$, *or*

(B) *if b is known, then a is identifiable if and only if* $x_p \in \cup\{P_n : n \in N - \{1\}\}$.

Proof. First we give

LEMMA 4. *Let* $\psi_n(x)$ *be the n-th eigenfunction in Lemma 2. Then,* $\psi_n(0) + \psi_n'(0) \neq 0$ *and* $\psi_n(1) - \psi_n'(1) \neq 0$ *for every* $n = 1, 2, \cdots$.

Proof of Lemma 4. See Appendix C. We shall prove only the case (i) of Result 12. Under the assumptions, difference of outputs is given by (27) as

$$e(t) = -\int_0^t \left(\sum_{n=1}^{\infty} (\psi_n(0) + \psi_n'(0))\psi_n(x_p)(e^{-k_n(t-\tau)} - e^{-k_n^m(t-\tau)}) \right) g_0(\tau)\, d\tau.$$

Let $e(t) \equiv 0$. Since $g_0(\cdot) \in L^1[0, \infty)$, the Laplace transformation of the above equation yields for all $s \geqq 0$

$$\sum_{n=1}^{\infty} (\psi_n(0) + \psi_n'(0))\psi_n(x_p)(k_n - k_n^m) \frac{1}{(s + k_n)(s + k_n^m)} \cdot \mathcal{L}[g_0(t)] = 0.$$

Note that $g_0(t)$ is not identically zero; hence, $\mathcal{L}[g_0(t)] \not\equiv 0$. Then there exists an s-interval $J \subset [0, \infty)$ such that $\mathcal{L}[g_0(t)] \neq 0$ for $s \in J$. Since $\mathcal{L}[g_0(t)]$ is an analytic function on $(0, \infty)$, the set $\{s > 0 : \mathcal{L}[g_0(t)] \neq 0\}$ is a dense, open subset in $(0, \infty)$. Consequently, $\mathcal{L}[g_0(t)] \neq 0$ for almost every $s \in (0, \infty)$. Then we have

$$\sum_{n=1}^{\infty} (\psi_n(0) + \psi_n'(0))\psi_n(x_p)(k_n - k_n^m) \frac{1}{(s + k_n)(s + k_n^m)} = 0 \tag{30}$$

for almost every $s \in (0, \infty)$, and further for all $s \in [0, \infty)$. Since the function on the lefthand side of (30) is a regular function of the complex variable s, (30) holds for

all $s \in \{z \in C: \operatorname{Re} z \geqq 0\}$. Applying the Laplace inverse transformation to (30), we have

$$\sum_{n=1}^{\infty} (\psi_n(0)+\psi_n'(0))\psi_n(x_p)(e^{-k_n t}-e^{-k_n^m t}) = 0$$

for every $t \geqq 0$. Since $\psi_n(0)+\psi_n'(0) \neq 0$ for all $n \in N$ by Lemma 4, we obtain from $x_p \in S_1$ that $(\psi_i(0)+\psi_i'(0))\psi_i(x_p) \neq 0$ and $(\psi_j(0)+\psi_j'(0))\psi_j(x_p) \neq 0$ for some i and $j (i \neq j)$. This implies from Lemma 3 that $k_i = k_i^m$ and $k_j = k_j^m$ for $i \neq j$. Thus, $a = a_m$ and $b = b_m$. The proof of the necessity is similar to that for the Result 10. Q.E.D.

RESULT 13. *Let* $u_0(x) = 0$, $g_0(t) = 0$ *and* $g_1(t) = 0$ *in* (1)–(3). *Let the input function be* $f(x, t) = f_1(x)f_2(t)$, *where* $f_2(t)$ *is not identically zero and* $f_2(\cdot) \in L^1[0, \infty)$ (*or* $f_2(\cdot)$ *belongs to the class of Laplace transformable functions*). *Two parameters a and b are identifiable if and only if* $(x_p, f_1) \in S$. *Moreover, if* $(\alpha_0, \alpha_1) \neq (0, 0)$ *and one parameter is known, then the other is identifiable if and only if* $(x_p, f_1) \in R$. *If* $(\alpha_0, \alpha_1) = (0, 0)$, *then*

(A) *if a is known, then b is identifiable if and only if* $(x_p, f_1) \in R$, *or*

(B) *if b is known, then a is identifiable if and only if* $(x_p, f_1) \in \cup\{P_n \times Q_n: n \in N-\{1\}\}$.

Proof.

$$\begin{aligned} y(t)-y_m(t) &= \int_0^t \left(\int_0^1 \sum_{n=1}^{\infty} e^{-k_n(t-\tau)}\psi_n(x_p)\psi_n(y)f_1(y)\,dy\right)f_2(\tau)\,d\tau \\ &\quad -\int_0^t \left(\int_0^1 \sum_{n=1}^{\infty} e^{-k_n^m(t-\tau)}\psi_n(x_p)\psi_n(y)f_1(y)\,dy\right)f_2(\tau)\,d\tau \\ &= \sum_{n=1}^{\infty} \psi_n(x_p)(f_1, \psi_n)\int_0^t (e^{-k_n(t-\tau)}-e^{-k_n^m(t-\tau)})f_2(\tau)\,d\tau. \end{aligned}$$

Let $y(t) \equiv y_m(t)$. Since $f_2(\cdot) \in L^1[0, \infty)$, we obtain by applying the Laplance transformation

$$\sum_{n=1}^{\infty} \psi_n(x_p)(f_1, \psi_n) \cdot \mathscr{L}[e^{-k_n t}-e^{-k_n^m t}] \cdot \mathscr{L}[f_2(t)] \equiv 0.$$

Since $f_2(t)$ is not identically zero, we obtain by a similar process as in the proof of Result 12

$$\sum_{n=1}^{\infty} \psi_n(x_p)(f_1, \psi_n)(e^{-k_n t}-e^{-k_n^m t}) \equiv 0.$$

The rest of the proof is similar to that for Result 10. Q.E.D.

Interpretations. Let us briefly discuss the relation between the parameter identifiability and the observability of the system state. The condition $x_p \in P_n$ in Results 10 and 13, where P_n is defined by (28), implies that the nth mode of the state is observable [7], [11, p. 132]. As shown in these results, it is necessary for the identifiability of two constant parameters that at least two modes are observable. The condition $x_p \in P_0$, where P_0 is defined by (29), implies that all modes are observable, that is, complete state observability. $x_p \in P_0$ means, of course, $x_p \in P_n$; however, note that complete observability is not necessarily required for the

identifiability of constant parameters. On the other hand, the initial function of the system must a priori be known for the parameter identifiability in Result 10. Actually, if the initial function is unknown, there exist cases where the parameter identifiability does not follow even if $x_p \in P_0$. Let us show this.

Example 4. First, note that $\sin \pi x = 2 \sin 3\pi x$ has two nonzero roots in $[0, 1]$ which are irrational numbers. Take one of these roots as measurement point x_p', and consider the system and model in Example 3 with $b = b_m = 0$. If $u_0(x) = \sin \pi x$ for the system and $u_{m0}(x) = 2 \sin 3\pi x$ for the model as initial functions, the solutions are given by

$$u(x, t) = e^{-a\pi^2 t} \sin \pi x, \qquad u_m(x, t) = e^{-9a_m \pi^2 t} 2 \sin 3\pi x.$$

Put $x = x_p'$, then $u(x_p', t) = u_m(x_p', t) \neq 0$ for $a = 9a_m$ and for all $t \geqq 0$.

We can see from this example that it would be impractical in applications to require, for the pointwise measurement, that the initial function be known or zero as in Results 10 and 13. However, this is in fact not so inconvenient from the engineering point of view. In the identification process in Fig. 1, normally, the time-varying input function ($f(x, t)$ or $g_i(t)$) will be required to identify the parameters exactly since the algorithm which makes the difference e tend to zero is in many cases slowly converging. Under such circumstances, the response due to the initial function dies out gradually and its influence can be neglected for sufficiently large time, i.e., the response to a nonzero initial function is approximately equal to that with zero initial function. Results 12 and 13 apply to these cases.

5. Conclusion. In this paper, the identifiability problem of the parameters in the distributed system described by a linear, 1-dimensional, parabolic partial differential equation is studied. The identifiability is defined, for the identification process using a model, as the uniqueness of the parameters determined using only the form of the system equation and the input-output data.

Several results for the parameter identifiability and nonidentifiability are presented. For the case of distributed measurements, the conditions for the identifiability depend on the profile of the state of the model (or the state of the system). In particular, the results in § 3 include the result by Chavent [3, p. 100] although only the one-dimensional case has been treated here. For the case of a pointwise measurement, the identifiability conditions depend on the position of a detector and the form of the initial function or the input functions to the system. The relation between the identifiability and the observability is also discussed and the results are related to the N-mode observability [7].

The definition of the parameter identifiability in § 2, and correspondingly the results in § 3, may be weaker than those obtained for lumped systems (e.g. [1]). However, it seems to be straightforward to extend the results in this paper to the multi-dimensional systems. For practical applications, it will be important to take into account the observation errors. The stochastic approach to the identifiability problem (e.g. [1]) may be used for distributed systems, too.

Appendix A. Proof of Lemma 1. Necessity of the statement is self-evident from the following equation (A.1). Sufficiency we obtain from (1), (5) and (7).

$$\begin{aligned}\frac{\partial e}{\partial t} &= \frac{\partial}{\partial x}\left(a\frac{\partial e}{\partial x}\right) + be + \frac{\partial}{\partial x}\left\{(a-a_m)\frac{\partial u_m}{\partial x}\right\} + (b-b_m)u_m \\ &= \frac{\partial}{\partial x}\left(a\frac{\partial e}{\partial x}\right) + be \quad \text{for all } x \in (0,1) \text{ and } t>0.\end{aligned} \tag{A.1}$$

The initial condition for (A.1) is given by $e(0) = u(0) - u_m(0) = 0$ and the boundary condition similar to (2) with $g_0(t) = g_1(t) = 0$. Thus, due to the uniqueness of the solution, we obtain $e(x, t) = 0$ for all $x \in [0, 1]$ and $t \geqq 0$. Q.E.D.

Appendix B. Derivation of (19) and (20). $e = u - u_m$ satisfies

$$\begin{aligned}&\frac{\partial e}{\partial t} = \frac{\partial}{\partial x}\left(a\frac{\partial e}{\partial x}\right) + \frac{\partial}{\partial x}\left(q\frac{\partial u_m}{\partial x}\right), \\ &e(0,t) = \frac{\partial e}{\partial x}(1,t) = 0\end{aligned} \tag{B.1}$$

where $q = a(x) - a_m(x, t)$. Define a Lyapunov functional by

$$V = \frac{1}{2}\int_0^1 e^2(x,t)\,dx + \frac{K}{2}\int_0^1 q^2(x,t)\,dx$$

where K is a positive constant; then

$$\begin{aligned}\dot{V}_{(B.1)} = &\left[(a+q)\,e\frac{\partial e}{\partial x}\right]_0^1 - \int_0^1 a(x)\left(\frac{\partial e}{\partial x}\right)^2 dx \\ &- \int_0^1 q\left\{K\frac{\partial a_m}{\partial t} + \frac{\partial e}{\partial x}\frac{\partial u_m}{\partial x}\right\} dx.\end{aligned} \tag{B.2}$$

The first term in the above equation vanishes by the boundary condition and the third term by (19). For the second term, we have

$$\int_0^1 a(x)\left(\frac{\partial e}{\partial x}\right)^2 dx \geqq \int_0^1 \alpha\left(\frac{\partial e}{\partial x}\right)^2 dx \geqq \alpha\left(\frac{\pi}{2}\right)^2 \int_0^1 e^2\,dx \geqq 0$$

where $a(x) \geqq \alpha > 0$ (α: positive constant) and Wirtinger's inequality [4, p. 79] are used. Thus, $\dot{V}$ is negative semi-definite. Further, the condition of (13) in Result 2 is always satisfied by the Neumann-type boundary condition in (18), and (14) holds for sufficiently smooth $a(x)$ and $a_m(x, t)$. Thus, the identifiability from Result 2 guarantees $q = 0$ for all x and t when $\dot{V} = 0$ for all x and t. Then, the system (B.1) and (19) is asymptotically stable with respect to L^2 norm [11, p. 107].

Appendix C. Proof of Lemma 4. If $(\alpha_0, \alpha_1) = (0, 0)$ and $n = 1$ we have

$$\psi_1(0) + \psi_1'(0) = 1 \quad \text{and} \quad \psi_1(1) - \psi_1'(1) = \cos(n-1)\pi \neq 0.$$

For other cases, $\psi_n(0) + \psi_n'(0) = \lambda_n \neq 0$ by Lemma 2. To show the latter relation,

we use a contradiction. Assume that $\psi_n(1)-\psi_n'(1)=0$; then

$$\text{(C.1)} \qquad \psi_n(1)-\psi_n'(1)=(\alpha_0+(1-\alpha_0)\lambda_n^2)\sin\lambda_n+(1-2\alpha_0)\lambda_n\cos\lambda_n=0.$$

By Lemma 2, λ_n satisfies the following equation.

$$\text{(C.2)} \qquad \{(1-\alpha_0)(1-\alpha_1)\lambda_n^2-\alpha_0\alpha_1\}\sin\lambda_n-(\alpha_0+\alpha_1-2\alpha_0\alpha_1)\lambda_n\cos\lambda_n=0.$$

Since $\sin\lambda_n$ and $\cos\lambda_n$ do not vanish simultaneously, the coefficient determinant, D, of (C.1) and (C.2) must be zero, i.e.,

$$-\frac{D}{\lambda_n}=(\alpha_0+\alpha_1-2\alpha_0\alpha_1)(\alpha_0+(1-\alpha_0)\lambda_n^2)+(1-2\alpha_0)((1-\alpha_0)(1-\alpha_1)\lambda_n^2-\alpha_0\alpha_1).$$

$$=\alpha_0+(1-\alpha_0)^2\lambda_n^2=0.$$

This yields a contradiction. Hence, $\psi_n(1)-\psi_n'(1)\neq 0$ for all $n\in N$. Q.E.D.

Acknowledgment. One of the authors, S. Kitamura would like to thank the Alexander von Humboldt Foundation who gave him an opportunity to stay at the University of Stuttgart.

REFERENCES

[1] M. AOKI AND P. C. YUE, *On certain convergence questions in system identification*, this Journal, 8 (1970), pp. 239–256.

[2] K. J. ÅSTROM, *System Identification, Stability of Stochastic Dynamical Systems*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[3] M. G. CHAVENT, *Anàlyse fonctionele et identification de coefficients répartis dans les equations aux dérivées partielles*, Thesis, Faculté des Sciences de Paris, 1971.

[4] J. B. DIAZ AND F. T. METCALF, *Variations of Wirtinger's inequality*, Inequality, O. Shisha, ed., Academic Press, New York, London, 1967, pp. 73–77.

[5] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, N.J., 1964.

[6] R. E. GOODSON AND M. P. POLIS, *Parameter identification in distributed systems; a synthesizing overview*, Identification of Parameters in Distributed Systems, American Society of Mechanical Engineers, New York, 1974, pp. 1–30.

[7] R. E. GOODSON AND R. E. KLEIN, *A definition and some results for distributed system observability*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 165–174.

[8] B. REISSENWEBER, *Parameteridentifikation bei örtlich verteilten Systemen mit Hilfe von bewegten Sonden*, Dissertation, Fakultät für Elektrotechnik, Universität Karlsruhe, Germany, 1975.

[9] J. H. SEINFELD, *Identification of parameters in partial differential equations*, Chem. Engrg. Sci., 24 (1969), pp. 65–74.

[10] J. H. SEINFELD AND W. H. CHEN, *Estimation of parameters in distributed systems*, Identification of Parameters in Distributed Systems, American Society of Mechanical Engineers, New York, 1974, pp. 69–89.

[11] P. K. C. WANG, *Control of distributed parameter systems*, Advances in Control Systems, vol. 1, C. T. Leondes, ed., Academic Press, New York, London, 1964, 75–172.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

ON THE SET OF ATTAINABILITY OF NONLINEAR NONAUTONOMOUS CONTROL SYSTEMS*

D. REBHUHN†

Abstract. An important tool in optimal control theory is the Pontryagin maximum principle. A necessary condition for optimality, the principle is an analytic description of a control whose response stays in the boundary of the attainable set. See [2], [23]. It is useful to know that the attainable set has a nonempty interior because the maximum principle gives no information otherwise.

If we consider nonlinear, nonautonomous control systems determined by the set $\mathscr{A}^k$ of time dependent C^k controllable vector fields on the m-dimensional manifold M in the Whitney C^k topology and if $k \geqq 2m+2$, then a few technical restrictions on the control system assure us that there is an open dense subset $\mathscr{O}$ of $\mathscr{A}^k$ such that for each $F \in \mathscr{O}$ the attainable set is contained in the closure of its own interior.

1. Introduction. In this section we give some basic definitions and connect the properties of a control system with the properties of its attainable sets.

DEFINITION 1.1. By a *nonautonomous C^k time optimal control system*, we mean a collection $(M, \Omega, \sigma, \mathscr{U}, F)$ such that:

(i) M is a finite dimensional, Hausdorff, connected, second countable manifold.

(ii) Ω is a Hausdorff topological space.

(iii) $\sigma: M \times \mathbb{R} \to 2^{\Omega}$ is a tracer function that assigns a subset of Ω to every point of $M \times \mathbb{R}$.

(iv) $\mathscr{U}$, the set of controls, is an admissible set of regulated paths in Ω. See [2] for definitions of these terms. It will be sufficient for our purposes to consider $\mathscr{U}$ to be the set of piecewise constant paths in Ω.

(v) F is a C^k controllable vector field parametrized by Ω, that is,

$$F: M \times \mathbb{R} \times \Omega \to TN,$$

$$(x, t, w) \to F(x, t, w) \in T_x M$$

such that the derivatives of F up to order k of F in (x, t) exist and are jointly continuous in (x, t, w).

DEFINITION 1.2. Let x', x'' be elements of M. Let $u \in \mathscr{U}$ be a control. Let $I_u = [t', t'']$ be an interval in $\mathbb{R}$. If there is a path $x_u: I_u \to M$ such that

(i) $x_u(t') = x'$,

(ii) $x_u(t'') = x''$,

(iii) $(d/dt)x_u(t) = F(x_u(t), t, u(t))$ and $u(t) \in \sigma(x_u(t), t)$ for all except countably many points of $[t', t'']$,

then we say that *u steers x' to x''*. The path x_u is called an *M response* of u. There may be no controls or there may be many controls steering x' to x''. We also say that (x'', t'') is *attainable from* (x', t') in the positive time $(t''-t')$.

DEFINITION 1.3. A control $u: I_u = [t', t] \to \Omega$ steering x' to x'' will be called *time optimal* if $\int_{t'}^{t} 1\, dt = t - t'$ is a minimum over all controls starting at time t' and

* Received by the editors August 14, 1975, and in final revised form October 12, 1976.

† Department of Mathematics, Vassar College, Poughkeepsie, New York 12601.

steering x' to x''. It is possible to replace 1 by a function $f: M \times \mathbb{R} \times \Omega \to \mathbb{R}$ and to try to minimize $\int_{t'}^{t} f(x_u(t), t, u(t))\, dt$.

DEFINITION 1.4. We will call a tracer function *admissible* if it admits at least two locally constant sections in a neighborhood of every point. By a section, we mean a map $s: M \times \mathbb{R} \to \Omega$ such that $s(x, t) \in \sigma(x, t)$.

DEFINITION 1.5. A control u' will be called *acceptable at* $(x', t') \in M \times \mathbb{R}$ if $u'(T) \in \sigma(x_{u'}(T), T)$ for all $T \geqq t'$. For each such u', we will assume the existence of an acceptable C^k vector field X with flow ν defined on an open subset of $M \times \mathbb{R}$ such that:

(i) Whenever τ is in the domain of $x_{u'}$ then $\nu((x', t'), \tau) = (x_{u'}(\tau), t' + \tau)$.

(ii) If (x, t) is in the domain of X, then there is a control u acceptable at (x, t) such that $\nu((x, t), \tau) = (x_u(\tau), t + \tau)$ for $\tau > 0$ in the domain of u.

For example, if the tracer function σ is constant, for any point $u \in \sigma(x, t)$, the vector field

$$\begin{pmatrix} F(x, t, u) \\ 1 \end{pmatrix}$$

is acceptable. If $\sigma(x, t)$ varies wildly from point to point, such a simple vector field may not be acceptable. See [22]. Condition (ii) guarantees that if (x'', t'') is attainable from (x', t') and if (x', t') is attainable from (x, t) with $t'' > t' > t$, then (x'', t'') is also attainable from (x, t). Conditions (i) and (ii) are geometric requirements for the existence of a maximum principle. See [15], [22] for more details.

If $\{X^i\}_{i \in \mathcal{I}}$ are the acceptable vector fields on $N = M \times \mathbb{R}$ with associated flows ν^i and if $i_1, i_2, \cdots, i_{q \in \mathcal{I}}$ and $t_1, t_2, \cdots, t_q \in \mathbb{R}^+$, we get a mapping from an open subset of N into N defined by

$$y \in N \to \nu^{i_q}(\cdots(\nu^{i_2}(\nu^{i_1}(y, t_1), t_2) \cdots, t_q))$$

when the composition exists. If we let the q tuples $(i_1, i_2, \cdots, i_q)$ and $(t_1, t_2, \cdots, t_q)$ be represented by I and T respectively, then we will use the symbol $\nu_T^I(y)$ to represent the image of y under this mapping.

DEFINITION 1.6. The $\{X^i\}_{i \in \mathcal{I}}$ are said to define a *local polydynamical system* or P.D.S. on $N = M \times \mathbb{R}$. Each X^i is an element of the P.D.S., and the P.D.S. is generated by its elements. This P.D.S. is called the P.D.S. associated to the time optimal control system $(M, \Omega, \sigma, \mathcal{U}, F)$.

DEFINITION 1.7. Consider the space $\mathcal{U}$ of local vector fields on N spanned by the $\{X^i\}_{i \in \mathcal{I}}$ and their successive Lie brackets whenever those are defined. If the $\{X^i\}$ were global and C^∞, this space would be the Lie algebra they generate. If we evaluate each element of $\mathcal{U}$ at $y \in N$, we get generators for a subspace $\mathcal{U}(y)$ of $T_y N$. The P.D.S. will be called *regular* at y if $\dim \mathcal{U}(y) = \dim N$. If the P.D.S. is regular at each point $y \in N$, we will call it a *regular* P.D.S.

MAIN RESULT 1.8. If σ is admissible, then the set of control systems whose associated P.D.S. is regular is open and dense for a topology which will be described in detail further on. See § 2.

The interest of this result in control theory stems in part from Chow's theorem. Given $y \in N$, let $G_y[0, \delta)$ be the set of all points attainable from y in any time $0 \leqq t < \delta$. Chow's theorem says if the P.D.S. associated to a control system is

regular at $y \in N$, then $G_y[0, \delta)$ contains y in the closure of its interior. See Krener [12] or Lobry [15] for a proof of this theorem.

It is only when a system has a regular P.D.S. that the maximum principle is of any use in choosing optimal controls. See [3], [14], [2], [23] for information about the maximum principle. The main result is related to work of Lobry [15], [16] who gets a similar result for simple autonomous control systems. Chow's theorem has been used in control theory by Hermann [8], Hermes [9], Jurdjevic and Sussmann [10]. Sussmann calls regularity the "accessibility property" and he uses it in realization theory. See [25].

When the attainable set is closed, the maximum principle actually picks out an optimal control. It is not always true that the attainable set is closed. See [6]. Indeed, the situation of [6] is stable under perturbation of the control system. Gronski [7] has classified closed sets of attainability in the plane and the author [21] has some results on existence of closed sets of attainability in the plane. Krener [11] has done some related work which has not yet been published.

2. The main theorem. Consider the set of all C^k controllable vector fields on the manifold M. Since these vector fields are not differentiable in all variables, we cannot consider them as topologized by the standard Whitney C^k topology. Instead, we take our "k-jet bundle" to involve derivatives in M and $\mathbb{R}$ only. This defines an analogous topology which we will call the C^k topology. We will use the symbol $\mathscr{A}^k$ to denote the set of vector fields in this topology. Note that $\mathscr{A}^k$ is a Baire space. See [18] or [17] for details.

From now on, we fix M, Ω, σ, $\mathscr{U}$ and assume that σ is admissible.

THEOREM 2.1. *Let $m = \dim M$ and let $k \geqq 2m+2$ be a finite integer. There is an open dense subset $\mathscr{O}$ of $\mathscr{A}^k$ such that, if $F \in \mathscr{O}$, then the P.D.S. associated to $(M, \Omega, \sigma, \mathscr{U}, F)$ is regular.*

Proof. Since σ is admissible, for any (x', t') in $M \times \mathbb{R}$ we have an open neighborhood V' of (x', t') and distinct w_1, w_2 in Ω such that $w_1, w_2 \in \bigcap \{\sigma(x, t)\colon (x, t) \in V'\}$. We wish to show that there is an open dense subset $\mathscr{O}'$ of $\mathscr{A}^k$ such that if $F \in \mathscr{O}'$, the vector fields X^1, X^2 defined by

$$X^1(x, t) = \begin{pmatrix} F(x, t, w_1) \\ 1 \end{pmatrix}, \qquad X^2(x, t) = \begin{pmatrix} F(x, t, w_2) \\ 1 \end{pmatrix}$$

generate a regular P.D.S. on a neighborhood $V \subset V'$ of (x', t'). Second countability of $M \times \mathbb{R}$ will then give us a residual subset of $\mathscr{A}^k$ whose elements have regular associated P.D.S.'s. It will follow from the construction of $\mathscr{O}'$ that our residual set is actually open and dense.

There is no loss of generality in assuming V', our neighborhood of (x', t'), is relatively compact and diffeomorphic to a relatively compact open subset of $\mathbb{R}^{m+1}$. We can describe $(x, t) \in V'$ by coordinates $(x_1, x_2, \cdots, x_m, t)$. Let W' be a relatively compact neighborhood of w_1 and w_2 that is homeomorphic to a relatively compact open subset of $\mathbb{R}^p$. We do not ask that W' be connected. Indeed, since $w_1 \neq w_2$, it may not be possible to find one coordinate neighborhood containing both points. We need a neighborhood homeomorphic to a relatively compact subset of $\mathbb{R}^p$ because we will want to apply approximation and extension theorems of Whitney that are stated for Euclidean space.

Let R be the set of $m \times (2m+1)$ matrices. For $0 \leqq j \leqq m-1$, let R^j be the subset of R of matrices of rank j. For each j in the given range, the codimension of R^j is $j(3m+1-j)$ which is greater than $m+1$. See [24] and [22] for further details.

Given $F \in \mathscr{A}^k$, we can define $\rho(F)$ a C^1 map from V' to R as follows:

(i) Let X^1, X^2 be defined as before; define $Y_F^1 = [X^1, X^2]$, and, inductively, define $Y_F^h = [X^1, Y_F^{h-1}]$ for $2 \leqq h \leqq 2n+1$.

(ii) Note that the $(m+1) \times (2m+1)$ matrix

$$[Y_F^1(x,t)\, Y_F^2(x,t) \cdots Y_F^{2m+1}(x,t)]$$

whose columns are the Y_F^h has a bottom row all of whose entries are zero. This follows from the definition of Lie bracket and fact that the last entries in X^1 and X^2 are constant.

(iii) We let

$$\rho(F)(x,t) = [Z_F^1 Z_F^2 \cdots Z_F^{2m+1}](x,t)$$

where the columns of $\rho(F)(x,t)$ are obtained from the columns of $[Y_F^1(x,t) \cdots Y_F^{2m+1}(x,t)]$ by dropping the bottom zeros.

When the rank of $\rho(F)(x,t)$ equals m, the P.D.S. generated by X^1 and X^2 is regular at (x,t). The rank of $\rho(F)(x,t)$ will equal m unless $\rho(F)(x,t) \in R^j$ for $0 \leqq j \leqq m-1$. Note that R^j is a manifold and that the codimension of R^j is greater than $m+1$ for each j. Thus, $\rho(F)(x,t) \notin R^j$ if and only if $\rho(F)$ is transversal to R^j at (x,t). In other words, our problem reduces to finding an open dense subset $\mathscr{O}'$ of $\mathscr{A}^k$ and a neighborhood V'' of (x,t) such that if $F \in \mathscr{O}'$ then $\rho(F)$ is transversal to each of the R^j on V''.

For the definition of transversality or details on transversal mappings, see [1] or [19].

We will modify the space we are considering so that we can apply a transversality theorem from [1].

Let V, W be open subsets of $M \times \mathbb{R}$ and Ω respectively such that $\{w_1, w_2\} \subset W \subset \mathrm{cl}(W) \subset W'$ and $(x', t') \in V \subset \mathrm{cl}(V) \subset V'$. Here $\mathrm{cl}(W)$ and $\mathrm{cl}(V)$ indicate the closures of W and V respectively.

We can get a modified Whitney C^k norm on $F|_{V \times W} \in \mathscr{A}^k|_{V \times W}$ by setting

$$\|F\|^k_{V \times W} = \sup\left\{ \sum \left| \frac{\partial^i F^j}{\partial x_1^{i_1} \cdots \partial x_{m+1}^{i_{m+1}}}(v,w) \right| : (v,w) \in V \times W, \right.$$
$$\left. 1 \leqq j \leqq q,\ 0 \leqq i \leqq k,\ i_1 + \cdots + i_{m+1} = i \right\}.$$

For k finite, since $\mathrm{cl}(V \times W)$ is compact in $M \times \mathbb{R} \times \Omega$, any $F \in \mathscr{A}^k|_{V \times W}$ has a finite norm. Furthermore, if we consider $\mathscr{A}^k|_{V \times W}$ in the topology induced by this norm, we can identify $\mathscr{A}^k|_{V \times W}$ with $\mathscr{A}^k|_{\mathrm{cl}(V \times W)}$.

We could *not* identify these two spaces if we were using the topology of uniform convergence on compact sets. The identification can be made because the modified C^k sup norm on $V \times W$ is the same as the modified C^k sup norm on $\mathrm{cl}(V \times W)$ for any element of $\mathscr{A}^k$. From now on, we will use the symbol $\mathscr{B}^k$ to denote $\mathscr{A}^k|_{V \times W} \cong \mathscr{A}^k|_{\mathrm{cl}(V \times W)}$ in the induced topology.

Note that the natural restriction map $r: \mathscr{A}^k \to \mathscr{B}^k$ sending $F \in \mathscr{A}^k$ to $r(F) = F|_{V \times W}$ (or to $F|_{\mathrm{cl}(V \times W)}$) $\in \mathscr{B}^k$ is a continuous open surjection. Thus, given any open

dense subset $\mathcal{O}''$ of $\mathcal{B}^k$, there is a corresponding open dense subset $\mathcal{O}' = r^{-1}(\mathcal{O}'')$ of $\mathcal{A}^k$. If we can find an open dense subset $\mathcal{O}''$ of $\mathcal{B}^k$ whose image under ρ is transversal to R^j on a fixed neighborhood of (x', t') in V, then $r^{-1}(\mathcal{O}'')$ is the open dense subset of $\mathcal{A}^k$ that we are looking for.

LEMMA 2.2. *$\mathcal{B}^k$ is a Banach space.*

Proof. Since $\mathcal{B}^k$ is a normed space, we will be done if we can show $\mathcal{B}^k$ is complete.

We identify $V' \times W'$ with an open, relatively compact subset of $\mathbb{R}^{m+1+p}$. Thus we automatically have an identification of $V \times W$ with an open subset of $\mathbb{R}^{m+1+p}$ such that $\text{cl}(V \times W) \subset V' \times W'$. By a result of Whitney [27], a function defined on a closed subset cl $(V \times W)$ of $\mathbb{R}^{m+1+p}$ can be extended to a differentiable function on an open subset of cl $(V \times W)$ if and only if the function on the closed set is differentiable in the sense of Whitney.

We observe that a Cauchy sequence of functions in $\mathcal{B}^k$, by our choice of norms, must converge to a limit function which is differentiable in the sense of Whitney on cl $(V \times W)$.

Thus if F is the limit of a Cauchy sequence in $\mathcal{B}^k$, then F must be an element of $\mathcal{B}^k$ and the lemma is proven.

We have reduced the problem to finding a neighborhood V'' of (x', t') and an open subset $\mathcal{O}''$ of $\mathcal{B}^k$ such that if $F \in \mathcal{O}''$, then $\rho(F)$ is transversal to the R^j on V''. This would guarantee that the P.D.S. associated to F is regular on V'' whenever $F \in \mathcal{O}''$.

Let V'' be an open neighborhood of (x', t') such that cl $(V'') \subset V$.

We will demonstrate the existence of $\mathcal{O}''$ in two steps.

(i) We will show that for each j, $0 \leqq j \leqq m-1$, the set $\mathcal{O}^j = \{F \in \mathcal{B}^k : \rho(F)$ is transversal to R^j on $V\}$ is a residual subset of $\mathcal{B}^k$.

(ii) We will show that the set $\mathcal{O}'' = \{F \in \mathcal{B}^k : \rho(F)$ is transversal to *all* of the R^j on cl $(V'')\}$ is an open subset of $\mathcal{B}^k$.

Note that $\mathcal{O}''$ contains the intersection of all the $\mathcal{O}^j$ for $0 \leqq j \leqq m-1$. Since $\mathcal{B}^k$ is a Banach space, this shows us that $\mathcal{O}''$ is open and contains a dense subset of $\mathcal{B}^k$.

Let us now demonstrate (i). By the transversal density theorem of [1], the set $\mathcal{O}^j$ of elements F of $\mathcal{B}^k$ such that $\rho(F)$ is transversal to R^j on V is residual in $\mathcal{B}^k$ if:

(a) $\mathcal{B}^k$ and V are second countable;

(b) ρ is a C^s representation for some $s > \max(0, \dim V - \text{codim } R^j)$;

(c) ev_ρ is transversal to R^j on V for each j between zero and $m-1$.

Recall that if $\mathcal{D}$ is a Hausdorff manifold, and $\mathcal{C}^s(M, N)$ is the set of C^s maps from M into N, then $\rho: \mathcal{D} \to \mathcal{C}^s(M, N)$ is a C^s representation if the map $ev_\rho: \mathcal{D} \times M \to N$ defined by $ev_\rho(x, y) = \rho(x)(y)$ is a C^s map.

In particular, if $\mathcal{D} = \mathcal{B}^k$ and ρ is the map we defined earlier, then ρ is a representation of $\mathcal{B}^k$ in $\mathcal{C}^1(V, R)$.

Since V is an open subset of $\mathbb{R}^{m+1}$, it is clear that V is second countable. We must check three things, that $\mathcal{B}^k$ is second countable, that ρ is a C^1 representation (remember, dim $V - \text{codim } R^j < 0$), and that ev_ρ is transversal to R^j on V.

LEMMA 2.3. *$\mathcal{B}^k$ is second countable.*

Proof. To show that $\mathcal{B}^k$ is second countable, we give coordinate representations of its elements in some chart. We then use the Weierstrass approximation theorem (see [19] or [27]) to show that the elements of $\mathcal{B}^k$ can be approximated as

closely as we wish by functions whose coordinates are polynomials with rational coefficients.

LEMMA 2.4. *The map ρ is a C^1 representation.*

Proof. Because $k \geqq 2m+2$, any map of the kind that sends $(F, (x, t)) \in \mathscr{B}^k \times V$ to $D^j_{(x,t)} F(x, t, w_i)$ is at least C^1 for $i = 1, 2, j = 1, 2, \cdots, 2m+1$. Here, D^j indicates the jth derivative of F with respect to (x, t). For verification that these maps are C^1, see [1, § 10].

The columns of $\rho(F)(x, t)$, namely $Z^1_F(x, t), \cdots, Z^{2m+1}_F(x, t)$ are analytic functions in the coordinate entries of the $D^j_{(x,t)}F$. This makes ev_ρ the composition of an analytic function with a C^1 function. The map ρ is a C^1 representation and the lemma is proven.

LEMMA 2.5. *The map ev_ρ is transversal to R^j on V.*

Proof. Note that the space $\mathscr{B}^k \times V$ is the product of a Banach space with an open subset of $\mathbb{R}^{m+1}$. Thus the tangent space to $\mathscr{B}^k \times V$ at any point is just $\mathscr{B}^k \times \mathbb{R}^{m+1}$. Similarly, the tangent space to R at a point is R itself.

We will prove the lemma by showing that if $(F, (x, t)) \in \mathscr{B}^k \times V$, then $D_{(F,(x,t))} ev_\rho$ is a split surjection.

We will show that $D_{(F,(x,t))} ev_\rho$ is a surjection by constructing C^1 paths $\beta^{j,\ell}: \mathbb{R} \to \mathscr{B}^k \times V$ such that $\beta^{j,\ell}(0) = (F, (x, t))$ and such that

$$\frac{d}{ds}\bigg|_{s=0} ev_\rho \beta^{j,\ell}(s) = \gamma^{j,\ell}$$

where the $\gamma^{j,\ell}$ span R. Indeed $\gamma^{j,\ell}$ will be an $m \times (2m+1)$ matrix with all columns before the ℓth column identically zero. The ℓth column itself will have one nonzero entry—precisely in the jth row. It is an elementary exercise to check that the $\gamma^{j,\ell}$ span R.

Let Φ: $M \times \mathbb{R} \to \mathbb{R}$ be a C^∞ function with compact support in V that is identical to one on a neighborhood of (x, t).

Let $\Phi' = \Omega \to \mathbb{R}$ be a continuous function with compact support in W such that Φ' is identical to one on a neighborhood of w_1 and identical to zero on a neighborhood of w_2. If

$$F(x'', t'', w'') = \begin{pmatrix} F_1(x'', t'', w'') \\ \vdots \\ F_m(x'', t'', w'') \end{pmatrix},$$

we set $\beta^{j,\ell}(s) = F^{j,\ell}((s), (x, t))$ where

$$F^{j,\ell}(s)(x'', t'', w'') = \begin{pmatrix} F_1(x'', t'', w'') \\ \vdots \\ F_j(x'', t'', w'') + s\Phi(x'', t'')\Phi'(w'')(t''-t)^\ell \\ \vdots \\ F_m(x'', t'', w'') \end{pmatrix}.$$

It is a matter of calculus and the definition of Lie bracket to verify that the ℓth column of $(d/ds)|_{s=0} ev_\rho \beta^{j,\ell}(s)$ is:

$$j\text{th row} \dashrightarrow \begin{pmatrix} 0 \\ \vdots \\ \ell!\Phi(x,t)\Phi(w_1) \\ \vdots \\ 0 \end{pmatrix}.$$

This demonstrates that $D_{(F,(x,t))} ev_\rho$ is a surjection.

Because the image of $D_{(F,(x,t))} ev_\rho$ is finite dimensional, the kernel has finite codimension and splits $\mathscr{B}^k \times \mathbb{R}^{m+1}$. Thus the lemma is proven and we know that ev_ρ is transversal to R^j on V for any $0 \leqq j \leqq m-1$. Thus we know that (i) holds.

It is much easier to see that (ii) holds. If $F \in \mathscr{B}^k$ such that $\rho(F)$ is transversal to the R^j on cl (V''), then $\rho(F)$ cl (V'') does not intersect $\cup\{R^j: 0 \leqq j \leqq m-1\}$ which is a closed set. It follows from the continuity of ev_ρ and the compactness of cl (V'') that there is a $\delta > 0$ such that if $G \in \mathscr{B}^k$ and $\|G-F\|^k < \delta$, then $\rho(G)(\text{ cl } (V''))$ does not intersect $\cup\{R^j: 0 \leqq j \leqq m-1\}$ either.

Thus we have an open dense subset $\mathscr{O}''$ of $\mathscr{B}^k$ whose extensions to $M \times \mathbb{R} \times \Omega$ form an open dense subset of $\mathscr{A}^k$ with associated P.D.S.'s regular on V''. We can thus conclude from prior reasoning that we have a residual subset $\mathscr{O}$ of $\mathscr{A}^k$ whose elements have regular P.D.S.'s associated to them.

From our choice of topology, $\mathscr{O}$ is actually open in $\mathscr{A}^k$ and Theorem 2.1 has been proved.

3. Conclusions. Theorem 2.1 is really a theorem about approximation. If $F \in \mathscr{A}^k$ and the P.D.S. associated to F is not regular, the theorem shows that there is a $G \in \mathscr{A}^k$ arbitrarily close to F in the Whitney C^k topology such that the P.D.S. associated to G is regular.

When $f \not\equiv 1$, that is when we have a general C^k optimal control system, we get a similar result by considering the set $\mathscr{E}^k$ of pairs (f, F) in the Whitney C^k topology. Here f is a C^k cost function, and F is a controllable C^k vector field on M.

THEOREM 3.1. *If σ is an admissible tracer function and if $k \geqq 2m+3$ is a fixed finite integer, then there is an open dense subset of $\mathscr{E}^k$ whose elements have regular associated P.D.S.'s.*

Proof. The proof is almost word for word the same as the proof of Theorem 2.1, the corresponding result for time optimal control systems. We merely substitute $\mathscr{E}^k$ for $\mathscr{A}^k$ and replace R by the space of $(m+1) \times (2m+2)$ matrices.

As a result about approximations, Theorem 3.1 is less satisfactory than Theorem 2.1. In Theorem 3.1, we may have to perturb the pair (f, F) to a pair (g, G) in order to get a regular P.D.S. There is no guarantee that we can get a regular P.D.S. without changing the cost function.

There is a theorem of Lobry connected to Theorems 2.1 and 3.1. See [15]. Lobry considered C^∞ vector fields on a manifold N in the Whitney C^k topology for k finite. He showed, using stratifications, that if k is a sufficiently large integer,

then there is an open dense subset of the set of pairs of such vector fields in the product topology such that the elements of the open dense set generate a regular P.D.S. on N.

Two questions arise naturally at this point. What happens when σ is not admissible, and what happens when F is C^k, but $k \leqq 2m+1$? The second question is, perhaps, the more important of the two. In many important examples, the set $\sigma(x, t)$ is fixed for all $(x, t) \in M \times \mathbb{R}$.

Both [4] and [26] consider controllability and "accessibility" questions. The "accessibility" property in some sense C^0 stable and this is closely related to the last question.

The need for so much differentiability in Theorems 2.1 and 3.1 arises because they are theorems about attaining a set of nonempty interior in a high dimensional manifold using only two fector fields. Clearly, the bigger the dimension of the manifold, the more we must ask of the two vector fields. If we could use more vector fields, we would not need so much differentiability for the controllable vector fields.

DEFINITION 3.2. The tracer function σ will be called *extremely admissible* if, for any $(x', t') \in M \times \mathbb{R}$, there is a neighborhood V of (x', t') and distinct $w_1, \cdots, w_{2m+2} \in \Omega$ such that $w_i \in \cap\{\sigma(x, t)\colon (x, t) \in V\}$, $1 \leqq i \leqq 2m+2$.

Recall that $\mathscr{A}^2$ is the set of controllable C^2 vector fields on M.

PROPOSITION 3.3. *If σ is extremely admissible, then there is an open dense subset of $\mathscr{A}^2$ whose elements each have a regular associated P.D.S.*

Proof. The proof is very like the proof of Theorem 2.1 but the columns of the matrix $\rho(F)(x, t)$ will be replaced by what is obtained when the terminal zero is dropped from

$$Z_F^h(x, t) = \left[\begin{pmatrix} F(x, t, w_h) \\ 1 \end{pmatrix}, \begin{pmatrix} F(x, t, w_{2m+2}) \\ 1 \end{pmatrix}\right].$$

There will also be a corresponding modification of the $\gamma^{i,\ell}$.

Even if σ is not admissible, we would like to know something about approximating the acceptable vector fields of a control system by vector fields on $M \times \mathbb{R}$ that generate a regular P.D.S.

Consider the set of C^k time-dependent vector fields on M in the Whitney C^k topology. Let us denote the set of pairs of such vector fields in the product topology by $\mathscr{T}^k$.

PROPOSITION 3.4. *Let m = dimension M be greater than one. If $k \geqq 2m+2$ is a finite integer, there is an open dense subset of $\mathscr{T}^k$ such that if (X^1, X^2) is any element of the open dense subset, then $\begin{pmatrix} X^1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} X^2 \\ 1 \end{pmatrix}$ generate a regular P.D.S. on $M \times \mathbb{R}$.*

Proof. This follows from the proof of the main theorem (Theorem 2.1). In fact, in the main theorem, things were much more complicated because acceptable vector fields were not necessarily globally defined as they are here.

There is a much simpler and more direct proof of Proposition 3.4 parallel to the work of [15]. We do not include that proof here.

We remark that in some ways the results of this section are unsatisfactory. For example, if $F \in \mathscr{A}^k$ happens to be autonomous, a controllable vector field that

does not depend explicitly on time, it would be of interest to know that we could approximate F arbitrarily closely by another autonomous controllable vector field whose associated P.D.S. is regular. By the results of this section, we could find a nonautonomous approximation to F whose associated P.D.S. was regular, but that is not a strong enough result.

Let $\mathscr{H}^k$ be the set of C^k autonomous controllable vector fields. An attempt to substitute $\mathscr{H}^k$ for $\mathscr{B}^k$ in the proof of the main theorem shows that the proof fails at Lemma 2.5. The paths $\beta^{i,\ell}$ we constructed in Lemma 2.5 lie in $\mathscr{A}^k - \mathscr{H}^k$.

Acknowledgment. I would like to thank Felix Albrecht for his unfailing encouragement and assistance while I worked on this material and to thank the department of mathematics at M.I.T. for the hospitality extended to me while I wrote up this paper for publication.

REFERENCES

[1] R. ABRAHAM AND J. ROBBIN, *Transversal Mappings and Flows*, W. A. Benjamin, New York, 1967.

[2] F. ALBRECHT, *Topics in Control Theory*, Springer-Verlag, Berlin, 1968.

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

[4] P. BRUNOVSKY AND C. LOBRY, *Controllabilité bang bang, controllabilité differentiable et perturbation des systèmes non lineaires*, Ann. Mat. Pura Appl., Sér. 105, 1975, pp. 93–119.

[5] J. DIEUDONNE, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[6] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[7] J. GRONSKI, Thesis, Dept of Math., Univ. of Illinois, Urbana, 1974.

[8] R. HERMANN, *The differential geometry of foliations, II*, J. Math. Mech., 2 (1962), pp. 303–315.

[9] H. HERMES, *Controllability and the singular problem*, this Journal, 2 (1964), pp. 241–260.

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

[11] A. KRENER, *Structural stability of control systems*, Notices Amer. Math. Soc., 22 (1975), p. A-407.

[12] ———, *A generalization of Chow's theorem and the bang bang theorem to nonlinear control systems*, this Journal, 12 (1974), pp. 43–52.

[13] S. LANG, *Introduction to Differentiable Manifolds*, Interscience, New York, 1967.

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

[15] C. LOBRY, *Quelques aspects qualitatifs de la theorie de la commande*, Thèse, L'Université de Grenoble, France, 1972.

[16] ———, *Dynamical polysystems and control theory*, Geometric Methods in System Theory, D. G. Mayne and R. Brockett, eds., D. Reidel, Boston, 1974.

[17] J. N. MATHER, *Stability of C^∞ mappings II*, Ann. of Math., 89 (1969), pp. 254–281.

[18] C. MORLET, *Seminaire Henri Cartan, Topologie differentielle*, École Normal Superieure, Paris, 1961/62.

[19] R. NARASIMHAN, *Analysis on real and complex manifolds*, North-Holland, Amsterdam, 1968.

[20] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Macmillan, New York, 1964.

[21] D. REBHUHN, *On the closure of sets of attainability in $\mathbb{R}^2$*, J. Optimization Theory Appl., to appear.

[22] ———, *On the set of attainability*, Thesis, Dept. of Math., Univ. of Illinois, Urbana, 1974.

[23] E. ROXIN, *A geometric interpretation of Pontryagin's maximum principle*, Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 303–324.
[24] S. STERNBERG, *Lectures on Differential Geometry*, Prentice-Hall, Englewood Cliffs, N J, 1964.
[25] H. SUSSMANN, *A generalization of the closed subgroup theorem to quotients of arbitrary manifolds*, J. Differential Geometry, 10 (1975), pp. 151–166.
[26] H. SUSSMAN AND N. LEVITT, *Controllability by means of two vector fields*, this Journal, 13 (1975), pp. 1271–1281.
[27] H. WHITNEY, *Analytic extensions of differentiable functions defined in closed sets*, Trans. Amer. Math. Soc., 36 (1934), pp. 63–89.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

A DECOMPOSITION THEORY FOR DIFFERENTIABLE SYSTEMS*

ARTHUR J. KRENER†

Abstract. A theory analogous to the Krohn–Rhodes theory of finite automata is developed for systems described by a finite dimensional ordinary differential equation. It is shown that every such system with a finite dimensional Lie algebra can be decomposed into the cascade of systems with simple or one dimensional algebras. Moreover, in some sense these systems admit no further decomposition. No knowledge of Krohn–Rhodes theory is assumed of the reader.

Introduction. The Krohn–Rhodes theory [1] of finite automata is a very elegant way of describing how a machine can be decomposed as the cascade of two or more simpler machines. Moreover it gives a complete classification of the fundamental building blocks of such cascades. We refer the reader to [2] and [3] for extensive treatment of this and related topics. This paper develops as far as possible a similar theory for differentiable systems, i.e., control systems defined by a nonlinear ordinary differential equation on a finite dimensional manifold.

The first step in this program is to view each constant input as not affecting a particular state but all possible states, that is, to consider the state transition map defined by the input. The family of these maps forms a semigroup which acts on the state space in the obvious fashion and this semigroup has a natural completion to a group. One can try to lift the dynamics from the state space to the group. For a finite state machine, this can always be done resulting in another finite state machine, but for a differentiable system the group of state transition maps need not be finite dimensional. This is a major difference between the two theories.

For a differentiable system it is natural to consider the infinitesimal version of this group, the Lie algebra of vector fields corresponding to constant inputs. This algebra, which has no analogue in Krohn–Rhodes theory, determines the local dynamics of the system. We shall show that if the algebra can be split into an ideal and finite dimensional subalgebra then the system can be split into a cascade. It follows that every system with finite dimensional Lie algebra can be decomposed into a cascade of systems with simple or one dimensional algebras. It is also shown that such systems admit no further decomposition into systems with less complicated Lie algebras. They do however admit decompositions where the controls are split between the elements of the cascade.

1. Nonlinear control systems. In the last few years it has become apparent through the work of Sussman, Brockett, Hermes, Elliott, Lobry, and others that the appropriate state space for nonlinear systems is not $\mathbb{R}^n$. For this paper we adopt a terminology and notation similar to that introduced by Sussmann in his important papers on the existence and uniqueness of minimal realizations of nonlinear systems [4], [5]. We restrict our discussion to real analytic systems for

* Received by the editors February 22, 1976, and in revised form August 30, 1976.

† Department of Mathematics, University of California, Davis, California 95616. This research was supported by the U.S. Office of Naval Research under the Joint Services Electronics Program by Contract N00014-75-C-0648, while the author was a research fellow at the Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts 1974–1975.

consistency of the hypothesis although many of the results hold for C^∞ or even C^k systems.

A *control system* is a 4-tuple $\Sigma = (M, \Omega, f, \mathcal{U})$ consisting of *state space* M, *control set* Ω, dynamics f and *class of admissible inputs* $\mathcal{U}$.

The state space M is a real analytic manifold (all manifolds are assumed to be connected and parcompact) of dimension n whose points are denoted by x. The control set Ω is a subset of $\mathbb{R}^k$ whose points are denoted by u.

The dynamics f is a map whose domain is $M \times \Omega$ such that for each $(x, u) \in M \times \Omega$, $f(x, u)$ is a tangent vector to M at x. Blurring the distinction between the points of M and their local coordinates allows us to express the dynamics by the familiar differential equation

$$\dot{x} = f(x, u).$$

We require that f be a real analytic function of x, continuous in u, which satisfies a local Lipschitz condition in x uniformly in u. (That is, for every compact subset $\Omega^c \subseteq \Omega$ and $\forall x_0 \in M$ there exists a neighborhood V with coordinates x and a constant K such that

$$|f(x_1, u) - f(x_2, u)| \leqq K|x_1 - x_2|$$

$\forall x_i \in V$ and $\forall u \in \Omega^c$). If the dynamics is not autonomous $f = f(t, x, u)$ we adjoin time as an extra state variable to make the system autonomous.

As for the class of inputs, consider the space $\mathcal{U}_m$ of all bounded measurable maps $u(\cdot) : [0, T] \to \Omega$ defined on any finite interval of the form $[0, T]$, $T \geqq 0$. This space can be regarded as a semigroup under concatenation, i.e., if $u_i(\cdot) : [0, T_i] \to \Omega$ are controls in $\mathcal{U}_m$ for $i = 1, 2$, then we define the concatenation $u_1 * u_2(\cdot) : [0, T_1 + T_2] \to \Omega$ also in $\mathcal{U}_m$ by $u_1 * u_2(t) = u_1(t)$ for $t \in [0, T_1]$ and $u_1 * u_2(t) = u_2(t - T_1)$ for $t \in (T_1, T_1 + T_2]$.

The class $\mathcal{U}_{pc}$ of all piecewise constant controls (with at most a finite number of jumps) is a subsemigroup of $\mathcal{U}_m$. We assume that the class of admissible controls $\mathcal{U}$ is also a subsemigroup of $\mathcal{U}_m$ satisfying $\mathcal{U}_{pc} \subset \mathcal{U} \subset \mathcal{U}_m$. For example, $\mathcal{U}$ could be the space of all piecewise C^k controls.

An *output system* is a 6-tuple $\Sigma = (M, \Omega, f, \mathcal{U}, N, g)$ where $(M, \Omega, f, \mathcal{U})$ is a control system, $N = R^n$ for some n and $g : M \to N$ is a real analytic map. We call N the *output space* and g the *output map*.

A system is *initialized* if there exists a distinguished state $x_0 \in M$ at which all trajectories start.

2. The Lie algebra, semigroup and group. A *vector field* h on M is a real analytic function defined on M such that $\forall x \in M$, $h(x)$ is a tangent vector to M at x. The set $V(M)$ of all such vector shields forms an infinite dimensional vector space over the reals with addition and scalar multiplication being defined pointwise. Moreover if $h_i \in V(M)$ we can define a new vector field $[h_1, h_2] \in V(M)$ by means of the Jacobi bracket, expressed in local coordinates by

$$[h_1, h_2](x) = \frac{\partial h_1}{\partial x}(x) h_2(x) - \frac{\partial h_2}{\partial x}(x) h_1(x).$$

It is easy to verify that this definition is independent of coordinates. Often the bracket is defined as the negative of the above but we have chosen this definition to agree with the commutator of two matrices.

The bracket is a noncommutative and nonassociative multiplication which instead satisfies the skewsymmetry and Jacobi relations,

$$[h_1, h_2] = -[h_2, h_1],$$
$$[h_1[h_2, h_3]] = [[h_1, h_2]h_3] + [h_2[h_1, h_3]].$$

This makes $V(M)$ into an infinite dimensional real *Lie algebra.* It is convenient to introduce the notation

$$ad^0(h_1)h_2(x) = h_2(x),$$
$$ad^l(h_1)h_2(x) = [h_1, ad^{l-1}(h_1)h_2](x).$$

Given a system $\Sigma = (M, \Omega, f, \mathcal{U})$ for each $u \in \Omega$ we define a vector field $f(\cdot, u)$ and define the Lie algebra L of Σ as the smallest subalgebra of $V(M)$ which contains all such vector fields.

To each vector field $h(\cdot) \in V(M)$ there exists a *flow* $\phi(t, x)$ defined as the family of solutions of the differential equation

$$\frac{d}{dt}\phi(t, x) = h(\phi(t, x))$$

satisfying the intial condition

$$\phi(0, x) = x.$$

Solutions of this equation could possibly escape from M in finite time and so the map ϕ is defined locally, i.e., for $x_0 \in M$ there exists a compact neighborhood $K \subset M$ containing x_0 and an interval $I \subset \mathbb{R}$ containing 0 such that $\phi : I \times K \to M$ is well defined. If $\phi : \mathbb{R} \times M \to M$ can be defined, the vector field $h(\cdot)$ is said to be *complete.*

Let Diff (M) be the group of all analytic diffeomorphisms of M, i.e., $\phi \in$ Diff (M), if $\phi : M \to M$ is 1–1, onto and both ϕ and ϕ^{-1} are analytic. On Diff (M) we put the topology of uniform convergence on compacta of a map and its derivatives. (A sequence $\phi_n \in$ Diff (M) converges to $\phi \in$ Diff (M) if every sequence of partial derivatives of ϕ_n including ϕ_n itself converges to the corresponding partial derivative of ϕ uniformly on every compact subset of M.) If $h(\cdot)$ is complete then for each t the map $\phi(t, \cdot) : M \to M$ is in Diff (M).

The system Σ is *complete* if every vector field $h(\cdot) \in L$ is complete. Palais [6, p. 95] has shown that if the Lie algebra L of Σ is finite dimensional and $f(\cdot, u)$ is complete for each constant $u \in \Sigma$ then Σ is complete. Henceforth we shall assume that Σ is complete.

Each constant control $u \in \Omega$ generates a flow $\phi_u(t, x)$ which can be viewed as a group homomorphism of the additive group $\mathbb{R}$ into Diff (M) in light of the following identity:

$$\phi_u(t_1 + t_2, x) = \phi_u(t_1, \phi_u(t_2, x)) = \phi_u(t_2, \phi_u(t_1, x)).$$

This allows us to define a semigroup homomorphism from $\mathcal{U}_{pc}$ into Diff (M) in the obvious fashion. The range of this homomorphism is a subsemigroup S of Diff (M) and we refer to this as the *semigroup of the system* Σ. The smallest subgroup G of Diff (M) containing S is called the *group of the system.*

Given a point $x_0 \in M$, we can consider the orbits of x_0 under the semigroup S and group G respectively

$$S(x_0) = \{\phi(x_0) : \phi \in S\},$$
$$G(x_0) = \{\phi(x_0) : \phi \in G\}.$$

$S(x_0)$ is often referred to as the set of points accessible from x_0 under Σ by controls in $\mathcal{U}_{pc}$. Because we have assumed $f(x, u)$ to be analytic in x it can be shown using the Hermann–Nagano theorem [17], [7] that $G(x_0)$ is an analytic submanifold of M. In some sense $G(x_0)$ is the natural submanifold of M on which to consider the problem for it contains all trajectories of the system emanating from x_0. Chow's theorem [8] tells us that every point in $G(x_0)$ can be reached from x_0 along trajectories of the system going both forward and backward in time. Moreover it is of minimal dimension among submanifolds of M containing $S(x_0)$ because $S(x_0)$ has a nonempty interior in the topology of $G(x_0)$ [9], [10]. (Note the topology of $G(x_0)$ is not necessarily its relative topology inherited from M.) If $S(x_0) = M(G(x_0) = M)$ then the system is said to be *controllable* (*weakly controllable*), see [18]. If $G(x_0) \neq M$ then we redefine the Lie algebra L of Σ to be the smallest subalgebra of $V(Gx_0)$ which contains all the vector fields $f(\cdot, u)$, $u \in \Omega$.

Let $u(\cdot) \in \mathcal{U}_m$; then given any $x_0 \in M$ there exists a compact neighborhood K of x_0, an open interval I containing 0 and a map $\phi_u : I \times K \to M$ satisfying

$$\frac{d}{dt}\phi_u(t, x) = f(\phi_u(t, x), u(t)),$$
$$\phi_u(0, x) = x. \tag{2.1}$$

Since $u(\cdot)$ is only a bounded measurable function the curve $t \mapsto \phi_u(t, x)$ is generally only absolutely continuous. For each $t \in I$ the map $\phi_u(t, \cdot) : K \to M$ is 1–1 and analytic. If it can be defined on all of M then it is an element of Diff (M). Since we have assumed that Σ is complete, if $u(\cdot) \in \mathcal{U}_{pc}$ then ϕ_u can be defined on $\mathbb{R} \times M$. However if $u(\cdot) \in \mathcal{U}_m - \mathcal{U}_{pc}$ this need not be true. (See Sussmann [4, p. 14] for a counterexample). The effect of $u(\cdot) \in \mathcal{U}_m$ can always be approximated by piecewise constant controls.

Approximation Lemma [4]. *Let $u(\cdot) \in \mathcal{U}_m$ and $\phi_u : I \times K \to M$ be its flow. Suppose $\{u^j(\cdot)\} \subset \mathcal{U}_{pc}$ is a sequence of piecewise constant controls such that $u^j(t) \to u(t)$ for almost all $t \in I$. If $\phi_j(\cdot, \cdot)$ is the flow of $u^j(\cdot)$ and J is a compact subinterval of I then $\phi_j \to \phi_u$ uniformly on $J \times K$.*

The above lemma indicates why S and G can be defined without regard to the class of admissible inputs, $\mathcal{U}$ as long as $\mathcal{U}_{pc} \subset \mathcal{U} \subset \mathcal{U}_m$.

The semigroup S plays a similar role in the theory of differentiable systems as the semigroup of a machine in the theory of finite automata. They both describe the action of the semigroup of inputs on all states of the system/machine. There are some important distinctions however; one is the problem of finite escape time.

Another is that the state transition map $\phi_u(t, \cdot)$ is always in Diff (M) for differentiable systems and hence S can naturally be extended to a subgroup G of Diff (M). On the other hand the state transition maps of a finite automaton are not necessarily invertible. The group of the machine is generated by the invertible ones and may not contain the semigroup of the machine.

The Lie algebra L has no analogue in the theory of finite automata; it is an infinitesimal version of G. By this we mean L completely determines $\phi_u(t, \cdot)$ for small t.

In both theories it is desirable to lift the dynamics from the state space to the group/semigroup of the system/machine. That is, we view inputs as not affecting a particular state but rather affecting all possible states. In the case of a finite automaton this results in the semigroup of the machine becoming the new state space. This state space is again finite and hence the new machine is again a finite automaton.

In the case of a differentiable system, G is not always a finite dimensional manifold and therefore the dynamics lifted to G is not described by a finite dimensional differential equation. We define Σ to be *finite dimensional* if L is finite dimensional. In this case G can be given the structure of a finite dimensional real analytic manifold compatible with the group operation (Palais [6]). This makes G into a Lie group and L can be viewed as the Lie algebra of right invariant vector fields on G. This allows us to lift the dynamics from the state space M where they are given by

$$\dot{x} = f(x, u), \qquad x(0) = x_0$$

to a new state space G where they are given locally by a matrix differential equation.

We shall elaborate on this in the next section but first we illustrate these points with some familiar examples, the first of which is a linear system.

Σ: Let the state space $M = \mathbb{R}^n$, the control set $\Omega = \mathbb{R}^k$, the initial point be x_0 and the dynamics be given by

$$\dot{x} = Ax + \sum_{i=1}^{k} u_i b_i \tag{2.2}$$

where A is $n \times n$ real matrix and b_i are n-vectors.

The Lie bracket is given by

$$[Ax, b_i] = Ab_i, \qquad [b_i, b_j] = 0,$$
$$ad^l(Ax)b_i = A^l b_i, \qquad [b_i, ad^l(Ax)b_j] = 0$$

and so the Lie algebra is finite dimensional with a particularly simple form which characterizes a locally linear system, [11]. The zero control defines a matrix differential equation

$$\dot{\Phi}(t) = A\Phi(t)$$

with $\Phi(0) = I$ the identity matrix. The solution is $\Phi(t) = \exp(tA)$. For any control

$u(\cdot) \in \mathcal{U}_m$ the flow ϕ_u is given by the variation of constants formula

$$\phi_u(t, x) = \Phi(t)x + \int_0^t \Phi(t-s) \sum_i u_i(s) b_i \, ds,$$

which for fixed t defines an invertible affine map $\phi_u(t, \cdot) : \mathbb{R}^n \to \mathbb{R}^n$. S and G are the smallest subsemigroup and subgroup of the group of invertible affine motions of $\mathbb{R}^n$ which contain all such $\phi_u(t, \cdot)$, $t \geqq 0$. Lifting the dynamics from the state space $\mathbb{R}^n$ to G means replacing the vector differential equation (2.2) by the matrix differential equation

$$\dot{X} = AX + \left(\sum_i u_i b_i \cdots \sum_i u_i b_i \right).$$

Now consider a bilinear system

Σ: Let the state space be $M = \mathbb{R}^n$, the control set be $\Omega = \mathbb{R}^k$, the initial point be x_0 and the dynamics be given by

$$\dot{x} = \left(A + \sum_i u_i B_i \right) x \tag{2.3}$$

when A and B_i are $n \times n$ real matrices.

If C and D are $n \times n$ real matrices the Lie bracket of the vector fields Cx and Dx is given by the commutator $[C, D] = CD - DC$,

$$[Cx, Dx] = [C, D]x.$$

Therefore, L is isomorphic to the smallest subalgebra of $gl(n, \mathbb{R})$ (the Lie algebra of all $n \times n$ real matrices) containing A and B_i. Each control $u(\cdot) \in \mathcal{U}$ defines a matrix differential equation

$$\dot{\Phi}_u(t) = \left(A + \sum_i u_i(t) B_i \right) \Phi_u(t) \tag{2.4}$$

where $\Phi_u(t) \in GL(n, \mathbb{R})$, the Lie group of invertible matrices in $gl(n, \mathbb{R})$ and

$$\phi_u(t, x) = \Phi_u(t)x.$$

S and G are the smallest subsemigroup and subgroup of $GL(n, \mathbb{R})$ containing Φ_u for each $u \in \Omega$. Lifting the dynamics from M to G again means replacing the vector differential equation (2.3) by the matrix differential equation (2.4). The finiteness of the Lie algebra locally characterizes bilinear systems [12].

It is well known that every linear control system can be turned into a bilinear system by the addition of an extra coordinate which is identically one, so henceforth when we refer to bilinear systems we include linear ones.

3. Simulation. Given an initialized system $I = (M, \Omega, f, \mathcal{U}, N, g, x_0)$ let $\mathcal{X}_{x_1}$ denote the space of all absolutely continuous functions $x(\cdot) : [0, T] \to M$ for any $T \geqq 0$ such that $x(0) = x_1$. Similarly define $\mathcal{Y}_{x_1}$ as the space of all absolutely continuous functions $y(\cdot) : [0, T] \to N$ for any $T \geqq 0$ such that $y(0) = g(x_1)$. Given any $x_1 \in M$ the system Σ defines a pair of maps $\mathcal{F}_{x_1} : \mathcal{U} \to \mathcal{X}_{x_1}$ and $\mathcal{G}_{x_1} : \mathcal{U} \to \mathcal{Y}_{x_1}$ in the obvious fashion:

$$\mathcal{F}_{x_1}(u(t)) = \phi_u(t, x_1) \quad \text{and} \quad \mathcal{G}_{x_1}(u(t)) = g(\phi_u(t, x_1)).$$

Two points $x_1, x_2 \in M$ are said to be *indistinguishable* if $\mathcal{G}_{x_1}(u(t)) = \mathcal{G}_{x_2}(u(t))$ for all $u(\cdot) \in \mathcal{U}_{pc}$. The system Σ is *observable* if x_1 and x_2 indistinguishable implies that $x_1 = x_2$. The system Σ is *minimal* if it is weakly controllable and observable.

Given a pair of initialized systems

$$\Sigma^i = (M^i, \Omega^i, f^i, \mathcal{U}^i, N^i, g^i, x_0^i) \quad \text{for } i = 1, 2,$$

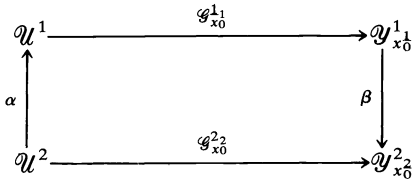
the maps $\mathcal{F}^i_{x_0^i}$ and $\mathcal{G}^i_{x_0^i}$ describe the state space and input/output behavior of the systems. A question of some importance is when the behavior of one system simulates the behavior of the other in either of the above senses. There are several alternative ways of approaching this problem. In the classical theory of minimal realizations of linear systems one assumes that $\Omega^1 = \Omega^2$, $N^1 = N^2$ and studies when two input-output equivalent systems differ by a homomorphism or isomorphism of the state space, $\mathbb{R}^n$. A similar theory has been developed by H. Sussmann for nonlinear systems which we shall discuss in a moment.

In the theory of finite automata the input and output spaces are allowed to differ by encoding and decoding functions. One wishes to know when an automaton can be made to simulate the input-output behavior of another automaton by a suitable choice of encoder and decoder. For algebraic reasons it is considerably easier to discuss when the state behavior of an automaton can be simulated by another automaton. We describe a similar theory for nonlinear systems.

Given a pair of initialized systems Σ^i and functions $\alpha : \Omega^2 \to \Omega^1$, $\beta : N^1 \to N^2$ with α continuous and β analytic we obtain induced maps (also denoted by α and β) $\alpha : \mathcal{U}^2 \to \mathcal{U}^1$ and $\beta : \mathcal{Y}^1_{x_0^1} \to \mathcal{Y}^2_{x_0^2}$ in the obvious fashion:

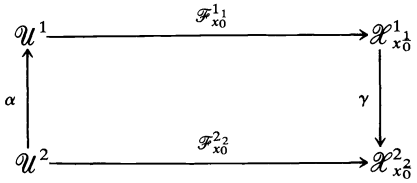
$$\alpha(u(\cdot))(t) = \alpha(u(t)) \quad \text{and} \quad \beta(y(\cdot))(t) = \beta(y(t)).$$

Σ^1 *simulates* Σ^2 with encoder α and decoder β if the following diagram commutes:



Σ^1 is *equivalent* to Σ^2 if Σ^1 simulates Σ^2 with $\alpha : \Omega^2 = \Omega^1$ and $\beta : N^1 = N^2$ the identity maps.

Suppose Σ^1 is weakly controllable, i.e., $G(x_0^1) = M^1$. Given a pair $\alpha : \Omega^2 \to \Omega^1$ and $\gamma : M^1 \to M^2$ with α continuous and γ analytic we say that Σ^1 is homomorphic to Σ^2 if the following diagram commutes:



Σ^1 is isomorphic to Σ^2 if Σ^1 is homomorphic to Σ^2 with $\alpha : \Omega^1 = \Omega^2$ the identity and $\gamma : M^1 \to M^2$ a diffeomorphism. If Σ^1 is not weakly controllable it is sufficient to find $\gamma : G(x_0^1) \to M^2$ such that the appropriate diagram commutes.

There is also a local form of the above definitions. For example suppose Σ^1 is weakly controllable, define $\mathcal{U}^i_T = \{u(\cdot) \in \mathcal{U}^i : u(\cdot)$ defined on $[0, t]$ when $t \leqq T\}$. Σ^1 is locally homomorphic to Σ^2 if there exists $T > 0$, a neighborhood V of x^1_0 in M^1 and a map $\gamma : V \to M^2$ such that the appropriate diagram commutes.

It is difficult to give conditions for Σ^1 to simulate Σ^2 because of the complications caused by the presence of the encoder and decoder. However if one is interested in equivalence these difficulties are somewhat mitigated and H. Sussmann has proved the following generalization of the existence and uniqueness theorem for minimal realization of linear systems.

SUSSMAN'S THEOREM [5]. *Every initialized analytic system is equivalent to a minimal system. Any two equivalent minimal systems are isomorphic.* For related results see [18].

We now focus in one the question of when Σ^1 is homomorphic to Σ^2, Suppose Σ^1 has the accessibility property and is homomorphic to Σ^2. Then it is easy to show that the Jacobian γ_*

$$\gamma_*(x^1) = \frac{\partial \gamma}{\partial x^1}(\cdot)$$

must define a homomorphism $\gamma_* : L^1 \to L^2$ of the Lie algebras L^i of Σ^i. Furthermore, the maps α and $\gamma_*(x^1)$ must satisfy the following commutative diagram

$$\begin{array}{ccc} \Omega^1 & \xrightarrow{f^1(x^1, \cdot)} & T_{x^1}M^1 \\ \uparrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \gamma_*(x^1)} \\ \Omega^2 & \xrightarrow{f^2(\gamma(x^1), \cdot)} & T_{x^2}M^2 \end{array}$$

where $T_{x^i}M^i$ is the tangent space to M^i at x^i.

Suppose $u_0, u_1, \cdots, u_l \in \Omega^2$ such that

$$f^1(x^1, \alpha(u_0)) = \sum_{i=1}^{l} \lambda_i f^1(x^1, \alpha(u_i));$$

then the linearity of $\gamma_*(x^1)$ implies that

$$f^2(\gamma(x^1), u_0) = \sum_{i=1}^{l} \lambda_i f^2(\gamma(x^1), u_i).$$

In particular if $\alpha(u_0) = \alpha(u_1)$ then $f^2(x^2, u_0) = f^2(x^2, u_1)$ for every $x^2 \in G^2(x^2_0)$. It follows that the map $f^1(\cdot, \alpha(u)) \mapsto f^2(\cdot, u)$ has a well-defined linear extension from span $\{f^1(\cdot, \alpha(u)) : u \in \Omega^2\}$ to span $\{f^2(\cdot, u) : u \in \Omega^2\}$.

Let $\alpha(L^2)$ denote the subalgebra of L^1 generated by $f^1(\cdot, \alpha(u))$ for each $u \in \Omega^2$. Let L^i_0 denote the *isotropy subalgebra* of L^i at x^i_0, i.e.,

$$L^i_0 = \{h(\cdot) \in L^i : h(x^i_0) = 0\}$$

and let $\alpha(L^2)_0 = L^1_0 \cap \alpha(L^2)$.

If L^1 and L^2 are arbitrary Lie algebras we say that L^2 *divides* L^1 if there exists a subalgebra $L \subset L^1$ and a Lie algebra homomorphism of L onto L^2. If L^i is the Lie algebra of system Σ^i we say L^2 G-divides L^1 if there exists a continuous map

$\alpha : \Omega^2 \to \Omega^1$ such that the map $f^1(\cdot, \alpha(u)) \mapsto f^2(\cdot, u)$ generates a Lie algebra homomorphism of $\alpha(L^2)$ onto L^2. If in addition this homomorphism carries $\alpha(L^2)_0$ into L_0^2 then we say L^2 Σ-*divides* L^1.

THEOREM 1 (Krener [11]). Σ^1 *is locally homomorphic to* Σ^2 *if* L^2 Σ-*divides* L^1. *Moreover if* L^2 Σ-*divides* L^1 *and* $G^1(x_0^1)$ *is simply connected then* Σ^1 *is homomorphic to* Σ^2.

Suppose L^2 Σ-divides L^1 but $G^1(x_0^1)$ is not simply connected; then we can ask if there is any way to lift Σ^1 to a system that is simply connected. The answer is yes; suppose for convenience $G^1(x_0^1) = M^1$. Then M^1 has a unique simply connected covering manifold M with covering map $\pi : M \to M^1$. Given any $x \in M$ we can find a sufficiently small neighborhood V of x such that the map $\pi : V \to M^1$ is a diffeomorphism, and so the Jacobian π_* is invertible. This allows us to lift the dynamics from M^1 to M by defining

$$f(x, u) = \pi_*^{-1}(x) f^1(\pi(x), u)$$

for $x \in M$ and $u \in \Omega = \Omega^1$. We choose any initial point $x_0 \in \pi^{-1}(x_0^1)$ and by Chow's theorem we know that $G(x_0) = M$. The result is a system $\Sigma = (M, \Omega, f, x_0)$ which is called the *simply connected cover* of Σ^1 and is homomorphic to Σ^1 under $\alpha : \Omega^1 \to \Omega$ the identity and $\pi : M \to M^1$ the covering map. If Σ^1 is locally homomorphic to Σ^2 then Σ is homomorphic to Σ^2.

For example consider the following system: $\Sigma^1 : M^1 = S^1$ the unit circle with angular coordinate θ, $\Omega = \mathbb{R}$, $\theta^0 = 0$ and $\dot{\theta} = u$. The simply connected covering space of S^1 is $\mathbb{R}$ and so the simply connected cover of Σ^1 is $\Sigma : M = \mathbb{R}$, $\Omega = \mathbb{R}$, $x^0 = 0$ and $\dot{x} = u$.

Suppose L^2 G-divides L^1 but does not Σ-divide L^1, then we can ask if there is any way to lift Σ^1 to a different system such that L^2 Σ-divides L. The answer is yes if the Lie algebra L^1 of Σ^1 is finite dimensional. The group G^1 of Σ^1 is then a finite dimensional Lie group (Palais [6]) and by Ado's theorem L^1 is isomorphic to some subalgebra L of $gl(m, \mathbb{R})$ for some m possibly different from $n^1 = \text{dimension } M^1$. (For bilinear systems L^1 is already a subalgebra of $gl(n^1, \mathbb{R})$.) For each $u \in \Omega^1$, let $F(u)$ be the matrix corresponding to $f^1(\cdot, u)$ under this isomorphism. Let H be the subgroup of $GL(m, \mathbb{R})$ corresponding to L; then some neighborhood of the identity in G^1 is isomorphic as a Lie group to some neighborhood of the identity in H. This isomorphism can be viewed as defining H-valued local coordinates on G^1 in which the dynamics of Σ^1 is described by

$$\dot{X} = F(u)X \tag{3.1}$$

where $X \in H \subset GL(m, \mathbb{R})$.

If H and G^1 are globally isomorphic as in the case of bilinear systems then the matrix differential equation describes the lifted dynamics throughout G^1. For bilinear systems the action of $G^1 = H$ on the state space $M^1 = \mathbb{R}^{n^1}$ is the natural linear action, however H and G^1 globally isomorphic does not necessarily imply that the action of $G^1 \subset \text{Diff}(M^1)$ on M^1 is linear in any coordinates.

If H and G^1 are only locally isomorphic then this isomorphism can be used to define H-valued coordinates in a neighborhood of every $\phi \in G^1$. In these coordinates the dynamics is locally given by a matrix differential equation similar to (3.1). We define a new system $\Sigma = (M, \Omega, f, \phi_0)$ where $M = G^1$, $\Omega = \Omega^1$, the

dynamics f is given locally by (3.1) and ϕ_0 is the identity of G^1, i.e., $\phi_0 : M^1 \to M^1$ the identity map. Σ is called the *group cover* of Σ^1.

Σ is homomorphic to Σ^1 under $\alpha : \Omega^1 \to \Omega$ the identity and $\gamma : G^1 \to M^1$ given by $\gamma(\phi) = \phi(x_0)$ for $\phi \in G^1$. Notice that the Lie algebra L if Σ is isomorphic to the Lie algebra L^1 of Σ^1 but the isotropy subalgebras need not be. If L^2 G-divides L^1 then L^2 Σ-divides L for $L_0 = 0$.

Recall that G^1 is the not necessarily closed subgroup of Diff (M) generated by the flows of constant controls. Suppose $u(\cdot) \in \mathcal{U}$ and for some t, the map $\phi_u(t, \cdot)$ defined by (2.1) is a diffeomorphism of M^1. The Approximation Lemma implies that $\phi_u(t, \cdot) \in$ closure G^1; however by lifting the dynamics to G^1 we have shown that $\phi_u(t, \cdot) \in G^1$.

Finally suppose L^2 divides L^1 but does not G-divide L^1, then we can ask if there is any way to lift Σ^1 to a different system Σ such that L^2 G-divides L. The answer is again yes provided L^1 is finite dimensional. Let $\Sigma = (M, \Omega, f, x_0)$ where $M = M^1$, $\Omega = L^1$, $x_0 = x_0^1$ and $f(x, h(\cdot)) = h(x)$ for $x \in M^1$ and $h(\cdot) \in L^1$. Define $\alpha : \Omega^1 \to \Omega$ by $\alpha(u) = f^1(\cdot, u)$ and $\gamma : M \to M^1$ the identity; then clearly Σ is homomorphic to Σ^1. Σ is called the *fully controllable cover* of Σ^1 and if $\Sigma = \Sigma^1$ then Σ^1 is said to be fully controllable. Notice that in a fully controllable system the dynamics is linear in the controls.

4. Cascades. Suppose $\Sigma^i = (M^i, \Omega^i, f^i, x_0^i)$ are control systems for $i = 1, 2$. Henceforth we assume $\mathcal{U}^i = \mathcal{U}_m^i$ and therefore do not mention it explicitly. Let $v : M^1 \times \Omega^1 \to \Omega^2$ be an analytic map of x^1, continuous with respect to u^1. We define the *cascade* $\Sigma^1 \circledcirc_v \Sigma^2$ of these two systems with *linking map* v as the system

$$(M^1 \times M^2, \Omega^1, f^1 \circledcirc_v f^2, (x_0^1, x_0^2))$$

where

$$f^1 \circledcirc_v f^2(x^1, x^2, u) = (f^1(x^1, u^1), f^2(x^2, v(x^1, u^1))).$$

$\Sigma^1 \circledcirc_v \Sigma^2$ is a *parallel cascade* if v is a function of u_1 alone and a *series cascade* if v only depends on x^1.

Cascades are a way of combining two or more systems to obtain a more complicated system. We would like to study when this technique can be used to represent a given system as the homomorphic image of a cascade of "simpler" systems. Of course any system is the homomorphic image of a cascade consisting of itself followed by an arbitrary system but this can hardly be called a cascade of "simpler" systems.

We must make rigorous the notion of "simpler"; the obvious choice is that Σ^i is "simpler" than Σ if Σ is homomorphic to Σ^i. However this is not the appropriate definition for if Σ is actually a cascade, $\Sigma^1 \circledcirc_v \Sigma^2$, then it is easy to see that Σ is homomorphic to Σ^1 but it need not be homomorphic to Σ^2. Therefore we are forced to a weaker definition—Σ^i is "simpler" than Σ if Σ is not a homomorphic image of Σ^i. A similar definition is used by Krohn and Rhodes.

A system Σ has a *nontrivial cascade decomposition* if there exists Σ^1, Σ^2 and v such that $\Sigma^1 \circledcirc_v \Sigma^2$ is homomorphic to Σ but neither Σ^1 nor Σ^2 alone is homomorphic to Σ.

We leave it to the reader to verify the following.

LEMMA. *Suppose* Σ *is homomorphic to* $\Sigma^1(\Sigma^2)$; *then* $\Sigma \circledcirc_w \Sigma^2(\Sigma^1 \circledcirc_w \Sigma)$ *is homomorphic to* $\Sigma^1 \circledcirc_v \Sigma^2$ *for some linking map* w.

COROLLARY. *Suppose* $\Sigma^1 \circledcirc_v \Sigma^2$ *is homomorphic to* Σ *and* $\Sigma^3 \circledcirc_w \Sigma^4$ *is homomorphic to* Σ^1 (Σ^2). *Then* $\Sigma^3 \circledcirc_w \Sigma^4 \circledcirc_v \Sigma^2$ $(\Sigma^1 \circledcirc_v \Sigma^3 \circledcirc_w \Sigma^4)$ *is homomorphic to* Σ.

We describe a way for decomposing a system into cascades which is based on a technique used by E. Wichmann [13] and K. T. Chen [19] and originally due to S. Lie [20]. Let $\Sigma = (M, \Omega, f, x_0)$ and suppose the Lie algebra L of Ω is a semidirect sum,

$$L = L^1 + L^2.$$

(L is a *semidirect sum* of L^1 and L^2 if L^1 is a subalgebra of L, L^2 is an ideal of L and L is the direct sum of L^1 and L^2 as vector spaces. We exclude the trivial case where either is 0.) For each $u \in \Omega$, define $f^1(\cdot, u) \in L^1$ and $g(\cdot, u) \in L^2$ by requiring that

$$f(\cdot, u) = f^1(\cdot, u) + g(\cdot, u).$$

Consider the control system $\Sigma^1 = (M^1, \Omega^1, f^1, x_0^1)$ where $M^1 = M$, $\Omega^1 = \Omega$ and $x_0^1 = x_0$. Let G^1 be the group of Σ^1. Clearly the Lie algebra of Σ^1 is L^1.

Define a second system $\Sigma^2 = (M^2, \Omega^2, f^2, x_0^2)$ where $M^2 = M$, $\Omega^2 = G^1 \times \Omega$, $x_0^2 = x_0$ and

$$f^2(x, \phi, u) = \phi_*^{-1}(x) g(\phi(x), u),$$

for $x \in M$, $\phi \in G^1$ and $u \in \Omega$. There is a problem with this definition for in general the control set Ω^2 is not finite dimensional since G^1 is not. However if we assume that L^1 is finite dimensional then G^1 is embeddable in some R^l by the Whitney theorem and hence Σ^2 is a control system according to our definition.

Moreover if L^1 is finite dimensional then we can redefine Σ^1 so it equals its group cover. This allows us to form the cascade $\Sigma^1 \circledcirc_v \Sigma^2$ where the linking map $v : G^1 \times \Omega \to \Omega^2$ is the identity.

For a fixed control $u(\cdot) \in \mathcal{U}$, let $x(t)$ be the trajectory in Σ, $\phi(t, \cdot)$ be the trajectory in Σ^1 and $x^2(t)$ be the trajectory in Σ^2. We claim $x(t) = \phi(t, x^2(t))$ and we show this by noting that $x(0) = x_0 = \phi(0, x^2(0))$ and using (2.1) we see that both satisfy the same differential equation

$$\begin{aligned}\frac{d}{dt}\phi(t, x^2(t)) &= f^1(\phi(t, x^2(t)), u(t)) + \phi_*(t, x^2(t)) x^2(t) \\ &= f^1(\phi(t, x^2(t)), u(t)) + g(\phi(t, x^2(t)), u(t)) \\ &= f(\phi(t, x^2(t)), u(t)).\end{aligned}$$

Therefore $\Sigma^1 \circledcirc_v \Sigma^2$ is homomorphic to Σ under $\alpha : \Omega \to \Omega$, the identity, and $\gamma : G^1 \times M^2 \to M$ given by $\gamma(\phi, x) = \phi(x)$. Since the Lie algebras of Σ^1 and Σ^2 are L^1 *and* L^2 it follows that neither system is homomorphic to Σ and hence the cascade decomposition is nontrivial.

If $g(\cdot, u)$ is independent of u then $\Sigma_1 \circledcirc_v \Sigma^2$ is a series cascade. On the other hand suppose L is a direct sum of L^1 and L^2 (that is, both are ideals of L). From

$[L^1, L^2] \subset L^1 \cap L^2 = 0$ it follows that for every $\phi \in G^1$, $f^2(x, \phi, u) = \phi_*^{-1}(x) g(\phi(x), u) = g(x, u)$. Therefore f^2 is independent of ϕ and $\Sigma^1 \circledcirc_v \Sigma^2$ is a parallel cascade. We sum this up in the following:

THEOREM 2. *If the Lie algebra of a system is the semidirect sum of a finite dimensional subalgebra and an ideal then it has a nontrivial cascade decomposition. If it is the direct sum of two ideals, then it has a parallel cascade decomposition.*

A particular application of the above result is when the system is finite dimensional, as in the case of a bilinear system. By Levi's theorem L is a semidirect sum of semisimple subalgebra L^1 and a maximal solvable ideal L^2. Therefore Σ has a cascade decomposition $\Sigma^1 \circledcirc_v \Sigma^2$.

Every finite dimensional semisimple Lie algebra L^1 is a direct sum

$$L^1 = L^{11} + \cdots + L^{1l}$$

of simple ideals L^{1i}. Therefore by repeated application of the above theorem Σ^1 can be decomposed into the parallel cascade of a family of systems Σ^{1i}, $i = 1, \cdots, l$, each with a simple Lie algebra, L^{1i}. Recall a Lie algebra is simple if it is not Abelian and contains no nontrivial ideals; therefore the Σ^{1i} admit no further decomposition using Theorem 2. However as we show by example in a moment systems whose Lie algebra is simple can admit nontrivial cascade decompositions.

We now turn to the system Σ^2 whose Lie algebra L^2 is solvable. This implies that $[L^2, L^2]$ is a proper ideal of L^2 and hence one can find a linear subspace L^{22} of codimension one in L^2 which contains $[L^2, L^2]$. Since L^{22} contains $[L^2, L^2]$ it is an ideal of L^2, and since it is of codimension one any vector field in $L^2 \backslash L^{22}$ generates a one dimensional subalgebra L^{21} such that $L^2 = L^{21} + L^{22}$ is a semidirect sum.

Using Theorem 2, Σ^2 can be decomposed into the cascade of a one dimensional system Σ^{21} and a system Σ^{22} with solvable Lie algebra of one lower dimension. By induction Σ^2 is cascade decomposition of a family of one dimensional systems.

Moreover there are, up to isomorphism, only two one dimensional systems, those on the circle and line described in § 3. Therefore we have shown the following.

THEOREM 3. *If the Lie algebra L of Σ is finite dimensional then Σ admits a decomposition into the parallel cascade of systems with simple Lie algebras followed by a cascade of one dimensional systems.*

This result is somewhat stronger than that of Brockett [14] since all the component systems of Theorem 3 are either simple or one dimensional. In Brockett's work the component systems are reductive. The one dimensional algebra and all simple algebras are reductive but $gl(n, \mathbb{R})$ is not simple but reductive.

Theorem 3 is highly reminescent of the Krohn–Rhodes theorem which states that every finite state machine can be broken up as a cascade of machines with simple groups and flip-flops. The system with simple Lie algebras are analogous to machines with simple groups but the analogy breaks down between one dimensional systems and flip-flop machines, since flip-flops correspond to the nongroup part of the machine.

Recall that in § 1 we suggested that time varying systems be made autonomous by the introduction of time as another state variable. Unfortunately this can

make the Lie algebra of a time varying bilinear system infinite dimensional and therefore not amenable to the application of Theorem 3. Instead suppose we consider time as another control variable and view the time varying bilinear system as the cascade of a trivial system and the bilinear system with the time control.

$$\Sigma^1: \quad \dot{x}^1 = 1,$$

$$\Sigma^2: \quad \dot{x}^2 = A(u_0)x^2 + \sum_{i=1}^{k} u_i B_i(u_0)x^2$$

where the linking map is $u_0 = x^1$. Then we can see Theorem 3 to decompose Σ^2 since its Lie algebra is contained in $gl(n, \mathbb{R})$ and hence is finite dimensional.

Actually the above technique can be used to generate a cascade decomposition even if L is not a semidirect sum. (For example see [19].) Instead of splitting the Lie algebra of Σ between Σ^1 and Σ^2, we split the controls. Let K denote the subspace of $V(M)$ which is the span $\{f(\cdot, u): u \in \Omega\}$ and suppose K admits a direct sum decomposition (as a real vector space)

$$K = K^1 + K^2.$$

For each $u \in \Omega$, define $f^1(\cdot, u) \in K^1$ and $g(\cdot, u) \in K^2$ by requiring that

$$f(\cdot, u) = f^1(\cdot, u) + g(\cdot, u).$$

Define Σ^1 as before; if its Lie algebra L^1 generated by K^1 is finite dimensional then we can also define Σ^2 and the cascade $\Sigma^1 \circledcirc_v \Sigma^2$. The same argument as before shows that $\Sigma^1 \circledcirc_v \Sigma^2$ is homomorphic to Σ.

We ask whether this is a nontrivial cascade decomposition. The Lie algebras L^1 and L^2 could each be equal to L so we must check if Σ^i is homomorphic to Σ. If the decomposition of K is nontrivial then K^1 is a proper subset of K so L could not possibly G-divide L^1. This implies that Σ^1 is not homomorphic to Σ.

On the other hand the generators of L^2 are contained in the orbit of K^2 under the group G^1 acting by conjugation. The Campbell–Baker–Hausdorff formula shows that this is equal to the orbit of K^2 under the Lie algebra L^1 acting by bracketing. Therefore if this orbit does not contain K then we can conclude that L does not G-divide L^2 and hence Σ^2 is not homomorphic to Σ.

THEOREM 4. *Let K be the linear span of the vector fields corresponding to constant controls of a system Σ and suppose K is a direct sum of linear subspaces K^1 and K^2. If the Lie algebra L^1 generated by K^1 is finite dimensional and orbit of K^2 under L does not include K then Σ admits a nontrivial cascade decomposition.*

Now we use this theorem to exhibit a system whose Lie algebra is simple but admits a nontrivial cascade decomposiition. Σ: Let $M = SL(2, \mathbb{R})$ the group of real 2×2 matrices of determinant 1, $\Omega = \mathbb{R}^3$, the initial point be any identity matrix and the dynamics be given by

$$\dot{X} = \sum_{i=1}^{3} u_i B_i X$$

where

$$B_1 = \begin{vmatrix} 1 & 0 \\ 0 & -1 \end{vmatrix}, \quad B_2 = \begin{vmatrix} 0 & 1 \\ 0 & 0 \end{vmatrix}, \quad B_3 = \begin{vmatrix} 0 & 0 \\ 1 & 0 \end{vmatrix}.$$

The Lie brackets are given by $[B_1, B_2] = 2B_2$, $[B_1, B_3] = -2B_3$ and $[B_2, B_3] = B_1$. If we let K^1 be the span of B_1X and K^2 be the span of B_2X and B_3X then Theorem 4 applies and we obtain the nontrivial cascade decomposition

$$\dot{X}^1 = u_1 B_1 X^1,$$
$$\dot{X}^2 = (X^1)^{-1}(u_2 B_2 + u_3 B_3) X^1 X^2$$

where

$$X = X^1 X^2.$$

Notice that if we defined K^1 to be the span of B_2X and the K^2 to be the span of B_1X and B_3X the corresponding cascade decomposition would be trivial because the orbit of K^1 under K^2 contains K.

A similar cascade decomposition is possible for the fully controllable system on $SO(3)$. The familiar polar decomposition of a matrix differential equation is another example of a cascade using Theorem 4.

5. Indecomposable systems. Krohn and Rhodes not only showed that every finite state machine can be decomposed into a cascade of flip-flops and machines with simple groups, but also that these building blocks are irreducible in a certain sense (see [3, p. 307]).

In this section we show that a similar result holds for the decomposition described in Theorem 3. (We use the term indecomposable since irreducible already has a well defined meaning in the context of Lie algebras.)

A finite dimensional system Σ is *indecomposable* if whenever Σ admits a finite dimensional cascade decomposition $\Sigma^1 \circledcirc_v \Sigma^2$ of finite dimensional systems Σ^i then L divides at least one L^i. This is weaker than requiring that Σ admit only trivial cascade decompositions, and perhaps some words of explanation are in order.

Since Σ has a finite dimensional Lie algebra it is reasonable to require the same of the Σ^i's and $\Sigma^1 \circledcirc_v \Sigma^2$. The Lie algebra L of an indecomposable system need only divide L^i rather than the stronger condition that Σ^i be homomorphic to Σ. This rules out decompositions based on splitting the controls which do not split the algebra, as in the $SL(2, \mathbb{R})$ example of § 4. Moreover, the one dimensional system on the line is indecomposable under the former condition while it is not under the latter as we now show.

Σ is the one dimensional system on $\mathbb{R}$ and Σ^1, Σ^2 are both the one dimensional system on S^1 described in § 3. We define a parallel cascade $\Sigma^1 \circledcirc_v \Sigma^2$ with the linking map $v(\theta^1, u) = cu$ where c is an irrational constant. Let $G^1 \circledcirc_v G^2$ denote the group of the cascade, the orbit $G^1 \circledcirc_{v_1} G^2(0, 0)$ is a dense winding line on the torus $T^2 = S^1 \times S^1$ and can be mapped in a standard fashion onto $\mathbb{R}$. This defines a homomorphism of $\Sigma^1 \circledcirc_v \Sigma^2$ onto Σ. On the other hand there is no map of the state space of Σ^i onto the state space of Σ so clearly Σ^i is not homomorphic to Σ.

Theorem 3 allowed us to split up a finite dimensional system into a cascade of one dimensional systems and those with simple Lie algebras. The next theorem tells us that this is in some sense the best we can do.

THEOREM 5. *Suppose Σ is a finite dimensional system; then Σ is indecomposable iff the Lie algebra of Σ is one dimensional or simple.*

Proof. One way follows from Theorem 3, so suppose the Lie algebra L of Σ is one dimensional or simple. Let $\Sigma^1 \circledcirc_v \Sigma^2$ be a finite cascade decomposition of Σ where the Σ^i's are finite dimensionable systems. If L is one dimensional then it clearly divides any nontrivial Lie algebra and at least one L^i must be nontrivial. Therefore we restrict to the case where L is simple.

If Σ is simulated by $\Sigma^1 \circledcirc_v \Sigma^2$ then Σ can be simulated by the cascade of the fully controllable covers of Σ^1 and Σ^2 using the above lemma (in § 4). Since Σ^i and its fully controllable cover have the same Lie algebra L^i it is with no loss of generality that we assume that each Σ^i is fully controllable. Therefore the controls of Σ^i enter linearly and the dynamics of the cascade must look like

$$\dot{x}^1 = \Sigma u_i a_i(x^1),$$
$$\dot{x}^2 = \Sigma v_j(u, x^1) b_j(x^2).$$

For each $u \in \Omega^1$ and each j, $v_j(u, \cdot)$ is a scalar-valued function of x^1 and any vector field $h^1(\cdot) \in L^1$ acts on it by partial differentiation. Let P denote the orbit of $\{v_j(u, \cdot) : \forall j \text{ and } u \in \Omega\}$ under the action of L^1. In general this is a real infinite dimensional vector space.

Consider the product $P \otimes L^2$ consisting of all finite linear combinations of elements of L^2 with coefficients from P. In an obvious fashion this can be identified with the subalgebra of $V(M^1 \times M^2)$ consisting of vector fields whose projection in the M^1 direction are identically zero. Similarly L^1 can be identified with the subalgebra of $V(M^1 \times M^2)$ consisting of vector fields whose projections in M^2 direction are identically zero. It follows that

$$L^1 \circledcirc_v L^2 \subset L^1 + P \otimes L^2$$

where the right side is the semidirect sum of the subalgebra L^1 and ideal $P \otimes L^2$.

Let Q denote the subspace of P consisting of all functions which actually appear in the expansion of a vector field of $L^1 \circledcirc_v L^2$. Each such vector field involves only a finite number of functions of P and since $L^1 \circledcirc_v L^2$ is finite dimensional it follows that Q is also finite dimensional. Moreover

$$L^1 \circledcirc_v L^2 \subset L^1 + Q \otimes L^2.$$

Since Q is a finite dimensional space of functions on M^1 and L^2 is a Lie algebra of vector fields on M^2, $Q \otimes L^2$, as a Lie algebra, is a direct sum of a finite number of copies of L^2.

Since the cascade is homomorphic to Σ there exists a Lie algebra homomorphism $\gamma_* : L^1 \circledcirc_v L^2 \to L$. Let π_* denote the projection $\pi_* : L^1 + Q \otimes L^2 \to L^1$; by restricting π_* we obtain the following diagram:

$$\begin{array}{ccc} L^1 \circledcirc_v L^2 & \xrightarrow{\gamma_*} & L \\ \Big\downarrow{\scriptstyle \pi_*} & & \\ L^1 & & \end{array}$$

Let $I = \ker \pi_*$ restricted to $L^1 \circledcirc_v L^2$ and $J = \gamma_*(I)$; these are ideals in $L^1 \circledcirc_v L^2$

and L respectively. Since L is simple there are only two possibilities, $J=0$ or $J=L$.

Suppose $J=0$; then $\ker \pi_* \subset \ker \gamma_*$. The first homomorphism theorem implies that there exists a natural homomorphism of $\pi_*(L^1 \ominus_v L^2)$ onto L, i.e., L divides L^i.

Suppose $J=L$; then we can conclude that $\gamma_*: I \to L$ is onto. Moreover $I \subset Q \otimes L^2$ so we conclude that L divides $Q \otimes L^2$. Since this is a direct sum of a finite number of copies of L^2, using the first homomorphism theorem as before shows that L divides L^2. Q.E.D.

Remark. For the reader familiar with Krohn–Rhodes theory, $L^1 + Q \otimes L^2$ plays the role of the wreath product. Notice that the choice of Q depends on the linking map and there is not one choice that works for all possible linking maps.

6. Conclusion. We would like to suggest two areas where extension of the current line of research might prove fruitful. One nice thing about cascade decomposition of systems is that it exposes the dynamic relations between the state variables. For example from Theorem 3 we know that a system with a finite dimensional solvable Lie algebra admits a cascade decomposition of one dimensional systems. When described in local coordinates this amounts to a lower triangular form for the dynamics. Kelley [15] has suggested that this form would be useful in applying singular perturbation techniques to nonlinear problems.

A second area of research would be in introducing dynamic compensation or state variable feedback as has been done in the linear case by numerous authors. (We refer the reader to [16] for an excellent treatment and bibliography.)

The feedback law $v: M \times \Omega \to \Omega$ would in general be a nonlinear function so that the dynamics becomes

$$\dot{x} = f(x, v(x, u)).$$

To a certain extent we have already considered this by allowing the control law of the second system of the cascade to depend on the control and state of the first system, that is, we have allowed state variable feedforward.

Such a dynamic compensator v completely changes the Lie algebra of the system as we have defined it. (If L were redefined as the Lie algebra generated by $f(\cdot, u(\cdot))$ for all analytic $u: M \to \Omega$ it would not change.) it would be of interest to know when two systems are equivalent under dynamic compensation or when a given system is equivalent to a simpler type of system, i.e., cascade, a bilinear, or a linear system.

Acknowledgment. I would like to thank Roger Brockett for introducing me to Krohn–Rhodes theory and acknowledge the help he gave me while writing this paper.

REFERENCES

[1] K. B. KROHN AND J. L. RHODES, *Algebraic theory of machines I. The main decomposition theorem*, Trans. Amer. Math. Soc., 116 (1965), pp. 450–464.

[2] M. A. ARBIB, ed., *The Algebraic Theory of Machines, Languages, and Semigroups*, Academic Press, New York, 1968.

[3] ———, *Theories of Abstract Automata*, Prentice-Hall, Englewood Cliffs, NJ, 1968.
[4] H. J. SUSSMANN, *Observable realizations of finite dimensional nonlinear autonomous systems*, Math. Systems Theory, to appear.
[5] ———, *Existence and uniqueness of minimal realizations of nonlinear systems, I: Initialized systems*, Ibid., to appear.
[6] R. S. PALAIS, *A global formulation of the Lie theory of transformation groups*, Amer. Math. Soc., (1957), no. 22.
[7] T. NAGANO, *Linear differential systems with singularities and an application to transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.
[8] W. L. CHOW, *Uber System von Linearen Partiellian Differentielgleichungen erster Ordnung*, Math. Ann., 117 (1939), pp. 98–105.
[9] H. J. SUSSMANN AND V. J. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
[10] A. J. KRENER, *A generalization of Chow's theorem and the bang-bang theory to nonlinear control problems*, this Journal, 12 (1974), pp. 43–52.
[11] ———, *On the equivalence of control systems and the linearization of nonlinear systems*, this Journal, 11 (1973), pp. 670–676.
[12] ———, *Bilinear and nonlinear realizations of input-output maps*, this Journal, 13 (1974), pp. 827–833.
[13] E. WICHMANN, *Notes on the algebraic aspect of the integration of a system of ordinary linear differential equations*, J. Mathematical Phys., 2 (1961), pp. 876–880.
[14] R. W. BROCKETT, *On the algebraic structure of bilinear systems*, Variable Structure Control Systems, Mohler and Ruberti, eds., Academic Press, New York, 1972, pp. 153–168.
[15] H. J. KELLEY, *State variable selection and singular perturbations*, Singular Perturbations: Order Reduction in Control System Design, Kokotovic and Perkins, eds., ASME, New York, 1972, pp. 37–43.
[16] W. M. WONHAM, *Linear Multivariable Controls, A Geometric Approach*, Springer-Verlag, New York, 1974.
[17] R. HERMANN, *On the accessibility problem in control theory*, Internat. Symp. on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325–332.
[18] R. HERMANN AND A. J. KRENER, *Nonlinear controllability and observability*, to appear.
[19] K. T. CHEN, *Decomposition of differential equations*, Math. Ann., 146 (1962), pp. 263–278.
[20] S. LIE, *Allgemeine Untersuchungen über Differentialgleichungen, die eine kontinvierliche, endliche Gruppe gestaten*, Ibid., 27 (1885), pp. 71–151.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

SOME REMARKS ON INFINITE-DIMENSIONAL NONLINEAR CONTROL WITHOUT CONVEXITY*

GREG. KNOWLES†

Abstract. The time-optimal control of certain nonlinear distributed systems is considered, and existence and bang-bangness of optimal controls is proven without convexity assumptions on the set of admissible controls.

Introduction. The time-optimal control problems considered in [6] are taken up again in this paper, but with the aim of applying these results to problems where the control appears nonlinearly, and the set of admissible controls is not assumed convex. This work could be regarded as a partial extension to infinite dimensions of the article of C. Olech [10], the main difference being, however, that the "normality" of the systems considered here guarantees that the optimal controls belong to the extreme points of the set of admissible controls, thus avoiding the use of some infinite-dimensional extension of Lyapunov's theorem (which in fact need not hold for the problem, e.g. [6, § 5]). These results are applied to nonlinear control of parabolic equations.

Firstly the general situation in [5] is reconsidered, this time for integrably bounded set-valued functions whose values are compact, convex sets.

1. Control systems. In order to avoid excessive repetition we refer to [5] and [6] for most of the notation used in this note. X will denote a quasi-complete l.c.t.v.s., T a set, $\mathcal{S}$ a σ-algebra of subsets of T and $m = (m_i)$ a control system of vector measures $m_i: \mathcal{S} \to X$, $i = 1, 2, \cdots$. The family of compact, convex (respectively compact) subsets of X will be denoted by CCX (respectively CX). A set-valued function $F: T \to CR^\infty$ is called *measurable* if for each $x' \in (R^\infty)'$, the function

$$t \mapsto \sup\{\langle x', x\rangle : x \in F(t)\}, \qquad t \in T,$$

is measurable. In the case considered here, T is a Lebesgue measurable subset of R^m, $\mathcal{S}$ is the σ-algebra of Lebesgue sets on T and l is Lebesgue measurable on $\mathcal{S}$, $F: T \to CR^n$ is measurable in the above sense if and only if F is l-measurable in the sense of Castaing. (See [5, Lemma VII.8.2] and [1, Thm. 3.2]). We shall assume this fact whenever the implicit function theorem in [1] ([1, Cor. 5.2′]) is used in this note.

The set-valued function F is *bounded* if its values are contained in a bounded set in R^∞, and F is called *m-integrably bounded* (or just *integrably bounded* if it does not cause confusion), if there exists a real-valued measurable function h on T, such that

$$\sum_{i=1}^{\infty} \int_{E_i} h\, dm_i \in X, \quad \text{for any set } E_i \in \mathcal{S}, \quad i = 1, 2, \cdots,$$

* Received by the editors April 23, 1976.

† Institut für Angewandte Mathematik der Universität Bonn, 53 Bonn, West Germany. Currently at Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. Research supported by the Sonderforschungsbereich 72.

and a bounded set $\mathcal{O} \subset R^\infty$, with $F(t) \subset \mathcal{O} \cdot h(t)$, $t \in T$. By virtue of [5, Lemma IX.9.1] (and the fact that m is a control system) we can suppose $h(t) \neq 0, \forall t \in T$.

For an m-integrably bounded set-valued function F, define

$$L_F(R^\infty, m) = \{f: f = (f_i) \in \mathcal{M}(R^\infty, \mathcal{S}), f(t) \in F(t), t \in T\}$$

and

$$A_F(m) = \left\{ \sum_{i=1}^{\infty} \int f_i \, dm_i : f = (f_i) \in L_F(R^\infty, m) \right\}.$$

Then we have

THEOREM 1. *If $m = (m_i)$ is a control system, $F: T \to CCR^\infty$ is a measurable, m-integrably bounded set-valued function, then $A_F(m)$ is a weakly compact, convex set in X.*

Proof. Let h be the function bounding F and satisfying the integrability assumptions above. Then the set-valued function $H: T \to CCR$ defined by

$$H(t) = F(t)/h(t), \qquad t \in T,$$

is well-defined, bounded, and measurable. Define vector measures $n_i: \mathcal{S} \to X$, by $n_i(E) = \int_E h \, dm_i$ (by definition, h is m_i integrable, for every $i = 1, 2, \cdots$). Each such measure is closed, ([5, Thm. IV.7.2]) and for any sets $E_i \in \mathcal{S}$, $i = 1, 2, \cdots$, we have

$$\sum_{i=1}^{\infty} n(E_i) = \sum_{i=1}^{\infty} \int_{E_i} h \, dm_i \in X,$$

by the properties of h. In other words, $n = (n_i)$, is a control system, so by [5, Thm. IX.1.1.], $A_H(n)$ is weakly compact and convex in X. The result follows as $A_F(m) = A_H(n)$.

Finally we consider some properties of the extreme points of the attainable set which will be of use in the next section. Suppose $n = (n_i)$ is a control system and $F: T \to CCR^\infty$ an n-integrably bounded, measurable set-valued function. We say F and n have the *extreme point property* if $n(f) \in \operatorname{ex} A_F(n), f \in L_F(R^\infty, n)$, implies that $f \in L_{\operatorname{ex} F}(R^\infty, n)$. By virtue of [5, Thm. VIII.4.1], for any $f \in L_F(R^\infty, n)/L_{\operatorname{ex} F}(R^\infty, n)$, there exists a measurable function $v: T \to R^\infty$, not n-null, such that $f \pm v \in L_F(R^\infty, n)$. Clearly F and n have the extreme point property if this function v can be chosen with $n(v) \neq 0$.

If F and n have the extreme point property for any control system n, then for simplicity, we say that F has the extreme point property. It follows from [1, Thm. 4] that if $F(t)$ is a product of intervals, for each $t \in T$, then it has the extreme point property, and from the proof of [5, Lemma VIII.4.1], any measurable, bounded $F: T \to CCR^1$ has the extreme point property. This latter case will be the situation considered in the examples in this note. In fact both these cases are a consequence of the following geometric property of the set $F(t)$. We say a set $Y \subset R^\infty$ has *property* (B) if for every $x = (x_i) \in Y$, $x \notin \operatorname{ex} Y$, there exists a nonzero number u (which may depend on x) such that for some coordinate i_x, $(x_1, \cdots, x_{i_x} \pm u, \cdots) \in Y$. (Alternatively, this says that the boundary of Y doesn't contain any oblique line segments.) Then we have

LEMMA 1. *If $F(t)$ has property* (B) *for each $t \in T$, then F has the extreme point property.*

Proof. The proof is similar to that of [5, Thm. VIII.4.1], and so we only sketch the differences.

Suppose $m = (m_i)$ is a control system, and $f \in L_F(R^\infty, m)/L_{\text{ex}F}(R^\infty, m)$. Let

$$B_i = \{t \in T: \exists\, u \neq 0 \text{ with } (f_1(t), \cdots, f_i(t) \pm u, f_{i+1}(t), \cdots) \in F(t)\}$$

$i = 1, 2, \cdots$, and assume each B_i is m_i-null. Define

$$\tilde{F}(t) = \{x \in F(t): x_k = f_k(t) \text{ for every } k \text{ such that } t \notin B_k\}$$

$t \in T$, and let $\{e_k\}_{k=1}^\infty$ be the usual coordinate functionals on R^∞, and $\tilde{f}(t)$ the lexicographic maximum of $\tilde{F}(t)$ ordered by $\{e_k\}$ [5, § VIII.2], $t \in T$. Then $\tilde{f}$ is $\mathcal{S}$-measurable and $\tilde{f}(t) \in \text{ex}\, \tilde{F}(t)$, $t \in T$. However, by property (B), $\tilde{f}(t) \in \text{ex}\, F(t)$, and since f and $\tilde{f}$ are m-equivalent, we must have $f \in L_{\text{ex}F}(R^\infty, m)$, a contradiction. Consequently, some B_{i_0} is not m_{i_0}-negligible, and by following the proof of [5, Thm. VIII.4.1] we can find a measurable (real-valued) function u on T, not m_{i_0}-null, with $(f_1(t), \cdots, f_{i_0}(t) \pm u(t), \cdots) \in F(t)$, $t \in T$. As u is not m_{i_0}-null, there exists a set $E \in \mathcal{S}$ such that $\int_E u\, dm_{i_0} \neq 0$. Letting

$$v(t) = (\underbrace{0, \cdots, u(t)}_{i_0})\chi_E(t), 0, \cdots), \qquad t \in T,$$

it is easy to see v is not m-null, $f \pm v \in L_F(R^\infty, n)$ and $m(v) = m_{i_0}(u\chi_E) \neq 0$. Hence F has the extreme point property.

Taking $T = [0, 1]$, $\mathcal{S}$ the Borel sets on T, l the Lebesgue measure on $\mathcal{S}$, and defining $F(t) = \{(x, x): 0 \leqq x \leqq 1\}$, $t \in T$, and

$$m_1(E) = l(E \cap [0, 1/2)) - l(E \cap [1/2, 1]),$$
$$m_2(E) = l(E \cap [0, 1/2)) + l(E \cap [1/2, 1]),$$

gives an example of a set-valued function F not having the extreme point property.

2. Normal systems. In this section the situation discussed in [6] is reconsidered, and applied to some problems in nonlinear control.

Suppose Ω is a set in R^n (possibly empty) and for every $t \in [0, t_0]$, a given time interval, $\mathcal{S}_t$ is a σ-algebra of subsets of $\Omega \times [0, t_0]$. For each $t \in [0, t_0]$ we are also given a control system $m(t) = (m_i(t))$, $m_i(t)$: $\mathcal{S}_t \to X$, $i = 1, 2$. The control system is called *essentially normal in X* if for every (nonzero) $x' \in X'$, $m_i(t) \ll \langle x', m_i(t)\rangle$, $i = 1, 2, \cdots, t \in [0, t_0]$.

In this situation a set-valued function F: $\Omega \times [0, t_0] \to CR^\infty$ is called *integrably bounded* if the bounding function h (cf. § 1) satisfies

$$\sum_{i=1}^{\infty} \int \int_{E_i} h(\omega, \tau)\, dm_i(t)(\omega, \tau) \in X, \tag{1}$$

for any measurable set $E_i \in \mathcal{S}_t$, $i = 1, 2, \cdots$, and any $t \in [0, t_0)$. For such a set-valued function, define for $f = (f_i) \in L_F(R^\infty, m(t))$, and $t \in [0, t_0]$

$$m(t, f) = \sum_{i=1}^{\infty} \int_0^t \int_\Omega f_i(\omega, \tau)\, dm_i(t)(\omega, \tau). \tag{2}$$

The integral boundedness of F guarantees that $m(t, f) \in X$.

For the remainder of this section we suppose such an F is given and consider the problem of steering the system, whose output is described by (2), where the control function $f(\omega, \tau) = (f_i(\omega, \tau))$ is restricted to be measurable and take values in $F(\omega, \tau)$, $(\omega, \tau) \in \Omega \times [0, t_0]$, to reach a fixed closed convex set W of X, with nonempty interior, in minimum time. (We adopt the convention, as in [6], that when we are interested in the output of the system up to a specified time $t < t_0$, we consider the controls as functions on only $\Omega \times [0, t]$.) Then we have

THEOREM 2. *Suppose* $F\colon \Omega \times [0, t_0] \to CR^\infty$ *is measurable, integrably bounded, and*

(i) $A_{\overline{\text{co}}F}(m(0)) \cap W = \varnothing$,

(ii) *for some* $t_1 > 0$, $A_{\overline{\text{co}}F}(m(t_1)) \cap W \neq \varnothing$ *(we can clearly assume* $t_1 < t_0$*),*

(iii) *for each* $f \in L_{\overline{\text{co}}F}(R^\infty, m(t_0))$, *the function* $t \mapsto m(t, f)$, $0 < t < t_0$, *is continuous into the given topology on* X,

(iv) *for each* $x' \in X'$, *and for each* $t \in (0, t_0)$,

$$\sup\{|\langle x', m(t, f) - m(t^*, f)\rangle| : f \in L_{\overline{\text{co}}F}(R^\infty, m(t_0))\} \to 0$$

as $t \downarrow t^*$,

(v) *for any (nonzero)* $x' \in X'$ *and* $t \in (0, t_0)$, $\overline{\text{co}}\, F$ *and* $\langle x', m(t)\rangle$ *have the extreme point property,*

(vi) $m(t)$, $t \in (0, t_0)$ *is essentially normal in* X.

Then W *is reached in minimum time* $t^* > 0$ *by an essentially unique optimal control* f^* *with* $f^*(\omega, \tau) \in \text{ex}\, F(\omega, \tau)$, $(\omega, \tau) \in \Omega \times (0, t^*)$. *Also, there exists a nonzero* $x' \in X'$ *such that*

$$\sum_{i=1}^{\infty} \int_0^{t^*} \int_\Omega f_i\, d\langle x', m_i(t^*)\rangle \leqq \sum_{i=1}^{\infty} \int_0^{t^*} \int_\Omega f_i^*\, d\langle x', m_i(t^*)\rangle \leqq \langle x', w\rangle \tag{3}$$

for all $f \in L_{\overline{\text{co}}F}(R^\infty, m(t^*))$ *and all* $w \in W$.

Proof. We firstly show that the problem of reaching W in minimum time by controls taking values in $\overline{\text{co}}\, F$ has a unique optimal solution. The essential normality of the system will then imply that this solution is in fact a solution of the original problem.

Consequently, define as in Theorem 1, the bounded measurable set-valued function $H\colon \Omega \times [0, t_0] \to CCR^\infty$ by $H(\omega, \tau) = \overline{\text{co}}\, F(\omega, \tau)/(h(\omega, \tau))$ (where h is the function bounding F in (1), chosen to be nonzero) and the control system $n(t) = (n_i(t))$ where $n_i(t)$ is the indefinite integral of h with respect to $m_i(t)$, $t \in [0, t_0]$, $i = 1, 2, \cdots$. As the measures $m_i(t)$ and $n_i(t)$, $i = 1, 2, \cdots$, are equivalent, and $A_H(n(t)) = A_{\overline{\text{co}}F}(m(t))$, $t \in [0, t_0]$, the problem of minimizing the time t for which $A_{\overline{\text{co}}F}(m(t)) \cap W \neq \varnothing$ is equivalent to the problem of minimizing the time t for which $A_H(n(t)) \cap W \neq \varnothing$. This latter problem has a solution by [6, Lemma 1], as the conditions (A) and (C) there, follow from (ii) and (iii) above, and (B) is a consequence of (iv) above, for

$$\sup\{|\langle x', n(t, q) - n(t^*, q)\rangle| : q \in L_H(R^\infty, n(t_0))\}$$
$$= \sup\{|\langle x', m(t, f) - m(t^*, f)\rangle| : f \in L_{\overline{\text{co}}F}(R^\infty, m(t_0))\}$$

for any $x' \in X'$ and $t, t^* \in [0, t_0]$.

Accordingly let t^* be the minimum time, and $f^* \in L_{\overline{\text{co}}F}(R^\infty, m(t^*))$ an optimal control. As in [6, Lemma 2] there exists a nonzero $x' \in X'$ such that (3) holds, and consequently

$$\sum_{i=1}^{\infty} \int_0^{t^*} \int_\Omega f_i^* d\langle x', m_i(t^*)\rangle \in \text{ex } A_{\overline{\text{co}}F}(\langle x', m_i(t^*)\rangle).$$

By the extreme point property (v) $f^* \in L_{\text{ex}(\overline{\text{co}}F)}(R^\infty, (\langle x', m_i(t^*)\rangle)) = L_{\text{ex}(\overline{\text{co}}F)}(R^\infty, m(t^*))$, as $m(t^*)$ is essentially normal. Hence f^* is an essentially unique optimal control and we may choose $f^*(\omega, \tau) \in \text{ex } \overline{\text{co}}\, F(\omega, \tau)$, $(\omega, \tau) \in \Omega \times (0, t^*)$. However the sets $\overline{\text{co}}\, F(\omega, \tau)$, and $F(\omega, \tau)$ are compact, and so by [2, Lemma V.8.5] $\text{ex } \overline{\text{co}}\, F(\omega, \tau) \subseteq \text{ex } F(\omega, \tau)$, $(\omega, \tau) \in \Omega \times (0, t^*)$, that is, f^* is, in fact, the essentially unique optimal control for the original problem.

As an example of the application of this theorem to nonlinear control, consider the following parabolic control problems.

Suppose Ω is a bounded domain in R^n, $\alpha = (\alpha_1, \cdots, \alpha_n)$ and $D^\alpha = D_1^{\alpha_1} \cdots D_n^{\alpha_n}$, where $D_j = \partial/\partial x_j$, and set $|\alpha| = \alpha_1 + \cdots + \alpha_n$. For a positive integer m consider the parabolic differential operator

$$Lu \equiv \frac{\partial u}{\partial t} - A(x, t, D)u \equiv \frac{\partial u}{\partial t} - \sum_{|\alpha| \leqq 2m} a_\alpha(x, t) D^\alpha u. \tag{4}$$

We assume that the boundary of Ω and the coefficients a_α are sufficiently smooth (in a sense to be made precise later) and that L is parabolic in the sense of Petrowski in $\bar{\Omega} \times [0, \infty)$.

Firstly take the case $m = 1$ and consider the second order boundary control problem

$$Lu = g \quad \text{in } \Omega \times (0, t_0), \tag{5}$$

$$u(x, 0) = \Phi(x), \qquad x \in \Omega, \tag{6}$$

$$\partial u/\partial v + a(x, t)u = \Psi(x, t, p(x, t)), \qquad (x, t) \in \partial\Omega \times (0, t_0), \tag{7}$$

where $\partial/\partial v$ is the outward transversal derivative along the lateral boundary (e.g. [4]), g and Φ are fixed smooth functions, p is regarded as the control function chosen to be measurable and have values in a compact (not necessarily convex) set $U \subset R$, and $\Psi: \partial\Omega \times (0, t_0) \times U \to R$ is a (fixed) measurable function which, for each $u \in U$, is Lebesgue measurable in (x, t), and for each $(x, t) \in \partial\Omega \times (0, t_0)$ is continuous in u. Suppose also that a $q > 1$ and a function $h \in L^q(\partial\Omega)$ are given and for which $|\Psi(x, t, u)| \leqq h(x)$, for all $(x, t, u) \in \partial\Omega \times (0, t_0) \times U$.

If $q > n - 1$, choose a number $s \in [1, \infty)$ and let W be a closed, convex subset of $L^s(\Omega \times [0, t_0])$ with nonempty interior such that $\Phi \notin W$. Suppose also that W is reached in some time $t_1 \in (0, t_0]$ in the sense that there exists a measurable function p for which $\Psi(x, t, p(x, t)) \in \overline{\text{co}}\,\{\Psi(x, t, v): v \in U\}$, $(x, t) \in \partial\Omega \times (0, t_1)$, and if u is the solution of (5), (6), (7) for this function p, then $u(\cdot, t_1) \in W$. In this case we have

THEOREM 3. *If $a(x, t)$ and the coefficients of L are analytic functions, and $\partial\Omega$ is an analytic manifold, then under the above assumptions on W, W is reached in minimum time t^* by an optimal control p^*, with*

$$\Psi(x, t, p^*(x, t)) \in \text{ex}\,\{\Psi(x, t, v): v \in U\}, \qquad (x, t) \in \partial\Omega \times (0, t^*). \tag{8}$$

Proof. As in [6, Thm. 5], we can represent the solution of (5), (6), (7) in the form

$$u(x,t)=\int_0^t\int_{\partial\Omega}\Psi(\omega,\tau,p(\omega,\tau))\,dm(t)(\omega,\tau),$$

$(x,t)\in\partial\Omega\times[0,t_0)$, where $m(t)\colon \mathcal{S}_t\to L^s(\Omega)$ is the vector measure derived from the Green's function G for this problem (see [6, § 5]) and $\mathcal{S}_t$ is the usual σ-algebra of subsets of $\partial\Omega\times[0,t]$. (Each measure $m(t)$ takes values in $L^s(\Omega)$ by [3, Lemma 1].) It is also a consequence of [6, Lemma 5] that $m(t)$ is normal in $L^s(\Omega)$ for any $0<t\leqq t_0$, and the assumptions on s imply that h is $m(t)$-integrable, $0<t\leqq t_0$, and that $u\in L^s(\Omega\times[0,t_0))$ [3, Lemma 1].

Define the measurable set-valued function $F\colon \partial\Omega\times(0,t_0)\to CR$ by $F(\omega,\tau)=\Psi(\omega,\tau,U)=\{\Psi(\omega,\tau,v)\colon v\in U\}$, and apply Theorem 2 to this set-valued function and the vector measures $m(t)$, $0<t<t_0$. The set-valued functions F and $\overline{\text{co}}\,F$ are integrably bounded by h, conditions (i) and (ii) hold from our initial assumptions, and (v) by the remarks above. We next prove (iii) and (iv).

Suppose $t\in(0,t_0)$ is given. Then for $f\in L_{\overline{\text{co}}F}(R,m(t_0))$ and $p>0$

$$\begin{aligned}\|m(t+p,f)-m(t,f)\|_{L^s(\Omega)}\leqq&\left\|\int_t^{t+p}\int_{\partial\Omega}G(\cdot\,,t+p,\xi,\tau)f(\xi,\tau)\,d\lambda\,(\xi)\,d\tau\right\|_{L^s(\Omega)}\\&+\left\|\int_0^t\int_{\partial\Omega}[G(\cdot\,,t+p,\xi,\tau)-G(\cdot\,,t,\xi,\tau)]f(\xi,\tau)\,d\lambda\,(\xi)\,d\tau\right\|_{L^s(\Omega)}\end{aligned}\tag{9}$$

where λ is the usual surface measure on $\partial\Omega$. To prove (iii) and (iv) of Theorem 2 we will only show that

$$\sup\left\{\left\|\int_0^t\int_{\partial\Omega}[G(\cdot\,,t+p,\xi,\tau)-G(\cdot\,,t,\xi,\tau)]f(\xi,\tau)\,d\lambda\,(\xi)\,d\tau\right\|_{L^s(\Omega)}\colon f,\ |f(\xi,\tau)|\leqq h(\xi)\ \text{ a.e.}\right\}\to 0\quad\text{as } p\downarrow 0,$$

as the convergence of the other integral in (9) follows similarly.

Since $1\leqq s<\infty$, the result will follow by dominated convergence, if we can show that for each $x\in\Omega$,

$$I_p(x)=\int_0^t\int_{\partial\Omega}|G(x,t+p,\xi,\tau)-G(x,t,\xi,\tau)|h(\xi)\,d\lambda\,(\xi)\,d\tau\to 0\tag{10}$$

as $p\downarrow 0$, and

$$I_p(x)\leqq K,\qquad K\text{ some constant},\tag{11}$$

for all $x\in\Omega$ and p sufficiently small.

Firstly we will prove that the function

$$\tau\to\int_{\partial\Omega}|G(x,\bar t,\xi,\tau)|h(\xi)\,d\lambda\,(\xi),\qquad\tau\in[0,t],\tag{12}$$

is integrable over $[0,t]$ for each (fixed) $\bar t\in[t,t+p]$ and each (fixed) $x\in\Omega$. From the properties of the Green's function it clearly suffices to suppose that x lies in a sufficiently small neighbourhood of $\partial\Omega$ and ξ lies in a neighborhood of x. That is, if Ω' is some $(n-1)$-dimensional domain, then (using the estimate [3, (2.3)] for G) it

suffices to show that, for each $x' \in \Omega'$ and each $s \in [0, s_0]$, say, the function

$$(13) \quad Q(x', s, \tau) = \int_{\Omega'} \frac{1}{(\bar{t}-\tau)^{n/2}} \exp\left(\frac{(x'-\xi')^2}{(\bar{t}-\tau)}\right) \exp\left(\frac{s^2}{(\bar{t}-\tau)}\right) h'(\xi')\, d\xi'$$

is integrable as a function of τ over $[0, t]$, where $h' \in L_q(\Omega')$. We may further suppose $s = 0$ in (13). Then by Hölder's inequality with exponents q, r $(1/q + 1/r = 1)$,

$$|Q(x', 0, \tau)| \leqq \|h\|_{L^q(\Omega')} \left(\int_{\Omega'} \left| \frac{1}{(\bar{t}-\tau)^{n/2}} \exp\left(\frac{|x'-\xi'|^2}{\bar{t}-\tau}\right) \right|^r d\xi' \right)^{1/r}$$

and substituting $\rho = |x-\xi|/\sqrt{\bar{t}-\tau}$, we have

$$(14) \qquad |Q(x', 0, \tau)| \leqq c/(\bar{t}-\tau)^{\mu}$$

where $\mu = (nr/2 - (n-1)/2) \cdot (1/r)$, and c is independent of x and p. As $q > n-1$, $\mu < 1$, and Q is integrable.

Now suppose $x \in \Omega$ is fixed and $\varepsilon > 0$. The integrability of the function (12) allows us to find a $t_\varepsilon \in (0, t)$ such that for all p sufficiently small (say $p < \delta_1$)

$$(15) \qquad \int_{t_\varepsilon}^{t} |G(x, t+p, \xi, \tau) - G(x, t, \xi, \tau)| h(\xi)\, d\lambda(\xi)\, d\tau < \varepsilon.$$

However, the function

$$(16) \qquad \int_0^{t_\varepsilon} \int_{\partial\Omega} |G(x, t+p, \xi, \tau) - G(x, t, \xi, \tau)| h(\xi)\, d\lambda(\xi)\, d\tau$$

is continuous in p and so there exists a $\delta_2 < \delta_1$ such that for $p < \delta_2$ the integral (16) is less than ε. Assertion (10) follows by adding the inequalities (15), (16), and assertion (11) can be proved similarly, as the constant c in (14) is independent of x and p.

By Theorem 2 the optimal time t^* exists, and there is an $(m(t^*)\text{- })$ essentially unique optimal control $f^* \in L_{\mathrm{ex}\,F}(R, m(t^*))$, and a nonzero $x' \in (L^s(\Omega))'$ such that

$$(17) \qquad \int_0^{t^*} \int_{\partial\Omega} f d\langle x', m(t^*)\rangle \leqq \int_0^{t^*} \int_{\partial\Omega} f^* d\langle x', m(t^*)\rangle$$

for any $f \in L_{\overline{\mathrm{co}}\,F}(R, m(t^*))$. By redefining f^* on an $m(t^*)$-null set, we have $f^*(\omega, \tau) \in \mathrm{ex}\, F(\omega, \tau)$, $(\omega, \tau) \in \partial\Omega \times [0, t^*]$, and then by [1, Cor. 5.2′] there exists an admissible control p^* such that $\Psi(\omega, \tau, p^*(\omega, \tau)) = f^*(\omega, \tau)$ all (ω, τ). Clearly this function p^* is an optimal control with the desired properties.

Remark 1. If, as in the proof of Lemma 5 in [6], we write the measure $\langle x', m(t^*)\rangle(E) = \iint_E K(\omega, \tau)\, d\lambda(\omega)\, d\tau$, where λ is the surface measure on $\partial\Omega$. $E \in \mathcal{S}_t$, and K is a nonzero integrable function, then (17) becomes,

$$\int_0^{t^*} \int_{\partial\Omega} \Psi(\omega, \tau, p(\omega, \tau)) K(\omega, \tau)\, d\lambda(\omega)\, d\tau \leqq \int_0^{t^*} \int_{\partial\Omega} \Psi(\omega, \tau, p^*(\omega, \tau)) K(\omega, \tau)\, d\lambda(\omega)\, d\tau$$

for all admissible controls p. It is then a consequence of [10, Thm. 7.1(i)], that there exists a nonzero real number η for which

$$\eta \cdot \Psi(\omega, \tau, p^*(\omega, \tau))K(\omega, \tau) = \max_{v \in U} (\eta \cdot \Psi(\omega, \tau, v)K(\omega, \tau)) \tag{18}$$

for almost all $(\omega, \tau) \in \partial\Omega \times [0, t^*]$. Equation (18) could be regarded as an analogue of the Pontryagin maximum principle. As f^* is unique (a.e.), the optimal control p^* will be unique (a.e.), if the function $v \to \Psi(\omega, \tau, v)$, $v \in U$, is invertible, for almost all (ω, τ).

Next we consider the action of distributed control, that is Ψ in (7) is kept fixed (and assumed to be smooth) and condition (5) is replaced by

$$Lu(x, t) = g(x, t, p(x, t)), \qquad (x, t) \in \Omega \times (0, t_0), \tag{19}$$

where p is regarded as the control function, measurable and taking its values in a compact set $U \subset R$, and g has analogous continuity and measurability assumptions to Ψ in Theorem 3, with the bounding function $h \in L^q(\Omega)$, for some q. Suppose the target set W is a closed convex subset of $L^s(\Omega)$ with nonempty interior, where $1 \leqq s < \infty$ if $q > n/2$ and $1 \leqq s \leqq q$ if $q \leqq n/2$, W is reached (in an analogous sense to that used in Theorem 3) in some time $t_1 \in (0, t_0]$, and $\Phi \notin W$. For this problem the regularity assumptions on the coefficients of L can be considerably weakened.

THEOREM 4. *Suppose $a(x, t)$ is smooth, the coefficients of L are $C^q(\Omega)$, $q = 2n + 2[(n+1)/2] + 11$, and L^* has the weak backward uniqueness property* [4, Chap. 9]. *Then under the above assumptions on W, W is reached in minimum time $t^* > 0$ by an optimal control p^*, and $g(x, t, p^*(x, t)) \in \operatorname{ex}\{g(x, t, v): v \in U\}$ a.e.*

Proof. As in [6, Thm. 6] we represent the solution of (19), (6), (7) in terms of the Green's function, G, for this system. Namely,

$$u(x, t) = \beta(x, t) + \int_\Omega \int_0^t G(x, t, \omega, \tau) g(\omega, \tau, p(\omega, \tau))\, d\omega\, d\tau \tag{20}$$

for $(x, t) \in \Omega \times [0, t_0]$, where β is a fixed smooth function which, without loss of generality, we take as zero. Define vector measures $m(t): \mathcal{S}_t \to L^s(\Omega)$, $\mathcal{S}_t$ the Lebesgue measurable subsets of $\Omega \times [0, t_0]$, by

$$m(t)(E) = \iint_E G(\cdot, t, \omega, \tau)\, d\omega\, d\tau, \qquad E \in \mathcal{S}_t,$$

and represent the solution (20) as

$$u(\cdot, t) = \int_\Omega \int_0^t g(\omega, \tau, p(\omega, \tau))\, dm(t)(\omega, \tau).$$

Then a similar argument to Theorem 3 shows $u \in L^s(\Omega \times [0, t_0])$, h is $m(t)$-integrable, $0 \leqq t \leqq t_0$, and (iii), (iv) of Theorem 2 hold. To complete the proof we will show the measures $m(t)$, $0 < t \leqq t_0$, are normal in $L^s(\Omega)$. This amounts to proving (see [6, § 5]), that for each $t^* \in (0, t_0]$, and each $x' \in [L^s(\Omega)]'$, $x' \neq 0$, the

function

$$K(\omega, \tau) = \int_\Omega x'(x) G(x, t^*, \omega, \tau)\, dx, \qquad (\omega, \tau) \in \Omega \times (0, t^*),$$

is nonzero almost everywhere.

Suppose t^* is fixed, and K is zero on some nonnull set $E \subset \Omega \times [0, t^*)$. Then for t sufficiently close to t^* $(t < t^*)$ K must also be zero on $D = \Omega \times [0, t) \cap E$. Hence by Fubini's theorem, there exists a set $\Delta \subset (0, t)$ of positive measure, such that for each $\tau \in \Delta$, $D^\tau = \{\omega : (\omega, \tau) \in D\}$ has nonzero n-dimensional Lebesgue measure. By the properties of the Green's function, $L^* K(\omega, \tau) = 0$ on $\Omega \times [0, t)$, and so by [7, Thm. 5] (and the remarks preceding this theorem), $K = 0$ a.e. on $\Omega \times \Delta$. Then, by weak backward uniqueness [4, Chap. 9], $K = 0$ a.e. on $\Omega \times [0, t)$. Taking a sequence $t \uparrow t^*$, we see that $x'(x) = 0$ a.e. on Ω, which contradicts our initial assumption.

Remark 2. The result of Theorem 4 holds for the general parabolic equation (4) with more general boundary conditions provided the coefficients of L and of the boundary operators are analytic, and $\partial\Omega$ is an analytic manifold. For exact details see [6, Thm. 6].

An analogue of the maximum principle also holds for this problem (cf. Remark 1), as do the uniqueness properties mentioned there.

The above examples have considered problems with only one control variable; however, problems with more than one boundary control or with boundary and distributed control can be solved using Theorem 2 (when condition (v) there holds), as the solution of such problems can be represented as a sum of the type (2), with, in general, as many terms in the summation as there are control variables. The verification of the other conditions of Theorem 2 follows as in the examples above.

The above considerations also extend to control governed by other performance indexes. For example consider the problem of minimizing the cost functional

$$I(p) = \int_\Omega \int_0^{t_0} |u(x, t_0, p) - z_d(x)|^2\, dx$$

where t_0 is a fixed time, $z_d \in L^2(\Omega)$ is a given function and $u(\cdot, t_0, p)$ is the solution at time t_0 of (5), (6) and the boundary condition

$$(\partial u / \partial v)(x, t, p) = \Psi(x, t, p(x, t)), \qquad (x, t) \in \partial\Omega \times (0, t_0).$$

The function p is the control function admissible if it is measurable and takes values in a compact set $U \subset R$, and Ψ is a measurable function: $\partial\Omega \times (0, t_0) \times U \to R$ satisfying the Carathéodory conditions of Theorem 3 and also $|\Psi(x, t, v)| \leqq h(x, t)$, $(x, t, v) \in \partial\Omega \times (0, t_0) \times U$, for some function $h \in L^2(\partial\Omega \times (0, t_0))$. We further suppose $I(f) > 0$ for all measurable functions f with $f(x, t) \in \overline{\text{co}}\, \Psi(x, t, U)$, $(x, t) \in \partial\Omega \times (0, t_0)$.

THEOREM 5. *If the coefficients of L, and* $\partial\Omega$, *are analytic, then an optimal control,* p^*, *for the above problem exists and is characterized by*

$$L^*y(p^*)=0 \qquad \text{in } \Omega\times(0,t_0),$$
$$(\partial y/\partial v)(p^*)=0 \quad \text{in } \partial\Omega\times(0,t_0),$$
$$y(x,t_0,p^*)=u(x,t_0,p^*)-z_d(x), \qquad x\in\Omega,$$

and

$$\int_{\partial\Omega}\int_0^{t_0} y(\omega,\tau,p^*)(\Psi(\omega,\tau,p^*(\omega,\tau))-\Psi(\omega,\tau,p(\omega,\tau)))\,d\omega\,d\tau\geqq 0$$

for all admissible controls p. The optimal control $p^*(\omega,\tau)\in \operatorname{ex}\Psi(\omega,\tau,U)$ *a.e.*

Proof. As before, set $F(\omega,\tau)=\Psi(\omega,\tau,U)$ $(\omega,\tau)\in\partial\Omega\times(0,t_0)$. Then by the implicit function argument used previously, the above problem is equivalent to minimizing $I(f)$, $f\in L_F=\{f: f \text{ measurable}, f(\omega,\tau)\in F(\omega,\tau), (\omega,\tau)\in\partial\Omega\times(0,t_0)\}$, where $u(x,t,f)$ is defined as the solution of

$$Lu=g \quad \text{in } \Omega\times(0,t_0) \tag{21}$$

$$u(x,0,f)=\Phi(x), \qquad x\in\Omega, \tag{22}$$

$$(\partial u/\partial v)(x,t,f)=f(x,t), \qquad (x,t)\in\partial\Omega\times(0,t_0). \tag{23}$$

The set $L_{\overline{\operatorname{co}}F}$ is, by our assumptions, a bounded subset of $L^2(\partial\Omega\times(0,t_0))$, and it is easy to show it is, also, closed and convex. Hence the problem of minimizing $I(f)$ over all functions $f\in L_{\overline{\operatorname{co}}F}$ subject to (21), (22), (23) has an optimal solution, f^*, which is characterized by the adjoint problem [9, § III.2.2],

$$L^*y(f^*)=0 \qquad \text{in } \Omega\times(0,t_0),$$
$$(\partial y/\partial v)(f^*)=0 \quad \text{on } \partial\Omega\times(0,t_0),$$
$$y(x,t_0,f^*)=u(x,t_0,f^*)-z_d(x), \qquad x\in\Omega,$$

and

$$\int_0^{t^*}\int_{\partial\Omega} y(\omega,\tau,f^*)(f^*(\omega,\tau)-f(\omega,\tau))\,d\omega\,d\tau\geqq 0 \tag{24}$$

for all $f\in L_{\overline{\operatorname{co}}F}$.

As $z_d\neq u(\cdot,t_0,f^*)$ the analyticity assumptions guarantee that $y(f^*)$ is nonzero a.e. in $\Omega\times(0,t_0)$ (cf. Theorem 3 or [9, Lemma III.3.1]). From (24), $f^*\in \operatorname{ex}L_{\overline{\operatorname{co}}F}=L_{\operatorname{ex}\overline{\operatorname{co}}F}\subseteq L_{\operatorname{ex}F}$ by our earlier remarks. The implicit function theorem then gives an admissible control p^* such that $f^*(\omega,\tau)=\Psi(\omega,\tau,p^*(\omega,\tau))$ a.e., $(\omega,\tau)\in\partial\Omega\times(0,t_0)$, and it is easily seen that p^* is the desired optimal control.

In problems with only one (real-valued) control function (e.g., Theorems 3, 4, 5) for which the set U (where the admissible controls take their values) is compact and convex, the existence of an optimal control can be proven without assuming normality or int $W\neq\varnothing$. For example, in Theorem 3, the assumed properties of the function Ψ imply that the set $F(\omega,\tau)=\Psi(\omega,\tau,U)$ is pathwise connected for each (ω,τ), but as this set is contained in R, it must be actually convex. The

existence of an optimal control (even if int $W = \varnothing$) follows under conditions (ii) and (iv) of Theorem 2 (cf. [6, Lemma 1]), and if int $W \neq \varnothing$ and (iii) holds, the necessary condition (17) also holds. Of course, this optimal control cannot be guaranteed to be bang-bang or unique.

Lastly we remark that all of the previous results extend to the case the set $U = U(\omega, \tau)$ varies with (ω, τ). For this it is sufficient to assume that the correspondence $(\omega, \tau) \mapsto U(\omega, \tau)$ is a measurable, compact-valued, set-valued function.

Acknowledgment. I would like to thank Professor Frehse for his help in some aspects of this work.

REFERENCES

[1] C. CASTAING, *Sur les multi-applications mesurables*, Rev. Française Informat. Recherche Operationnelle, 1 (1967), pp. 91–126.
[2] N. S. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Part I, Interscience, New York, 1958.
[3] A. FRIEDMAN, *Optimal control for parabolic equations*, J. Math. Anal. Appl., 18 (1967), pp. 479–491.
[4] ———, *Partial Differential Equations*, Holt, Reinhart and Winston, New York, 1969.
[5] I. KLUVÁNEK AND G. KNOWLES, *Vector Measures and Control Systems*, North-Holland, Amsterdam, 1975.
[6] G. KNOWLES, *Time optimal control of infinite-dimensional systems*, this Journal, pp. 919–933.
[7] O. A. LADYŽENSKAJA, V. A. SOLONNIKOV, AND N. N. URAL'CEVA, *Linear and Quasi-linear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.
[8] E. M. LANDIS AND O. A. OLEINIK, *Generalized analyticity and some related properties of solutions of elliptic and parabolic equations*, Russian Math. Surveys, 29 (1974), pp. 195–212.
[9] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin-Heidelberg-New York, 1971.
[10] C. OLECH, *Extremal solutions of a control system*, J. Differential Equations, 2 (1966), pp. 74–101.
[11] I. KLUVÁNEK, *The range of a vector-valued measure*, Math. Systems Theory, 7 (1973), pp. 44–54.
[12] L. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

DIFFERENTIAL SEARCH GAMES WITH MOBILE HIDER*

JOE G. FOREMAN†

Abstract. In differential search games with a mobile hider of the princess and the monster type formulated by Isaacs, two blind players P, the pursuer, and E, the evader, are initially distributed in a playing space S and may move therein on paths determined by their control variables, P being constrained to move no faster than a given maximum speed. The game terminates at a given time T, and the payoff is the capture probability which P maximizes and E minimizes.

Two such games are analyzed; in the first S is a circle, and in the second S is a region of the plane. The game where S is a circle is solved for very general initial relative distributions of the players in S and all termination times T.

The game where S is a region of the plane is solved for the equiprobable relative initial distribution of the two players in a region far from the boundary of S relative to the distance that P can travel in time T (that is for situations in which the boundary is sufficiently far away as to make no difference).

1. Introduction. Differential search games with a mobile hider are examples of games with limited information and as such differ considerably from the differential games with perfect information that have been successfully analyzed [1]. Games with limited information differ in that mixed strategies are required for optimal play. These games occur frequently and have practical value; yet there is no systematic method for their solution.

A representative search game with mobile hider that embodies the quintessence of the problem is due to Isaacs [1, Chap. 12].

The princess and the monster. The monster P searches for the princess E, the time required being the payoff. They are both in a totally dark room (of any shape), but they are each cognizant of its boundary (possibly through small light-admitting performations high in the walls). Capture means the distance $PE \leqq l$, a quantity small in comparison with the dimension of the room. The monster, supposed highly intelligent, moves at a known fixed speed. We permit the princess full freedom of locomotion.

The version of the princess and the monster game that is of concern here is the one in which P has a given maximum speed which we normalized to unity, and the game terminates at time T if capture has not occurred (such a game is said to have a fuel constraint), and the payoff is the probability of capture. In § 3 this game is analyzed where the players are initially distributed equiprobably in a region of the plane far from the boundary. The game for large T (in which the boundaries play a role) is unsolved.

Literature. As an aid toward the solution of the princess and the monster game Isaacs also introduced [1, Chap. 12] the one-dimensional game on the (boundary of) the circle. The continuous version of this game has been studied [2], [3] for the case in which the players are distributed equiprobably on the circle at the start, and the payoff is the expected length of the game. In a discrete version of this game [1, Chap. 12], each player at the start occupies one of $N(\geqq 3)$ points on the circle, and on each move the players move one point clockwise or

* Received by the editors June 16, 1975, and in revised form July 27, 1976.
† Space Systems Division, Naval Research Laboratory, Washington, D.C. 20375.

counterclockwise. Capture occurs when the players occupy the same point or exchange positions during a move. The discrete game has been studied for the case where the payoff is the expected length of the game [4], [5]. In [4] the game is analyzed for the case in which both players know the starting positions of each.

The name search game or search-hide game is used to distinguish the subject matter from search optimization problems or one-player games. Frequently this distinction has not been made in the literature and the term search theory has been applied to various search problems including two-player games. During World War II a large amount of work was done in the general area of search theory, but for the most part this work was not approached from a game theoretic point of view. The results of this early work are summarized in the works of Bernard O. Koopman. A bibliography of the literature through 1965 is available [6]. A min–max pursuit problem (not actually a game, but similar to the present problem) in which the evader knows the itinerary of the pursuer has been studied [7].

The payoff structure and the differential nature of the game. The first payoff structure to be considered is the one in which a partie (a partie is a particular playing of the game) of the game terminates only when capture occurs, and the payoff is the expected length of the game. For the discrete game this payoff is

$$\text{pay}(\phi, \psi) = \sum_{i=1}^{\infty} i \,\text{cap}(i, \phi, \psi) \tag{1.1}$$

where i is the move number and cap (i, ϕ, ψ) is the probability that capture occurs on the ith move, given the strategies ϕ and ψ for P and E respectively. Here and throughout this paper P is the maximizing player and E is the minimizing player.

This payoff is a particular case of the more general payoff

$$\text{pay}(\phi, \psi) = \sum_{i=1}^{\infty} U(i) \,\text{cap}(i, \phi, \psi) \tag{1.2}$$

where $U(i)$ is the weight to be given to capture on the ith move. In some games with payoff (1.2) the optimum strategies are independent of the function U. This is the case if, for example, there exists strategies $\bar{\phi}$ for P and $\bar{\psi}$ for E such that

$$\text{cap}(i, \bar{\phi}, \psi) = \text{cap}(i, \phi, \bar{\psi}) = k_i \tag{1.3}$$

for every ϕ and ψ. Then $\bar{\phi}$ and $\bar{\psi}$ are optimum strategies, the value of the game $V = \sum_{i=1}^{\infty} U(i)k_i$, and the payoff $= V$ if either P uses $\bar{\phi}$ or E uses $\bar{\psi}$.

Another case in which the optimum strategies are independent of the U function occurs in a continuous game on the circle within which the players are initially equiprobably distributed [3]. In this game the continuous version of (1.3) does not hold since E can, by playing poorly, do worse than the value of the game. Furthermore, in this game the optimum strategies are no longer independent of $U(i)$ for more general initial distributions.

The discrete game which terminates after a given number n of moves (fuel constraint) and in which the payoff is the capture probability also fits into the general scheme (1.2). For this case we have $U(i) = 1$ for $1 \leqq i \leqq n$, and $U(i) = 0$ for

$i > n$. The payoff is

$$\begin{aligned} \text{pay}\,(\phi, \psi) &= \sum_{i=1}^{n} \text{cap}\,(i, \phi, \psi) \\ &= \text{Cap}\,(n, \phi, \psi) \end{aligned} \tag{1.4}$$

the cumulative probability of capture. In the following sections we are concerned with games whose payoffs are continuous versions of (1.4).

We are generally concerned here with differential search games in which both players may move at finite speeds to adjacent parts of the playing space. This playing space may be discrete, continuous, or a combination of both. It is the constraint on the motion of the players induced by the connectedness of the playing space that makes many of these problems difficult to solve.

For example, consider the "completely connected" discrete game in which to start the game both players move into one of N discrete cells and capture occurs if they choose the same cell. If capture does not occur on the first move, then for the second move the players may again move into any of the N cells, and so on until capture occurs. If the payoff is given by (1.2), then the optimum strategy for both players is to move equiprobably to each cell on each turn regardless of the values of $U(i)$. This is because each move is independent of the previous move. In this game $k_i = (1/N)\,(1-1/N)^{i-1}$.

However, if the N cells are arranged on, say, a rectangular grid and the players are permitted to move only to one of the at most 5 adjacent cells (counting the position of the player before the move) on each turn, then the game has become a discrete princess and monster game. Different games are obtained depending on the shape of the playing space. The boundaries of the playing space produce particular difficulties in obtaining solutions. Other factors in the definition of a game are the payoff structure, the initial starting conditions, and the speed constraints.

Initial conditions. A game has not been defined until the initial placement of the players in the playing space and the information given to each about the other's whereabouts has been specified (usually probabilistically). We always assume that, once the game starts, no further information is given the players, although each is constantly aware of his own position. The game may be started by allowing the players to select their starting positions or such may be governed by a prescribed probability distribution. The amount of information granted to the players about the other's initial placement may vary from none at all to exact information. A case sufficiently general for us is the one in which the players are given that the initial positions are chosen from a joint probability distribution. Then at the start of the partie each player, knowing his own position, can calculate the conditions probability distribution of the other.

In the game on the circle of § 2 it is assumed that the players are distributed independently in their initial placement on the circle. Thus, due to the symmetry of the problem, only the relative positions of the players matter, and a single variate distribution is sufficient to describe the relative initial placement. In § 3 the game in the plane is analyzed with an equiprobable initial distribution.

A game in which no information is given to each player about the location of the other may not have optimum strategies or value; the payoff matrix cannot be contructed. However, in some cases, such as a game treated in § 2, the game may have optimum strategies and value because the optimum strategies are optimum for every initial distribution.

Kinematic constraints and allowable pure strategies. In the continuous games analyzed in the following section, the pursuer is contrained to a maximum possible speed (normalized to unity) and the evader is unconstrained. If the evader is constrained to move no faster than a maximum speed which is less than the pursuer's, the game is changed considerably. The extreme case where the evader cannot move at all (immobile hider) has been analyzed by Isaacs [1, Chap. 12].

A pure strategy for E is required to be continuous and have a finite forward derivative at each $t \in [0, T]$. The class of all such is denoted by F. A pure strategy for P has the further requirement that the forward time derivative is $\leqq 1$ in absolute value. This class of pure strategies is denoted F_1.

The motivation for the above class of pure strategies is that the motion of the players is determined by the velocity controls. However, the results obtained here remain valid if a pure strategy for E is allowed to be continuous and for P is allowed to be Lipschitzian with Lipschitz constant 1.

2. The continuous game on the circle with fuel constraint. This section concerns the continuous game on the circle S with termination time T and given arbitrary initial distributions. It will be shown that the solution breaks up into two parts, termed a global part and a local part. The global part of the optimal strategies is a sequence of cohato (or nearly so) moves. The cohato or coin-half-tour move is the equiprobable selection of proceeding clockwise or counterclockwise around the circle from the starting position to the opposite pole.

2.1. General description of games on the circle with fuel constraint. Two blind players, the pursuer P and the evader E, are allowed to move on a circle S in a continuous manner by choosing an instantaneous velocity (forward time derivative). The player P has a given maximum speed while E is allowed any speed. Capture occurs when the two players occupy the same point. The partie terminates when capture occurs or at a given time T whichever occurs first. The payoff is the probability of capture, which P maximizes and E minimizes. We normalize the problem by taking P's maximum speed as unity and the circumference of S as 2.

We assume that the players are initially distributed on S independently; thus due to the symmetry of the problem, we may replace the two probability functions by a single variate function giving the relative initial positions of the two players. Furthermore, we take a coordinate system in S with P initially at the origin; then E's initial position is specified by a probability function C on S. That is, the probability that E is initially in a given set $A \subseteq S$ is $C(A)$.

Pure strategies for P and E. We adopt the convention that the counterclockwise direction, which we also call right, is positive, and thus a pure strategy e for E is a function whose value $e(t)$ is the displacement of E at time t from E's initial position $E_0 \in S$. Here E_0 is the random variable with probability distribution C. Similarly, a function p whose value $p(t)$ gives P's displacement is a pure strategy

for P. Note that this makes $p(t)$ P's actual position in S at time t. We have $p(0) = e(0) = 0$ since neither player has yet moved at time $t = 0$. The points of S are reckoned modulo the circumference 2. Thus, for example, $e(t) = t$, $t \in [0, T]$ means that E, starting from E_0, goes right at speed 1 until the termination time (or capture occurs) which may be several trips around S if T is sufficiently large.

For $a \leqq b$, by the cyclic (closed) interval $[a, b]$ we mean $[a, b]$ in the usual sense if $a \leqq b < a + 2$ and S if $b \geqq a + 2$. That is $[a, b]$ is not to be enhanced by redundant overlaps. Accordingly $[a + n, b + n] = [a, b]$ for every even intger n. Open and half-open intervals are denoted similarly. Furthermore, we frequently represent a point $X \in S$ by any of its coordinates $x \in \{\text{reals}\}$.

The capture probability. For pure strategies p for P and e for E let

$$w(t) = p(t) - e(t), \tag{2.1}$$

$$a(t) = \min_{\tau \leqq t} w(\tau) \leqq 0, \qquad b(t) = \max_{\tau \leqq t} w(\tau) \geqq 0, \tag{2.2}$$

$$A(t) = [a(t), b(t)]. \tag{2.3}$$

LEMMA. *The probability that capture has occurred at time t is*

$$\text{Cap}\,(t) = C(A(t)). \tag{2.4}$$

Proof. Let $e_0 \in (0, 2)$ be a coordinate of the initial position of E in S. The distance right from P to E at τ is

$$D(\tau) = e(\tau) + e_0 - p(\tau).$$

Capture occurs at the earliest time τ such that either $D(\tau) \leqq 0$ or $D(\tau) \geqq 2$. Thus capture has occurred by time t if

$$\min_{\tau \leqq t} D(\tau) \leqq 0 \quad \text{or} \quad \max_{\tau \leqq t} D(\tau) \geqq 2.$$

The first relation becomes $e_0 \leqq b(t)$, and the second becomes $e_0 \geqq a(t) + 2$. Thus capture has occurred if E is initially in the interval $[a(t), b(t)]$. This probability is given by (2.4). □

For the case in which the initial placement may be described by a density function r, we have

$$C(A(t)) = \int_{A(t)} r(x)\, dx. \tag{2.4a}$$

Example partie. Figure 1 shows three different graphical presentations of an example partie of the game with $T = 1.5$. Part (a) of the figure shows pure strategies p for P and e for E assuming that E's initial position was at e_0 shown. In this example capture occurs at the time indicated by the arrow. Part (b) of the figure shows p, $-e$, the resultant $w = p - e$, and the set A in S. In this case since e_0, the indicated initial position of E, is contained in $A(\tau)$ we see that capture occurs by time T. When P and E both use the same pure strategies as shown, the probability that capture occurs by time T depends on the distribution function of e_0 and is given by (2.4). The portion of the curve $x(t) = w(t)$ above the line $x = 1$ is

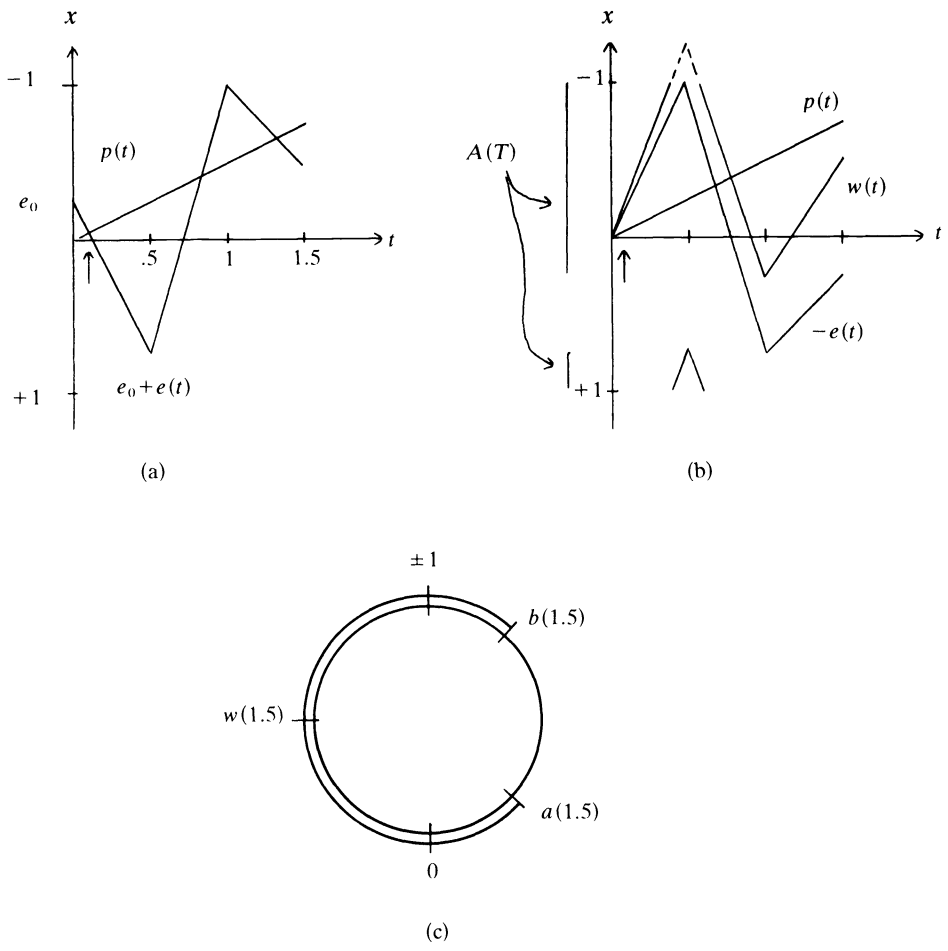


FIG. 1. *Three different graphical presentations of an example partie*

shown in a dotted line and its image in the region $-1 \leqq x \leqq 1$ is shown as a solid line. Part (c) of the figure shows the same situation as (a) and (b) but on the circle without time as an explicit coordinate.

The reduced space Π. The problem as stated is now replaced by an equivalent problem in a reduced space Π which, like S, is a circle of circumference 2. A single point W moves in Π, with coordinate $w(t)$ given by (2.1), corresponding to the motion of P and E in S. Here the coordinate system in Π is like the coordinate system in S, and points are reckoned modulo 2. In this formulation of the problem we no longer regard capture as an event which terminates the game; rather the game terminates at time T and the payoff is Cap (T) given by (2.2), (2.3), and (2.4) where $A(t) \subseteq \Pi$ is the interval swept out by the motion of W in Π. The probability of escape, that is the probability that capture has not occurred, is given by

$$\text{Esc}(t) = C(\Pi \sim A(t)) = 1 - C(A(t)). \tag{2.5}$$

The state of the game. The triplet $(t, w(t), A(t))$ constitutes the state of the game. The entire state of the game is of course unknown by the players since each knows only his own contribution.

2.2. The cohato move and games with initial distribution given as a probability density function.

The cohato move. Consider the two pure strategies that are half-tours of the circle (in realistic space) traveling either left or right at speed unity from the initial position to the opposite pole of the circle. The *coin half-tour* (cohato) move is the mixed strategy which is the equiprobable selection of the left or right half-tour. When used, the cohato move causes the escape probability to decrease by a factor of at least $\frac{1}{2}$, and as such is an important element of P's optimum strategy when $T \geqq 1$.

The cohato move is also an important element of E's optimum strategy when the initial distribution can be represented using a density function as in (2.4a). We first state a theorem about the cohato move; then we analyze games in which (2.4a) holds.

The cohato Theorem.

THEOREM 2.1. *Given that the game is in state* $(s, w(s), A(s))$ *then*:

1) *If P makes a cohato move at maximal speed then*

$$\text{Esc}\,(s+1) \leqq \tfrac{1}{2}\,\text{Esc}\,(s) \tag{2.6}$$

and if E's speed $\leqq 1$, *and* (2.4a) *holds, we have the equality*

$$\text{Esc}\,(s+1) = \tfrac{1}{2}\,\text{Esc}\,(s). \tag{2.7}$$

2) *If* (2.4a) *holds and E makes a cohato move at speed unity, then the equality* (2.7) *is satisfied.*

Proof. The proof of this theorem is given in [8]. However this proof may be simplified by shifting the time scale so that $s = 0$ and the state is $(0, w(0), A(0))$, and further simplified by rotating the coordinate system so that $w(0) = 0$, the origin of the new coordinate system for Π; $A(0)$ is still the same set in Π. □

The value of the game. Let $V(T)$ denote the value of the game (the probability of capture under optimum play) with termination time T. Write $T = n + \eta$ where n is a nonnegative integer and $0 \leqq \eta < 1$. If $V(\eta)$ and optimum strategies exist, then the game with termination time T may be reduced to a game with termination time η.

THEOREM 2.2. *If* (2.4a) *holds, then an optimum strategy for a game with termination time* η *becomes an optimum strategy for the game with termination time* $T = n + \eta$ *if it is followed by n cohato moves. The value of the game (probability of capture)*

$$V(T) = 1 - (\tfrac{1}{2})^n + (\tfrac{1}{2})^n V(\eta). \tag{2.8}$$

Let $\bar{V}(t) = 1 - V(t)$, *the probability of escape under optimum play by both players*; *then* (2.8) *is equivalent to*

$$\bar{V}(T) = (\tfrac{1}{2})^n \bar{V}(\eta) \tag{2.9}$$

which is in a form frequently more convenient.

Proof. We prove the theorem by showing (2.9).

1) Let P play the strategy of the theorem so that at time η, $\text{Esc}(\eta) \leqq \bar{V}(\eta)$. Then from Theorem 2.1

$$\text{Esc}(T) \leqq (\tfrac{1}{2})^n \text{Esc}(\eta) \leqq (\tfrac{1}{2})^n \bar{V}(\eta)$$

or $\bar{V}(T) \leqq (\tfrac{1}{2})^n \bar{V}(\eta)$.

2) On the other hand, if E plays the strategy of the theorem, then $\text{Esc}(\eta) \geqq \bar{V}(\eta)$ and from Theorem 2.1

$$\text{Esc}(T) = (\tfrac{1}{2})^n \text{Esc}(\eta) \geqq (\tfrac{1}{2})^n \bar{V}(\eta)$$

or $\bar{V}(T) \geqq (\tfrac{1}{2})^n \bar{V}(\eta)$. □

We have then that the value of the game with termination time T and arbitrary distribution r, is between $1-(\tfrac{1}{2})^n$ and $1-(\tfrac{1}{2})^{n+1}$ which gives a good approximation for large n.

The local problem. The original problem on the circle with distribution r and termination time $T = n + \eta$ has now been reduced to solving the game for termination time η. That is, we have reduced the global game where the searcher has time to search at least half the realistic space to the local game where the searcher has less than enough time to search half the space. The solution to the local game depends on the given function r.

A particular distribution. A particular distribution r for which the game can be solved exactly is

$$r(x) = \begin{cases} r_1 & \text{for } x \in [0, 2\eta], \\ r_2 & \text{for } x \in [-2\eta, 0]. \end{cases}$$

That is, $r(x)$ is constant for a distance 2η on each side of the origin.

THEOREM 2.3. *For the given distribution, an optimum strategy for P is a mix with probabilities $r_2/(r_1+r_2)$ and $r_1/(r_1+r_2)$ of the two pure strategies which have velocities $+1$ and -1 respectively. An optimum strategy for E is also a mix of the two, speed 1, constant velocity, pure strategies; but, going left with probability $r_2/(r_1+r_2)$ and right with probability $r_1/(r_1+r_2)$. The value of the game*

$$V(\eta) = 2r_1 r_2 \eta/(r_1+r_2). \tag{2.10}$$

Proof. Suppose P uses the strategy ϕ of the theorem, and E uses any pure strategy e; then the probability of capture is

$$\text{Cap}(\phi, e) = (r_2/(r_1+r_2))C(I^1(e)) + (r_1/(r_1+r_2))C(I^2(e)) \tag{2.11}$$

where $I^1(e)$ is the interval swept out if P goes right and $I^2(e)$ is the interval swept out if P goes left.

Let

$$\hat{I}^1 = I^1 \cap [0, 2\eta],$$

$$\hat{I}^2 = I^2 \cap [-2\eta, 0].$$

Then

$$C(I^1) \geqq r_1 L(\hat{I}^1), \tag{2.12}$$

$$C(I^2) \geqq r_2 L(\hat{I}^2)$$

where $L(\cdot)$ is the function on intervals whose value is the length of the interval. From (2.9) and the following relations, the probability of capture

$$\text{Cap}(T) \geqq r_1 r_2 [L(\hat{I}^1(e)) + L(\hat{I}^2(e))]/(r_1 + r_2) = r_1 r_2 L(\hat{I}^1 \cup \hat{I}^2)/(r_1 + r_2)$$
$$\geqq 2r_1 r_2 \eta/(r_1 + r_2).$$

Thus $V(\eta) \geqq 2r_1 r_2 \eta/(r_1 + r_2)$.

Now suppose that E uses the strategy ψ of the theorem and P uses any pure strategy p; then since P's speed $\leqq 1$, the two possible intervals swept out by w (corresponding to the right and left motion of E) intersect only at the origin and we have

$$\text{Cap}(p, \psi) = 2r_1 r_2 \eta/(r_1 + r_2).$$

Thus (2.10) holds because both players may achieve the value of the game. □

The equiprobable distribution.

THEOREM 2.4. *For the equiprobable distribution $r(x) = \frac{1}{2}$, for all x, the value of the game $V(\eta) = \frac{1}{2}\eta$. Thus for termination time T*

$$V(T) = 1 - (\tfrac{1}{2})^n + (\tfrac{1}{2})^{n+1}\eta.$$

An optimum strategy for P is the partial cohato move going right or left equiprobably until time η followed by n cohato moves. This strategy is also optimum for E, but E may substitute for the partial cohato move any strategy which is an equal mix of the pure strategy $e(t) = \alpha t$ or $-\alpha t$ for $t \in [0, \eta]$ where $\alpha \in [0, 1]$. Note that this includes the pure strategy $e(t) = 0$.

Furthermore, we state without proof that the players may interchange the order in which the $n+1$ segments of the move are performed. For example, an optimum move for E is a cohato move followed by remaining stationary until time $1+\eta$ followed by $n-1$ cohato moves.

Proof. The value of the game, and the fact that an optimum strategy for both players is the partial cohato move followed by n cohato moves is obtained from Theorem 2.3. That an equal mix of pure strategies of the form $e(t) = \alpha t$ or $-\alpha t$ where $\alpha \in [0, 1]$ is also optimum for E is shown in the standard manner. □

2.3. The game with exact initial information. We now consider a game in which (2.4a) does not hold, that is, the initial distribution has a "spike". Specifically, we consider the game with exact initial information. In this game the cohato move is still an element of an optimal strategy for P, but the corresponding move for E is a modified cohato move. A modified (or ε-modified) cohato move is like the cohato move in that the player proceeds at speed 1 toward his polar point but stops a distance ε from his polar point and then remains stationary for time ε. The cohato move will not be an element of E's optimal strategy for the game because using the cohato move, E arrives at his polar point at time $t = 1$, and P could guarantee capture by being there then. It is this aspect which makes games with a spike in the initial distribution more complicated than games in which (2.4a) holds.

In this game the initial positions of both players in S are known by each, and it is clear that E can avoid capture until time $t < 1$ by staying closer to P's polar point than P can achieve. This maneuver constitutes the local part of an optimum strategy for E. For games with termination time $T \geqq 1$ we need a theorem

comparable to Theorem 2.1. Combining this result with the solution for the game with termination time $\eta < 1$ gives the solution for games with termination time $T = n + \eta$.

Let

$$e_\varepsilon(t) = \begin{cases} t & \text{for } t \leqq 1-\varepsilon, \\ 1-\varepsilon & \text{for } 1-\varepsilon \leqq t \leqq 1. \end{cases}$$

Then E's ε-modified cohato move is the equiprobable selection of one of the pure strategies $e_\varepsilon(\cdot)$ or $-e_\varepsilon(\cdot)$.

THEOREM 2.5. *Suppose the game is in a state such that E is at least distance* $\delta_0 \in (0, 1]$ *from P. If E makes an* ε*-modified cohato move with* $\varepsilon = \frac{1}{2}\delta_0$, *then at time* 1 *the probability is at least* $\frac{1}{2}$ *that E has escaped with* $\delta_1 \geqq \frac{1}{2}\delta_0$, *where* δ_1 *is the distance between P and E at time* 1.

Proof. Figure 2 shows an ε-modified cohato move for E in which for convenience E is initially taken to be at the origin of the coordinate system. Both the L (left) and the R (right) paths for E are shown, and the diagram allows us to analyze the outcome for both possible paths simultaneously. Also shown in the figure is a possible starting position for P ($p_0 = \delta_0$) and P's two extreme paths (dotted lines) starting from there. From the figure it can be seen that if P starts at least distance $\varepsilon = \frac{1}{2}\delta_0$ away from E at time $t = 0$, it is impossible for his trajectory to intersect both the R and L paths for E since due to P's speed constraint the slope of P's trajectory must be $\leqq 1$ in magnitude. Furthermore, if P starts at least distance δ_0 away from E at time $t = 0$ it is impossible for P to intersect one of the two paths (R or L) and end up closer than distance $\frac{1}{2}\delta_0$ from one of the paths. □

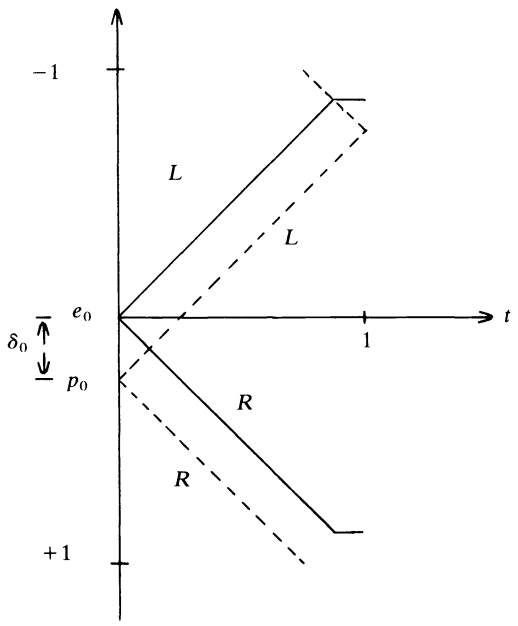


FIG. 2. *The* ε*-modified cohato move*

Combining results for termination time $\eta < 1$ and Theorem 2.5 we obtain

THEOREM 2.6. *The value of the game with exact initial information is*

$$V(T) = 1 - (\tfrac{1}{2})^n.$$

An optimum strategy for P is any strategy containing n cohato moves. An optimum strategy for E is to avoid capture until time η, then make n modified cohato moves with $\varepsilon_1 = \frac{1}{2}\delta_0$ where δ_0 is the closest distance that P can obtain to E at time η, and

$$\varepsilon_{i+1} = \tfrac{1}{2}\delta_i = \tfrac{1}{2}\varepsilon_i, \qquad i = 1, \cdots, n.$$

Proof. Certainly $V(T) \geqq 1 - (\frac{1}{2})^n$ since from Theorem 2.1 the n cohato moves guarantee at least the payoff $1 - (\frac{1}{2})^n$.

To show $V(T) \leqq 1 - (\frac{1}{2})^n$ we observe that if E uses the strategy of the theorem, then $\text{Esc}(\eta) = 1$ and repeated application of Theorem 2.5 yields $\text{Esc}(T) \geqq (\frac{1}{2})^n$. □

3. The game in the plane without boundaries.

The game. In this section we are concerned with the following princess and monster game. Two blind players P and E may move in the playing space S, which is a region of the plane; P moves with a maximum speed normalized to unity. The game terminates at a given time T. Each player is initially distributed equiprobably relative to the other in a region of the playing space far enough from any boundaries so that the boundaries play no role. Capture means that the distance $PE \leqq l$, a quantity small in comparison to the dimension of S; in other words, there is a capture disk of radius l about P. The payoff is the probability of capture with P maximizing and E minimizing.

We note that the requirement that each player is distributed equiprobably relative to the other is not equivalent to distributing the players independently and equiprobably in a region $B \subseteq S$ since, if a player found himself placed near or on the boundary of B, he would know that the other is not outside B and thus not equiprobably around him. Initially then, E_0 is distributed with constant density τ in the annulus with center at P_0, inside radius l, and outside radius $2T + l$. Thus if B is a set contained in the annulus, then

$$\text{Prob}(E_0 \in B) = \tau \text{ Area}(B). \tag{3.1}$$

By far from the boundary, we mean far relative to the distance that the slowest player P can travel in time T. Thus the boundaries play no role in determing the value and optimal strategies. Such might be the case for games involving ships and/or aircraft at sea.

In one sense then the game above is a local game (T small), whereas for games on the circle (§ 2) we were able to obtain solutions for global games (T large relative to the dimensions of the playing space).

Coordinates. Since only the relative positions of the players matter in this game with symmetric distributions and information, we may pick a coordinate system with P at the origin. A pure strategy for P is a vector function $p(\cdot)$ that gives P's position $p(t)$ at time t. A pure strategy for E is similarly a vector function $e(\cdot)$, but $e(t)$ is E's position at time t relative to E's starting position E_0. Thus $p(0) = e(0) = 0$ as in § 2.

The fatal set. Suppose that P selects a pure strategy p, and E selects a pure strategy e; consider the set of points in S such that if E starts in this set, the combined motion of the players causes capture to occur. This set of possible initial positions for E is a function of p and e and is called the fatal set. The fatal set is vacuous only if $p = e$.

In a partie of the game (given p and e), the probability that capture occurs is just the probability that E starts in the fatal set. If the fatal set lies entirely within the region of uniform density r (which will certainly be the case if E does not excxeed speed 1) then the probability of capture is just the area of the fatal set times r.

Figure 3(a) shows an example of the fatal set in S. In this example P uses the speed 1 constant velocity strategy going right in the figure; E uses the speed 1 constant velocity strategy going downward in the figure. The fatal set is the tubular area. If E's initial position e_0 is anywhere in this fatal set, capture will occur to the right of P's initial position and below E's initial position. The figure shows P's velocity vector $\dot{p}$, and E's velocity vector $\dot{e}$ drawn from two representative starting positions.

The reduced space Π. It is convenient to work in a reduced space Π similar to the one used for the game on the circle. Corresponding to the motion of P and E in the realistic space, a single point W moves in Π with position vector

$$w(t) = p(t) - e(t) \tag{3.2}$$

and $w:[0, T] \to \Pi$ is the path of W in Π.

The curve corresponding to the path w is denoted by $I(w) \triangleq I(w, T)$ where $I(w, t) = \{w(\tau) | \tau \in [0, t]\}$. It is clear that there are many different paths w corresponding to the same set of points or curve $I(w)$.

In the reduced space we regard E as remaining stationary and capture occurs if the disk around the moving point W overlaps E_0. The fatal set in S corresponds to the set in Π swept out by the moving disk around W. This set of points in Π is

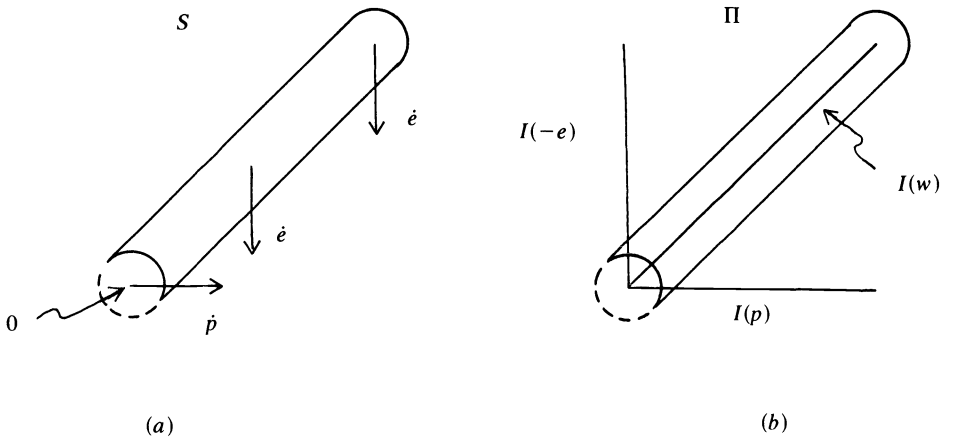


FIG. 3. *An example partie showing the fatal set in S, and the corresponding set (the swath of w) in* Π

called the swath of the path of W. Figure 3(b) shows the swath in Π swept out by the disk around W for the same example partie as was shown in Figure 3(a) in S. Figure 3(b) shows the curves $I(p)$, $I(-e)$, and $I(w)$ as well as the swath of W.

We denote the swath area of curve I by $A(I)$ where the area of the initial disk around W_0 is not included in $A(I)$. The area of this initial disk is excluded because it is assumed that E is not in this disk (i.e., capture has not occurred before motion of the players). Furthermore, this convention prevents the area of this disk from being counted twice when a curve is regarded as being composed of two sections. Thus if the curve I is composed of I_1 followed by I_2 we have

$$A(I) \leqq A(I_1) + A(I_2)$$

and the equality holds provided that the radius of curvature of I is $\geqq l$ in a sufficiently large neighborhood of the point where I_2 joins I_1 and that the swath of I_1 does not intersect the swath of I_2 at other portions of the curve.

As long as the swath of W in Π remains within the region of constant density r, the problem of determining the capture probability reduces to determining the area of this swath; P desires to maximize and E to minimize this area. We denote the swath area by $A(w, t)$ and when $t = T$ we write cap $(T) = rA(w, T) = rA(w)$. The use of the symbol A as a function of curves and as a function of paths should cause no confusion.

If w is smooth and the radius of curvature of w is $\geqq l$ thoughout, and the swath of w does not cut back upon itself, then $A(w) = 2lL(w)$ where $L(w)$ is the length of the path w. However, for more general paths w, the swath area is not so simply related to the path length, and we have in general

$$A(w) \leqq 2lL(w). \tag{3.3}$$

We now state a useful theorem about the minimum swath area between two points.

THEOREM 3.1. *Among all the curves I between two fixed points, the one with minimum swath area is the curve I^* consisting of the straight line segment between the two fixed points. All other curves have swaths of larger area.*

Proof. This theorem seems intuitively obvious, but due to possible overlaps in the swath of w, the proof cannot rely upon the length $L(w)$.

Figure 4 shows the straight line curve I^* and another curve I between the two fixed points on the X-axis. We define $h^*(\xi)$ and $h(\xi)$ as the distances across the

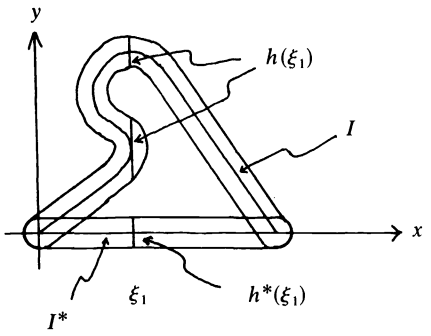


FIG. 4. *The curves I and I^* and their swaths in* Π

swaths of I^* and I. In case the swath of I "doubles back" as in the figure at ξ_1, the distance $h(\xi_1)$ is the sum of more than one segment length. We have $h(\xi) \geqq h^*(\xi) \geqq 0$ for all ξ. Now since

$$\int_{-\infty}^{\infty} h(x)\,dx \geqq \int_{-\infty}^{\infty} h^*(x)\,dx$$

we have

$$A(I) \geqq A(I^*). \tag{3.4}$$

To show the strict inequality when $I \neq I^*$ we observe that at a point on I where the slope of the curve is nonzero we have that $h(\xi) > h^*(\xi)$ in a neighborhood. Thus

$$A(I) > A(I^*) \tag{3.4a}$$

for $I \neq I^*$. □

We observe that a translation of a path w by a constant vector c to path $w+c$ preserves the swath area of the path. That is

$$A(w+c, t) = A(w, t) \tag{3.5}$$

where we still retain the convention of omitting the initil area of the capture disk.

The solution. It is clear at the outset that the value of the game (if it exists) is no greater than $2lrT$ since E can prevent the probability of capture from being greater than this by remaining stationary. Thus

$$V \leqq 2lrT. \tag{3.6}$$

It remains to show that P can guarantee achieving this probability of capture.

THEOREM 3.2. *The value of the game is $2lrT$. An optimum strategy for P is any mixed strategy which is the equiprobable selection of the pure strategies p and $-p$ where p is any full-speed, constant velocity strategy. That is, p is such that*

$$p(t) = p(T)t/T \quad \text{and} \quad |p(T)| = T. \tag{3.7}$$

The optimum strategy for E is to remain stationary.

Proof. We need only show

$$V \geqq 2lrT \tag{3.8}$$

since we have shown the other inequality (3.6). Assume that P uses the strategy ϕ of the theorem, and let e be any pure strategy for E. Then

$$\text{pay}(\phi, e) = \tfrac{1}{2}rA(p-e) + \tfrac{1}{2}rA(-p-e)$$

if the paths $p-e$ and $-p-e$ remain in the region of constant density r. For the case in which E moves so far and fast as to cause either $p(t)-e(t)$ or $-p(t)-e(t)$ to exit the region of constant density then at least one of $A(p-e)$ or $A(-p-e)$ is $\geqq 4lT$ and thus (3.8) holds for this case.

For the case in which both paths $p-e$ and $-p-e$ remain in the region of constant density we have $A(-p-e) = A(p+e)$ since the path $-(p+e)$ is a reflection of $p+e$ through the origin. From (3.4), $A(p-e) =$

$A(p(T)+e(T)+p-e)$ since $p(T)+e(T)$ is a constant vector. We have

$$\text{pay}\,(\phi, e)=\tfrac{1}{2}rA(p+e)+\tfrac{1}{2}rA(p(T)+e(T)+p-e).$$

The two paths $p+e$ and $p(T)+e(T)+p-e$ taken successively form two legs of a "triangular" path between the origin and the point $2p(T)$. See Figure 5.

From Theorem 3.1, the path $2p$ is a path of minimum swath area between the origin and the point $2p(T)$. Thus

$$A(p+e)+A(p(T)+e(T)+p-e)\geqq A(2p)=4lT.$$

And pay $(\phi, e)\geqq\frac{1}{2}r(4lT)=2rlT$ so that (3.8) is satisfied. □

We notice that if E's motion is parallel to P's line of travel and E's speed $\leqq 1$, then pay$(\phi, e)=2rlT$. On the other hand, if E has a component of motion perpendicular to P's line of travel, the strict inequality holds pay $(\phi, e)>2rlT$. Thus the strategy ϕ of Theorem 3.2 does not necessarily penalize E for moving. However, the strategy for P that is the equiprobable selection of all p that satisfy (3.7) is optimum since it is a mixture of optimum strategies, and at the same time any motion by E whatsoever will increase the payoff.

Comparison to game on circle. In the game on the circle analyzed in § 2 we found that with the equiprobable initial distribution, E's optimum strategy for termination time $\eta<1$, was not unique. One of E's optimum strategies was to remain stationary until time η. For the game in the plane that we have just analyzed, however, E does not have a choice of optimum strategies. E's only optimum strategy is to remain stationary.

Extensions of results. It is clear that the results of this section also apply in N-space. However, the difficulties due to the boundary of the playing space S still remain. It is also clear that the results apply where S is a region of the sphere, but global results have not been obtained.

It is conjectured that for the game on the sphere, a result similar to (2.6) holds. That is, with optimum play, the probability of escape decreases by a factor of $g<1$ in the time required to search out g of the space.

Memory and information. For the class of pure strategies allowed, a player needs to have the capability of "remembering" a trajectory of length T in order to be able to employ any of the pure strategies. He need not have an alternate plan about what to do if a certain region has been explored and capture has not occurred; he simply continues searching along the prescribed path. However, in

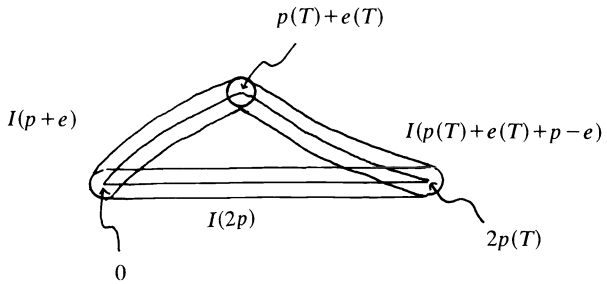


FIG. 5. *The "triangle" through* $0, p(T)+e(T)$, *and* $2p(T)$

situations in which the pure strategies that make up the optimal strategy mix are a small subset of the set of allowed pure strategies the players may be able to perform an optimal strategy with considerably less memory. For example in the game on the circle with large T the players need only a memory large enough to hold a short initial maneuver plus a sequence of decisions about which way to go at "nodal" times (times at which the cohato moves are initiated). Furthermore, the decisions as to which way to go at nodal times can be generated and stored before the game starts or be generated by a random device at the nodal times, eliminating the need for that storage.

In games with "blind" players, no more information about the state of the game is ever obtained than was initially given. In fact the players know less, generally speaking, as the game developes since each player is unaware of the maneuvers of the other. The current state of the game is a myriad of possible initial positions together with possible trajectories of the opponent.

Acknowledgment. The results presented here are from the author's Ph.D. dissertation, Johns Hopkins University, Baltimore, MD, 1974. The author gratefully acknowledges the guidance of his advisor, Prof. Rufus Isaacs.

REFERENCES

[1] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[2] S. ALPERN, *The search game with mobile hider on the circle*, Differential Games and Control Theory, E. O. Roxin and R. Sternberg, eds., Marcel Dekker, New York and Basel, 1974, pp. 181–200.

[3] M. I. ZELIKIN, *On a differential game with incomplete information*, Soviet Math. Dokl., 13 (1972), no. 1, pp. 228–231.

[4] D. J. WILSON, *Isaac's princess and monster game on the circle*, J. Optimization Theory Appl., 9 (1972), pp. 265–288.

[5] R. H. WORSHAM, *A discrete game with a mobile hider*, Differential Games and Control Theory, E. O. Roxin and R. Sternberg, eds. Marcel Dekker, New York and Basel, 1974, pp. 201–230.

[6] P. ENSLOW, *A bibliography of search theory and reconnaissance theory literature*, Naval Res. Logist. Quart., June (1966), pp. 177–202.

[7] B. HALPERN, *The robot and the rabbit—a pursuit problem*, Amer. Math. Monthly 76 (Feb. 1969), pp. 140–144.

[8] J. G. FOREMAN, *The princess and the monster on the cirle*, Differential Games and Control Theory, E. O. Roxin and R. Sternberg, eds. Marcel Dekker, New York and Basel, 1974, pp. 231–240.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

COMMENTS ON "GENERALIZED PREDICTION-CORRECTION ESTIMATION"*

L. B. WEINER†

Abstract. A recently discovered error in the original paper (this Journal, 1969) leads to incorrect results and numerical instability in the application discussed in Example 2. A correct approximation yielding a stable algorithm is suggested.

The paper by Krasnakevitch and Haddad [1] correctly and properly derives a technique to reject second order bias resulting from applying local linearization to nonlinear filter problems. However, inconsistent accounting of higher order terms in the Taylor series expansions in Example 2 (also in [2] and [3, § 3]) yields an incorrect equation, which is likely to produce numerical instability in repeated iteration. As discussed in conversation with the authors, the approximation given in [3, § 6], matching the approximation represented by (8) below, is valid and will reduce the likelihood of numerical instability. This error was discovered recently during the study of the application of the methodology presented to the problem of tracking maneuvering vehicles reentering the Earth's atmosphere.

The error occurs in combining the equation (107) where fourth order terms in $e(k-1/k-1)$ are neglected in the formulation of $P(k)$, with (95) where the fourth order term $E\{e(k/k-1)\}E\{e'(k/k-1)\}$ is subtracted. To demonstrate the effect of this omission, consider the scalar case where (102) becomes:

$$e(k/k-1)=A(k)e(k-1/k-1)+B(k)[e(k-1/k-1)]^2 + C(k)[e(k-1/k-1)]^3+J(k) \tag{1}$$

where $J(k)$ contains terms of order 4 and higher. Then, to fourth order (with $e(k-1/k-1)$ considered to be zero mean Gaussian with variance σ^2_{k-1}), equation (106) still holds true as:

$$E\{e(k/k-1)\}=B(k)\sigma^2_{k-1} \tag{2}$$

to second order.

$P(k)$, computed correct to fourth order, proceeds as follows, with $P(k)$ still defined by (84):

$$P(k)=E\{[e(k/k-1)]^2\} \tag{3}$$

or

$$P(k)=E\{A^2(k)[e(k-1/k-1)]^2+2A(k)B(k)[e(k-1/k-1)]^3 + B^2(k)[e(k-1/k-1)]^4+2A(k)C(k)[e(k-1/k-1)]^4+G(k)\} \tag{4}$$

where $G(k)$ contains terms at order 5 and above.

Then, equivalent to (107), with the term for $u(k)$ suppressed for simplicity:

$$P(k)=A^2(k)\sigma^2_{k-1}+3[B^2(k)+2A(k)C(k)]\sigma^4_{k-1} \tag{5}$$

* Received by the editors April 23, 1976, and in revised form October 13, 1976.
† Teledyne Brown Engineering, Huntsville, Alabama 35807.

and combining with (95) yields, correct to fourth order:

$$\text{Cov}\{e(k/k-1)\} = A^2(k)\sigma_{k-1}^2 + 2B^2(k)\sigma_{k-1}^4 + 6A(k)C(k)\sigma_{k-1}^4. \tag{6}$$

This compares to the authors' equations (95) and the incorrect equation (107) combined to give:

$$\text{Cov}\{e(k/k-1)\} = A^2(k)\sigma_{k-1}^2 - B^2(k)\sigma_{k-1}^4. \tag{7}$$

While the difference in the original incorrect form (7) and the correct form (6) is of second order, the iteration of the incorrect equations thousands of times as in the suggested maneuvering reentry vehicle tracking application can, and indeed does, render the Cov $\{e(k/k-1)\}$ nonpositive definite through repeated subtraction of the small quantity $B^2(k)\sigma_{k-1}^4$, yielding incorrect results and numerical instability in the filter equations.

Approximating of (6) by its second order approximation:

$$\text{Cov}\{e(k/k-1)\} \simeq A^2(k)\sigma_{k-1}^2 \tag{8}$$

as used in linearized Kalman filtering, or justifying deleting the term $6A(k)C(k)\sigma_{k-1}^4$ to eliminate computation of the matrices of third partials, merits further study for specific applications. The approximation represented by the original (107) combined with (95) is, however, invalid for the reasons demonstrated by the example.

REFERENCES

[1] J. R. KRASNAKEVITCH AND R. A. HADDAD, *Generalized prediction-correction estimation*, this Journal, 7 (1969), pp. 496–511.

[2] ———, *The Kindler formulation for unbiased BRV/MaRV tracking*, Rep. 7612-1, Kindler Assoc., Cambridge, MA, revised Feb. 1976.

[3] ———, *Filter equations—The Kindler formulation for unbiased BRV/MaRV tracking*, Rep. 7612-2, Kindler Assoc., Cambridge, MA, Feb. 1976.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

SURVEY OF MEASURABLE SELECTION THEOREMS*

DANIEL H. WAGNER†

Abstract. Suppose $(T, \mathcal{M})$ is a measurable space, X is a topological space, and $\emptyset \neq F(t) \subset X$ for $t \in T$. Denote $\operatorname{Gr} F = \{(t, x): x \in F(t)\}$. The problem surveyed (reviewing work of others) is that of existence of $f: T \to X$ such that $f(t) \in F(t)$ for $t \in T$ and $f^{-1}(U) \in \mathcal{M}$ for open $U \subset X$. The principal conditions that yield such f are (i) X is Polish, each $F(t)$ is closed, and $\{t: F(t) \cap U \neq \emptyset\} \in \mathcal{M}$ whenever $U \subset X$ is open (Kuratowski and Ryll-Nardzewski and, under stronger assumption, Castaing), or (ii) T is a Hausdorff space, $\operatorname{Gr} F$ is a continuous image of a Polish space, and $\mathcal{M}$ is the σ-algebra of sets measurable with respect to an outer measure, among which are the open sets of T (primarily von Neumann). The latter result follows from the former by lifting F in a natural way to a map into the closed sets of a Polish space. This procedure leads to the theory of set-valued functions of Suslin type (Leese), which extends the result (i) to comprehend a considerable portion of the results on the problem surveyed. Among the topics addressed, measurable implicit functions and the case where X is a linear space and each $F(t)$ is convex and compact are particularly important to control theory, for example. With $T = X = [0, 1]$ and $\operatorname{Gr} F$ Borel, an elegant partition of $\operatorname{Gr} F$ into Lebesgue measurable maps from T to X, parameterized by Borel functions, has been found (Wesley) via Cohen forcing methods. Other topics discussed include pointwise optimal selections, selections of partitions, uniformization, non-σ-algebras in place of $\mathcal{M}$, Lusin measurability, and set-valued measures. Substantial historical comments and an extensive bibliography are included. (See addenda (i)–(iii).)

1. Introduction. This paper surveys the subject of existence of a measurable function which is a selection of a given set-valued function mapping a measurable space into subsets of a topological space. The subject has undergone considerable development in the past decade. We attempt to review the principal results currently available and to give a history of prior work, dating primarily from[0] a 1949 lemma of von Neumann [NE] and from precursors on the subject by Lusin [LS], Novikov [NO1], and others of the 1930 era. (See addenda (i), (ii).)

Measurable selection problems arise in a variety of ways in control theory, mathematical economics, probability theory, statistics, and operator theory, among other fields. For example, Aumann's influential 1965 paper [AU1] was motivated by economics. Numerous applications are given in the referenced papers. Although we will have little to say about applications, let us note two examples.

Suppose $d: R^2 \to R^n$ and $D(q) = \int_R d(t, q(t))\, dt$ for all $q: R \to R$ for which the (Lebesgue) integral is finite. Suppose it is known that $\lambda \in R^n$ and q^* have the property that

$$\lambda \cdot D(q^*) \geqq \lambda \cdot D(q) \quad \text{for all admissible } q, \tag{1.1}$$

the dot being ordinary inner product. Do we then have

$$\lambda \cdot d(t, q^*(t)) \geqq \lambda \cdot d(t, y) \quad \text{for } y \in R, \quad \text{a.e. } t \in R? \tag{1.2}$$

In other words, does satisfaction of a functional multiplier rule imply satisfaction of a pointwise multiplier rule? The answer was shown by Aumann and Perles

* Received by the editors June 21, 1976, and in revised form December 14, 1976.

† Daniel H. Wagner, Associates, Paoli, Pennsylvania 19301. This work was supported in part by the Office of Naval Research under Contract N 00014-70-C-0232.

[0] Additional credit to Jankov [JN], Novikov [NO2], and Rokhlin [RK2], and other recent information are given in addenda in proof at the end of the paper.

[AP], and in more generality by Wagner and Stone [WS], to be affirmative providing d is a Borel function. An example in [WS] also shows that it does not suffice for d to be Lebesgue measurable. In this application, one defines the set-valued function F by

$$F(t) = \{x : x \in R^n \text{ and } \lambda \cdot d(t, x) > \lambda \cdot d(t, q^*(t))\} \quad \text{for } t \in R.$$

If (1.2) fails, one finds a Borel $T \subset R$ of positive measure such that $F(t) \neq \varnothing$ for $t \in T$. Since d is a Borel function, von Neumann's theorem mentioned above (e.g., Corollary 5.2 below) assures the existence of a Lebesgue measurable function f on T such that $f(t) \in F(t)$ for $t \in T$, i.e., f is a measurable selection of $F|T$, from which a contradiction to (1.1) is easily deduced.

As the second example, consider the problem of generalizing the LaSalle bang–bang principle of control theory: Any output attainable via an admissible control function is attainable by a control which utilizes only extreme points of each instantaneous (compact convex) set of possible choices. Proving such statements usually involves recognizing that each such choice is a convex combination of extreme points, and one needs to find such a representation in a measurable way (see § 8 below, [WG1], [AU1], [CA5], [HV5], or [VA3], for example).

The preliminaries in § 2 include some instructive counterexamples due to Dauer and Van Vleck and to Kaniewski. Early history is discussed in § 3, primarily work of Lusin, Novikov, and Saks.

The main fundamentals of measurable selections are given in § 4 on closed-valued functions, § 5 on set-valued functions with measurable graph, § 6 on set-valued functions of Suslin type, and § 7 on implicit functions. Of foremost importance is pioneering work of von Neumann, Kuratowski and Ryll-Nardzewski, and Castaing. Prominent in these sections is the prolific work of Castaing, Himmelberg and Van Vleck, and Leese. Section 6, based on work of Leese, unifies much of the developments in § 4, § 5, and § 7. Numerous additional authors have contributed to these developments. In particular, the papers on graph-conditioned theorems by Aumann and Sainte-Beuve are quite interesting and some expositions by Rockafellar and Himmelberg are especially helpful. The initial contribution to the implicit function topic of § 7 was Filippov's.

Convex-valued functions are discussed in § 8, notably Valadier's work on scalarly measurable selections and results of Himmelberg and Van Vleck and of Leese on extreme point selections. We note applications of selection theory for convex-valued functions to bang–bang problems and, primarily by Rockafellar, to optimization of convex integral functionals by duality methods. In § 9 we review results on pointwise optimal measurable selections, initiated by Dubins and Savage.

In § 10 we discuss decomposition of the graph of a set-valued function into measurable selections, notably an elegant result of Wesley which appears to be the most profound result to date in measurable selection theory, judging by its proof via Cohen forcing methods.

Regarding a partition of a set as a set-valued function, in § 11 we have an alternative approach to selection problems, used in early results by Mackey and Dixmier and later more extensively by Hoffman-Jørgensen, Kuratowski, Maitra, and Rao, among others. The subject of uniformization, discussed in § 12, usually

treats selections with emphasis on their properties as subsets of a product space; this subject is older than that of emphasizing *function* properties of selections, and our coverage here is less complete than that of most topics discussed. It includes theorems relating these two types of properties. Replacing the σ-algebra of the measurable space with other structures is discussed in § 13. Lusin measurability is reviewed in § 14, including generalizations of Lusin's theorem involving set-valued functions. In § 15, we discuss work on set-valued measures, led by Godet-Thobie and Artstein. A few works which do not come directly under our other topic headings are noted in § 16; the final work discussed is the very recent "measurable fields" approach by Delode, Arino, and Penot, which appears to be quite promising.

A sequence of recommended initial reading is given in § 17—some readers may wish to turn to this first.

Numerous results come under more than one of these topic headings. We have tried to give or discuss each in the section where its greatest interest appears to lie.

A significant special topic that we do not discuss is that of differential equations involving set-valued functions, in particular orientor fields. Our only discussion of *continuous* selections, an important topic related to measurable selections, is to cite a few general references in § 13.

An extensive bibliography is provided, categorized as described in its introduction.

An acknowledgement to several sources of help is given at the end. Regarding accreditation, let us emphasize the well-known fact that "superseded" results have usually contributed to the development of the subject by their earlier appearance. By including considerable historical comments, we have tried to do some justice to this point, but certainly very inadequately. Indeed even with fairly recent literature, the heavy volume of results on the subject has required that much excellent work be reviewed in only a superficial way, e.g., our discussion of set-valued measures in § 15.

2. Preliminaries. For every set S we define $\mathcal{P}(S) = \{A : \varnothing \neq A \subset S\}$. When $\mathcal{L}$ is a set of sets, by $\mathcal{L}_\sigma$ we mean the set of countable unions of members of $\mathcal{L}$. If S is topologized, by $\mathcal{B}(S)$ we mean the σ-algebra of Borel sets of S, i.e., that generated by the open sets of S, and for $A \subset S$, by cl A we mean the closure of A. If $\mathcal{A}$ and $\mathcal{D}$ are σ-algebras, by $\mathcal{A} \otimes \mathcal{D}$ we mean the smallest σ-algebra containing $\{A \times D : A \in \mathcal{A}$ and $D \in \mathcal{D}\}$. We denote the set of real numbers by R and Euclidean n-space by R^n.

We make considerable use of fixed notations denoting fundamental objects in the structure of the problem. Definitions stated with respect to T, μ, $\mathcal{M}$, X, or F as fixed below apply in obvious ways to counterpart other objects.

We fix $T \neq \varnothing$ as a set, not necessarily topologized, and μ as a nonnegative (possibly infinite) measure over T. *Measurability always refers to μ unless stated otherwise.* Often we specify that μ is an outer measure, meaning that $\mu(S)$ is defined for *each* $S \subset T$ and μ is countably sub-additive; then measurability of $S \subset T$ is defined by Carathéodory's criterion [FE, § 2.12] and the set of measurable sets is a σ-algebra. At other times μ is merely defined on a given σ-algebra,

i.e., the family of measurable sets, and is countably additive. Often it will not matter which of these measure concepts is used. *In all cases we fix* $\mathcal{M}$ as the σ-algebra of measurable sets. If Z is a topological space and $f: T \to Z$, we say f is a *measurable function* if $f^{-1}(U) \in \mathcal{M}$ whenever $U \subset Z$ is open. (Many of the results reviewed here are taken from papers based on the measure foundations of Bourbaki [BO2] who defines an integral as a linear functional and the measure of a set as the integral of its characteristic function.)

It should be recognized that our measure conventions include the case where no measure is present, i.e., when one is dealing with a measur*able* space $(T, \mathcal{M})$, meaning $\mathcal{M}$ is an arbitrary σ-algebra of subsets of T and $T \in \mathcal{M}$; one may let μ be the trivial measure given by $\mu(S) = 0$ for $S \in \mathcal{M}$ to bring this case into our framework. For theorem statements which do not mention properties of μ, in fact, one may just as well consider that μ is not present.

We fix X as a topological space (except in Theorems 5.8 and 12.1), and we reserve F to mean $F: T \to \mathcal{P}(X)$, i.e., F is a *set-valued function* (also called multifunction, multivalued function, in French, multiapplication, or in German, Multiabbildung).

We say F is (adjective)-valued if $F(t)$ is (adjective) for $t \in T$, and we apply operations on sets to operations on set-valued functions in an obvious fashion, e.g., if $G: T \to \mathcal{P}(X)$, then $(F \cap G)(t) = F(t) \cap G(t)$ for $t \in T$.

A *selection* (also called selector, section, uniformization, or, in German, Schnitt) of F is a function $f: T \to X$ such that $f(t) \in F(t)$ for $t \in T$. We denote

$$\mathcal{S}(F) = \{f: f \text{ is a measurable selection of } F\}.$$

We say f is an *a.e. measurable selection* of F if for some $S \in \mathcal{M}$, $\mu(T \backslash S) = 0$ and f is a measurable selection of $F|S$. The problems considered here are: when does one have $\mathcal{S}(F) \neq \varnothing$ (i.e., there exists a measurable selection of F) or when does there exist at least an a.e. measurable selection of F? Of course, from the axiom of choice, every set-valued function has a selection.

Following [RC6], we say $\{f_1, f_2, \cdots\}$ is a *Castaing representation* of F if $f_i \in \mathcal{S}(F)$ for $i = 1, 2, \cdots$, and $\{f_1(t), f_2(t), \cdots\}$ is dense in $F(t)$ for $t \in T$. Under weak conditions (see Theorem 4.2 below), existence of a Castaing representation, which is an additional problem of interest, is equivalent to measurability of F as defined next; this fact lends itself to manipulation of closed-valued functions in ways which help to solve our primary problem of proving $\mathcal{S}(F) \neq \varnothing$, as shown particularly well by Rockafellar [RC2, 6]. (See addenda (ii), (v).)

For $A \subset X$, we define $F^-(A) = T \cap \{t: F(t) \cap A \neq \varnothing\}$. We say F is *measurable*, as a set-valued function, if $F^-(K) \in \mathcal{M}$ whenever $K \subset X$ is closed, and *weakly measurable* if $F^-(U) \in \mathcal{M}$ whenever $U \subset X$ is open. Early uses of variations on these concepts of measurability were made by Rokhlin [RK2], Berge [BG], Pliś [PL1], Debreu [DE], and Kuratowski and Ryll-Nardzewski [KRN]. The definition of a measurable set-valued function was formalized and exploited by Castaing in his thesis [CA4, 5]. The term "weak measurability" (although not the concept) was introduced by Himmelberg, Jacobs, and Van Vleck [HJV].

We define the *graph* of F, denoted Gr F, by

$$\text{Gr } F = (T \times X) \cap \{(t, x): x \in F(t)\}.$$

For a function $f: T \to X$ we do not refer to the graph of f (which we regard as the same as f). It should be clear when we are referring to properties of f as a subset of $T \times X$ (such as being a Borel set) or as a map on T to X (such as being a Borel function). If Y and Z are topological spaces, we say $f: T \times Y \to Z$ is a *Carathéodory map* if $f(t, \cdot)$ is continuous for $t \in T$ and $f(\cdot, x)$ is measurable for $x \in Y$. We denote $\pi_T(t, x) = t$ and $\pi_X(t, x) = x$ for $t \in T, x \in X$.

If T is topologized and F is closed-valued, we say F is *upper{lower} semi-continuous*, abbreviated usc{lsc}, as a set-valued function, if for each closed{open} $A \subset X$, $F^-(A)$ is closed{open} (see [KU1, Chap. 1, § 18]); F is usc implies Gr F is closed. One says F is *continuous* if F is usc and lsc. The abbreviations usc and lsc are also applied to $f: T \to R$.

When F is compact-valued and X is separable metric, measurability and continuity of F as a set-valued function are respectively equivalent to measurability and continuity as a "point-valued" function with respect to the Hausdorff metric on the set of compact subsets of X. This is applied, e.g., by Castaing [CA4, 5, Chap. 4] and in earlier work by Debreu [DE].

An excellent source for measurability properties of set-valued functions is Himmelberg [HM2]; see also Rockafellar [RC6, 2, 3] and Castaing [CA5]. Several references in the bibliography are additional sources; those marked with a single prime are included because, in this respect, they augment the unprimed references (sources on existence of measurable selections), in some cases peripherally. Debreu's [DE] (1965) was a pioneering paper on measurability properties of F, without going into selection questions.

If every closed subset of X is a G_δ (e.g., if X is metrizable), then measurability of F implies weak measurability; the converse fails as shown by Example 2.4 below. We cannot omit the condition on X: Let $T = X = R$, $\mathcal{M} = \{\varnothing, T\}$, the open sets of X be the open right half-lines, and $F(t) = \{x: x \leqq t\}$ for $t \in T$. In most theorems below where F is weakly measurable, we also have X metrizable, so measurability may be substituted for weak measurability in those cases.

If $F_i: T \to \mathcal{P}(X)$ is {weakly} measurable for $i = 1, 2, \cdots$, then so is $\bigcup_{i=1}^{\infty} F_i$. Unfortunately the same cannot be said for intersections—see Example 2.3 below. However, if each F_i is weakly measurable and closed-valued, and *either* (i) X is σ-compact and metric, (ii) X is separable metric and for $t \in T$, for some i, $F_i(t)$ is compact, (iii) X has a countable base and for some i, F_i is measurable and compact-valued, or (iv) $X = R^n$, then $\bigcap_{i=1}^{\infty} F_i$ is measurable [HM2], [LE3], [RC6].

The principal additional properties of closed-valued F are summarized in Theorem 4.2 below.

By a *Polish space* is meant a (not necessarily complete) homeomorph of a complete separable metric space. We say that S is a *Suslin{Lusin} space* if S is topologized as a Hausdorff space and there exist a Polish space P and a continuous surjective {bijective} $\varphi: P \to S$. We define a *weakly* Suslin space in the same way without the Hausdorff requirement. A *{weakly} Suslin set* in a topological space is a subset which is a {weakly} Suslin space. Suslin sets play important roles in measurable selection theory. Probably the most thorough treatment of them is [HJ]. Other excellent references include [FE, § 2], [KU1, Chap. III, § 38], [BO1, Chap. IX, § 6], and [CH]. A subset of a Hausdorff space which is a Lusin space is a Borel subset, and in a Polish space the converse holds [FE, § 2.2.10]. A Borel

subset of a Suslin space is a Suslin set [HV3, Lem.]. If μ is an outer measure, T is Hausdorff and $\mathcal{M}$ contains the open sets of T, then $\mathcal{M}$ contains the Suslin sets of T [FE, § 2.2.12]. Any Suslin space is a continuous image of the set of irrational numbers.

The definition of Suslin set given here is more general than that given in [KU1] (there called analytic set) and [BO1] and less general than that given in [FE]. It is important to note that [BO1] requires Suslin spaces and Lusin spaces to be metrizable by definition, but we are advised that a forthcoming edition of [BO1] will use the definitions employed here. Note that Castaing and his colleagues at l'Université du Languedoc, Montpellier, have consistently considered Suslin spaces to be Hausdorff, not necessarily metrizable, although that has not been explicit in their earlier publications. What we term Suslin and Lusin spaces are respectively called analytic and standard spaces in [HJ], [CH], and [MG]. Not much can be said about properties of weakly Suslin sets ([LE5] calls them "classical analytic")—that definition merely affords a weaker hypothesis which suffices for some theorems in non-Hausdorff spaces. Still weaker hypotheses, related variations under the term "analytic," are used in, e.g., [LE3, 5] and [SN] (see below: Theorem 4.11, remarks before Theorem 5.6, and Theorem 12.3).

One says μ is *complete* if $S' \subset S \in \mathcal{M}$ and $\mu(S) = 0$ imply $S' \in \mathcal{M}$ (always true if μ is an outer measure). We say $S \subset T$ is *universally measurable* (w.r.t. $\mathcal{M}$ and without reference to μ) if S is measurable for each bounded (equivalently, σ-finite) outer (equivalently, complete) measure whose set of measurable sets contains $\mathcal{M}$. If $\mathcal{M}$ contains all of its universally measurable sets, it is said to be *complete*. Of course, if μ is σ-finite and complete, then $\mathcal{M}$ is complete.

We shall frequently employ an assumption that is weaker than $\mathcal{M}$ being complete, viz., that $\mathcal{M}$ is a Suslin family, defined next. This definition employs the Suslin operation, which has been central to the classical development of the theory of Suslin sets. We make little use of the Suslin operation other than to define "Suslin family"; however, it is used in several papers to prove results cited below.

We fix $\mathcal{V}$ and $\mathcal{V}^*$ as the respective sets of infinite and finite sequences of positive integers. Let $\mathcal{F}$ be a family of sets, and $A: \mathcal{V}^* \to \mathcal{F}$. For $\sigma \in \mathcal{V}$, denote $(\sigma_1, \cdots, \sigma_n)$ by $\sigma|n$, following [RG]. Then

$$\bigcup_{\sigma \in \mathcal{V}} \bigcap_{n=1}^{\infty} A_{\sigma|n}$$

is said to be *obtained from $\mathcal{F}$ by the Suslin operation*. If every set obtained from $\mathcal{F}$ in this way is also in $\mathcal{F}$, we say $\mathcal{F}$ is a *Suslin family* ([RB] and [LE2–5], for example, say $\mathcal{F}$ admits the Suslin operation, and [DL] and [DAP2] say $\mathcal{F}$ is "souslinienne"). We always have $\{D: D$ is obtained from $\mathcal{F}$ by the Suslin operation$\}$ is a Suslin family ("generated by $\mathcal{F}$") [HF, § 19]. In a Hausdorff space, each Suslin set is in the Suslin family generated by the set of closed sets [RGW, Theorem 2]; in a Suslin space the converse holds (adapt the proof in [KU1, § 39, II]).

If μ is an outer measure, then $\mathcal{M}$ is a Suslin family, e.g., [SK, p. 50]. Consequently, if $\mathcal{M}$ is complete, $\mathcal{M}$ is a Suslin family. Also, as noted in [LE2], if μ is a Radon measure on a locally compact Hausdorff space, then it follows from [KU1, p. 95] that $\mathcal{M}$ is a Suslin family. These observations obviate most of the

complication which appears to be introduced by considering the Suslin operation, in contrast to consideration of continuous images of Polish spaces.

Let us consider some cases where measurable selections do not exist, i.e., $\mathcal{S}(F)=\varnothing$. The most elementary example is the case where $f: T\to X$ is not a measurable function and $F(t)=\{f(t)\}$ for $t\in T$. Then f is obviously the only selection of F. Suppose in particular that $T=X=[0,1]$, μ is outer Lebesgue measure over T, $T\supset S\notin\mathcal{M}$, and f is the characteristic function of S. Now Gr F is measurable with respect to 2-dimensional (outer) Lebesgue measure, since it has measure 0 (but Gr F is not Borel). Thus, one can have $\mathcal{S}(F)=\varnothing$ even when Gr F has fairly nice measurability.

We shall see in Theorem 5.3, for example, that if Gr F is a Borel, or even Suslin, subset of R^2, then F has a selection which is a Lebesgue measurable function, and which will also (by Lusin's theorem, Theorem 14.1 below) be a.e. equal to a Borel function. However, with Gr F Borel in R^2 there need not exist a selection of F which is a Borel function on *all* of T, i.e., if $\mathcal{M}=\mathcal{B}(T)$, we may have $\mathcal{S}(F)=\varnothing$. This was shown by an example given in Novikov's [NO1] and in [LS] (see § 3) and a later example by Blackwell [BL].

Where assumptions on Gr F are not made, it helps for F to be measurable and to have closed values. The following three examples of Dauer and Van Vleck [DV] illustrate some bad behavior of measurable set-valued functions which are not closed valued. For Examples 2.1, 2.2, and 2.3, we let $T=X=[0,1]$, μ be outer Lebesgue measure and $S\subset T$ have inner measure 0 and outer measure 1.

Example 2.1 [DV]. Let Q, Q' be disjoint dense countable subsets of $[0,1]$. Let $F(t)=Q$ for $t\in S$ and $F(t)=Q'$ for $t\in T\backslash S$. Then F is not measurable, since $F^-(\{a\})=S$ for $a\in Q$. However, F is weakly measurable, since $F^-(U)$ is $\varnothing$ or T for open $U\subset X$. Also, F is countable-valued. In [DV] it is shown that $\mathcal{S}(F)=\varnothing$.

Example 2.2 [DV]. Let $F(t)=X\backslash\{t\}$ if $t\in S$ and $F(t)=X$ if[1] $t\in T\backslash S$. Then F is measurable but Gr $F\notin\mathcal{M}\otimes\mathcal{B}(X)$.

Example 2.3 [DV]. Let F be as in Example 2.2 and let $G(t)=\{t,1\}$ for $t\in T$. Then[1] $(F\cap G)(t)=\{1\}$ for $t\in S$ and $(F\cap G)(t)=\{t,1\}$ for $t\in T\backslash S$. While F and G are measurable, $F\cap G$ is not even weakly measurable.

The following example of Kaniewski (privately communicated via Kuratowski and Himmelberg) shows that a weakly measurable *closed-valued* function need not be measurable, even when T and X are Polish. Leese [LE3, p. 73, Example (vi)] had independently shown that this is true (with the same T and $\mathcal{M}$) without exhibiting an example.

Example 2.4 (Kaniewski). Let $T=[0,1]$, Z be the set of irrationals, $X=T\times Z$, $p(t,n)=t$ for $(t,n)\in X$, $F=p^{-1}$, and $\mathcal{M}=\mathcal{B}(T)$. Since p is an open mapping, F is weakly measurable, in fact lsc. That F is not measurable is seen by taking a closed $K\subset X$ such that $p(K)$ is not Borel.

3. Pre-1949 history. A reasonable starting point for an historical discussion of measurable selections appears to be Lusin's 1930 book [LS, Chap. IV] and Novikov's 1931 paper [NO1]. Reference [LS] is a classic treatment of the theory of Suslin sets in R^n, the early development of which is primarily due to Suslin,

[1] Typographical error in [DV].

Sierpinski, and Lusin. Both [LS] and [NO1] make nonspecific reference to the other author's work, but neither cites these references. Both treat implicit functions in a way which constitutes a setting for the subject of Borel function selections. They consider a Borel $f\colon R^m \times R^p \to R^q$ from which one may define (we are rendering usages)

$$F(t) = R^p \cap \{y\colon f(t, y) = 0\} \quad \text{for } t \in R^m,$$

$$E = R^m \cap \{t\colon F(t) \neq \varnothing\}.$$

Both showed the following:

(i) If each $F(t)$ is countable, then E is a Borel set and $F|E$ has a Borel function selection.

(ii) Without the requirement that F be countable-valued, E need not be a Borel set and $F|E$ need not have a Borel function selection.

In giving (i), Lusin showed more, by way of decomposing Gr F—see § 10. Note that the assumption that F is countable-valued is a severe restriction. Achievement of (ii) centered on showing that there exist disjoint complementary Suslin (i.e., CA) sets which cannot be separated by Borel sets, the original demonstration of which Lusin credits to Novikov. Incidentally, since Gr F is Borel, we now know that F has a Lebesgue measurable selection (Corollary 5.2 below), which by other work of Lusin (Theorem 14.1 below) agrees a.e. with a Borel function.

Lusin also addressed the question: Given $g\colon E \to R^p$ such that $f(t, g(t)) = 0$ for $t \in E$, does there exist a Borel $h\colon R^m \to R^p$ such that $g = h|E$? This is a case of the extension problem noted in § 16 below and discussed in [HM2].

The usages "uniforme" (i.e., single-valued) and "multiforme" (i.e., multi-valued) functions in [LS], [NO1], and earlier works appear to have given rise to the term "uniformization," which is conceptually the same as "selection," but with a different emphasis on properties of the selections. This topic affords additional early history—see § 12.

Another early result is the following of Saks [SK, Lem. 7.1, p. 282] (first edition was 1933): If $X \subset R$ is compact, and $f\colon X \to T$ is continuous, then there exists a Borel $A \subset X$ such that $f(A) = f(X)$ and $f|A$ is one-to-one. Then $[f|A]^{-1}$ is a Borel function selection of $F = f^{-1}$—see Kuratowski [KU1, Chap. III, § 39, V, Thm. 3] (first edition was 1933). Also X could be any compact metric space, since such is a continuous image of a Cantor set. Saks' lemma was generalized to Lebesgue measurable f by Federer and Morse [FM] (1943), but in a way which does not appear to generalize the measurable selection consequence just stated. Mackey [MC1, Lem. 1.13] (1952) applied [FM] as noted in § 11 below. Baker [BA, Lem. 3] (1965) adapted Mackey's argument with [FM] to generalize Saks' lemma to the case where T and X each have a topology with countable base and are "almost Hausdorff" as defined in Theorem 12.4 below; the same Borel selection consequence follows.

The earliest result on existence of measurable selections without assuming countability or compactness of the values of F is von Neumann's in 1949, which we will come to in § 5. (See addenda (i), (ii).)

4. Closed-valued functions. In this section we survey selection results when F is closed-valued, generally without assumptions on Gr F. We remind the reader that (see § 2) μ is a measure over T for which $\mathcal{M}$ is the σ-algebra of measurable

sets, X is topologized, $\varnothing \neq F(t) \subset X$ for $t \in T$, and $\mathcal{S}(F)$ is the set of measurable selections of F.

The assumption that F is closed-valued is not as restrictive as might first appear. For example, if T is topologized as a T_1 space, $f: X \to T$ is continuous, and $F = f^{-1}$, then F is automatically closed-valued. We shall see in the next section how this observation may be used to derive graph-conditioned selection results from results of the sort given in this section.

Probably the most important result to date in the entire theory of measurable selections is the following theorem. Its hypotheses are sufficiently weak that it suffices for most applications, and numerous measurable selection results have been derived from it, including the earlier result of von Neumann [NE], Theorem 5.1 below. It has also been generalized somewhat.

THEOREM 4.1. *If F is weakly measurable and closed-valued and X is Polish, then $\mathcal{S}(F) \neq \varnothing$.*

Because Theorem 1 is so important, we discuss its origin in detail. This result was given in 1965, by Kuratowski and Ryll-Nardzewski in stronger form as Theorem 1 of [KRN] (see also [KU2, p. 74]), and independently by Castaing in more restricted form as Théorème 3 of [CA1]. In [KRN], $\mathcal{M}$ is permitted to be $\mathcal{L}_\sigma$, where $\mathcal{L}$ is a field (i.e., Boolean algebra) of subsets of T; this hypothesis is weaker than the requirement that $\mathcal{M}$ be a σ-algebra. Castaing's statement in [CA1] is an announcement, with proof deferred to Théorème 3 of [CA2] (1966) and Théorème 5.2 of his thesis [CA4, 5] (1967). In [CA1, 2, 4, 5] the assumption is made and utilized in proof that F is measurable, not just weakly measurable. Characteristic of Bourbaki measure foundations, it is also hypothesized that μ is a Radon measure on T, a locally compact space (in [CA1, 2] a compact space), but the method of proof (which uses a sifting, i.e., "criblage," of X) requires neither a topology nor a measure on T. The proofs in [KRN] and [CA2, 4, 5] construct in different ways a Cauchy sequence of functions which converges uniformly to a selection. Castaing was the first to show, by Théorème 5.4 of [CA4, 5] (same hypothesis as Théorème 5.2), that one can in fact obtain what has been termed a Castaing representation of F—see Theorem 4.2 below. (See addendum (ii).)

Subsequent to the appearance of [KRN] and [CA5], workers in the field became aware of the existence of Rokhlin's 1949 statement [RK2, § 2.9, Lem. 2] which was similar to Theorem 4.1 except that $\mathcal{M}$ was specialized to be isomorphic to the σ-algebra generated by the Lebesgue measurable subsets of [0, 1] and a countable family of atoms. In recent years Rokhlin has often been credited with the origination of, in effect, Theorem 4.1. However, although the statement in [RK2] is correct, the proof is not—the recursive construction does not satisfy $(10n)$.[2] (See addendum (iii).)

A special variant of Theorem 4.1 was given in 1962 by Dixmier [DI, Lem. 2]: assuming also $T = X$, $\mathcal{M} = \mathcal{B}(T)$, and $\{F(t): t \in T\}$ is a partition of T, he obtained a selection f of F with range f Borel (Corollary 11.2(ii) below). Now there is a fairly easy metric argument in [HM2, Theorem 3.3] showing that where X is separable metric, F is weakly measurable only if $\operatorname{Gr} \operatorname{cl} F \in \mathcal{M} \otimes \mathcal{B}(X)$. This argument may be used (i) to deduce from [DI], Theorem 4.1 with the added conditions that T is

[2] For confirming our finding on this point we are indebted to Roman Pol and Pawel Szeptycki, who reviewed the original Russian version, and to Fred Van Vleck who reviewed the English translation.

Polish and $\mathcal{M} = \mathcal{B}(T)$ (by applying [DI] to the partition $\{\{t\} \times F(t): t \in T\}$ of Gr F), and (ii) to deduce this special case of Theorem 4.1 from von Neumann's earlier theorem, discussed in § 5 below. Dixmier used sifting methods. Plausibly [DI, Lem. 1] could be used with Castaing's argument to prove [CA4, 5, Thm. 5.2] with F weakly measurable rather than measurable.

An additional source for the proof of Kuratowski and Ryll-Nardzewski of Theorem 4.1 is [PR1], the first text on measurable selections.

That a closed-valued F is well-behaved is seen in the following theorem, which summarizes properties of such F given by Castaing [CA9], Rockafellar [RC3], Himmelberg [HM2], Himmelberg and Van Vleck [HV6], Leese [LE3], and Delode, Arino, and Penot [DAP2].

THEOREM 4.2. *Suppose F is closed-valued. Consider the following*:

(i) $F^-(B) \in \mathcal{M}$ *for* $B \in \mathcal{B}(X)$;
(ii) $F^-(K) \in \mathcal{M}$ *for closed* $K \subset X$, *i.e., F is measurable*;
(iii) $F^-(U) \in \mathcal{M}$ *for open* $U \subset X$, *i.e., F is weakly measurable*;
(iv) *for some metric d on X, $d(x, F(\cdot))$ is a measurable function for* $x \in X$;
(v) $\operatorname{Gr} F \in \mathcal{M} \otimes \mathcal{B}(X)$;
(vi) $\operatorname{Gr} F$ *is the Suslin family generated by* $\mathcal{M} \otimes \mathcal{B}(X)$;
(vii) $\pi_T(A \cap \operatorname{Gr} F) \in \mathcal{M}$ *for* $A \in \mathcal{M} \otimes \mathcal{B}(X)$;
(viii) $\pi_T(A \cap \operatorname{Gr} F) \in \mathcal{M}$ *for A in the Suslin family generated by* $\mathcal{M} \otimes \mathcal{B}(X)$;
(ix) *F has a Castaing representation*;
(x) *there exists a measurable $f_i: T \to X$ for $i = 1, 2, \cdots$, such that $\{f_1(t), f_2(t), \cdots\} \cap F(t)$ is dense in $F(t)$ for $t \in T$ and $T \cap \{t: f_i(t) \in F(t)\}$ is measurable for $i = 1, 2, \cdots$*;
(xi) $F^-(C) \in \mathcal{M}$ *for compact* $C \subset X$.

We then have the following:

(a) (ix)$\Leftrightarrow$(x).
(b) *If X has a countable base, then* (iii)$\Rightarrow$(v).
(c) *If X is regular and a continuous image of a space with a countable base, then* (ii)$\Rightarrow$(v).
(d) *If X is separable metric, then* (ii)$\Rightarrow$(iii)$\Leftrightarrow$(iv)$\Rightarrow$(xi), (iii)$\Rightarrow$(v), *and* (ix)$\Rightarrow$(xi). *If also X is σ-compact, then* (ii)$\Leftrightarrow$(iii)$\Leftrightarrow$(ix)$\Leftrightarrow$(xi).
(e) *If X is separable metric and F is complete-valued, then* (iii)$\Leftrightarrow$(ix)$\Leftrightarrow$(xi).
(f) *If $\mathcal{M}$ is a Suslin family and X is regular and a weakly Suslin space, then* (ii)$\Leftrightarrow$(v)$\Rightarrow$(ix).
(g) *If $\mathcal{M}$ is a Suslin family and X is metric Suslin, then* (i) *through* (x) *are equivalent.*

Proof. The proof of [RC6, Thm. 1B] proves (a); (b) and (c) are given as [LE3, Thms. 3.6 and 3.7]; (d) and (e) come from [HM2, Thms. 3.5 and 5.6] and [HV6, Thm. 1']; (f) follows from Theorem 6.1 below and [LE3, Thm. 3.9], observing that F is of Suslin type.

It remains to prove (g). From what has been proved and obvious observations, (viii)$\Rightarrow$(vii)$\Rightarrow$(i)$\Rightarrow$(ii)$\Rightarrow$(iii)$\Rightarrow$(iv)$\Rightarrow$(v)$\Rightarrow$(vi) and (ix)$\Leftrightarrow$(x). The proof of [RC6, Thm. 1B] shows (ix)$\Rightarrow$(ii). Leese (personal communication) has deduced (ix) and (viii) from (vi) as follows. One shows that $\mathcal{M} \otimes \mathcal{B}(X)$ and hence the Suslin family generated by $\mathcal{M} \otimes \mathcal{B}(X)$ are contained in the Suslin family generated by $\{S \times K: S \in \mathcal{M}$ and $K \subset X$ is closed$\}$. Hence (vi) implies (viii) by [LE5, Thm. 5.5]. It

also follows that (vi) implies that F is of Suslin type by [LE2, Thm. 6], from which (ix) follows by Theorem 6.1 below; thus (vi)$\Rightarrow$(ix). □

The most comprehensive conclusion in Theorem 4.2 is (g). For σ-finite complete μ and Polish X, Castaing [CA9, Lem. 1] gave (i)$\Leftrightarrow$(iii)$\Leftrightarrow$(iv)$\Leftrightarrow$(v)$\Leftrightarrow$(ix) (and (iv)$\Rightarrow$(ii)$\Rightarrow$(iii) is elementary). Rockafellar [RC3, Thm. 1] added that these are equivalent to (x). Very recently, Delode, Arino, and Penot [DAP2] added (vi), (vii), and (viii) to the equivalence and weakened the requirement on μ to $\mathcal{M}$ being a Suslin family. Leese (and subsequently Delode [DL, 2.9]) observed that Polish X could be weakened to metric Suslin X. Under separable metric X, Himmelberg [HM2] has given (b), (c), (d), (e), and various related facts.

The equivalence (iii)$\Leftrightarrow$(ix) in Theorem 4.2(e) is useful in both directions. For example, Rockafellar [RC2] applies this equivalence with $X = R^n$ to show measurability of the intersection and the closed vector sum of measurable closed-valued functions. He gives a more comprehensive treatment of related manipulations in [RC6]. (An even more powerful manipulative tool is Leese's theory of Suslin type—see § 6.) The set values involved in Rockafellar's [RC1–6] are primarily epigraphs of convex real-valued functions on a separable reflexive Banach space, parameterized on a complete σ-finite measure space. Measurability of the epigraph-valued function is a key criterion for a convex "integrand" to be "normal" (see [RC3, 6]). The above (iii)$\Leftrightarrow$(ix) is used in showing, for instance, that a convex integrand which is a finite Carathéodory map is normal, and that conjugation of convex normal integrands is reflexive. These facts are, in turn, useful to optimization of convex integral functionals by duality methods. Following is an application of (iii)$\Leftrightarrow$(x). (See addendum (v).)

COROLLARY 4.3 (Rockafellar [RC6, Cor. 1D]). *Suppose $\mathcal{M}$ is a Suslin family, X is Polish, and for a.e. $t \in T$, $F(t) = \text{cl interior } F(t)$ (as is true, for instance, if $X = R^n$ and $F(t)$ is an n-dimensional closed convex set). Then F is measurable iff $F^-(\{x\})$ is measurable for $x \in X$.*

In the remainder of this section we give a chronological review of additional results with closed-valued F.

Himmelberg and Van Vleck [HV2, Thm. 5] observed that the completeness requirement in Theorem 4.1 could be put on the values of F (as in Theorem 4.2(e)) rather than on a homeomorph of X. They obtained a precursor to [HV3] (see § 5) with X a metric Lusin space, and various results pertaining to (xi) in Theorem 4.2 and to $\mathcal{M}$ being merely a σ-ring. Reference [HV2] supersedes [HV1].

In extending Scorza-Dragoni's generalization of Lusin's theorem (see § 14), Castaing [CA13, Thm. 5] gave a result to the effect that if T is compact, μ is Radon, X is metric, and F is "approximately lower semi-continuous" and complete-valued, then F has an a.e. measurable selection. Jacobs [JC1] and Himmelberg, Jacobs, and Van Vleck [HJV] gave related results.

In the following theorem, Castaing has substituted existence of a suitable sifting [BO1, Chap. IX, § 6.5] of X for some of the assumptions in Theorem 4.2, motivated by his proof of [CA5, Thm. 5.2].

THEOREM 4.4 [CA15, 16, Thm. 1]. *Suppose X is a Hausdorff space, $((C_1, p_1, \varphi_1), (C_2, p_2, \varphi_2), \cdots)$ is a sifting of X, $F^-(\varphi_n(c)) \in \mathcal{M}$ for $c \in C_n$ and $n = 1, 2, \cdots$, and F is closed-valued. Then there exists a selection of F which is a pointwise limit of measurable functions on T to X with countable range.*

Following is a corollary to this theorem.

Mägerl [MG, Kapitel IV, Korollar 2.4] independently obtained $\mathcal{S}(F) \neq \varnothing$ under the bracketed hypothesis of Corollary 4.5. Valadier [VA4, 5, Lem. 1] obtained the conclusion of Theorem 4.4 in Castaing representation form assuming F is closed-valued, $\operatorname{Gr} F \in \mathcal{M} \otimes \mathcal{B}(X)$, X is Suslin, and μ is complete and σ-finite.

COROLLARY 4.5 [CA16, Cor. 6] *Suppose X is a Suslin {Lusin} space, F is closed-valued, and $F^{-}(A) \in \mathcal{M}$ for every Suslin {Borel} $A \subset X$. Then F has a Castaing representation.*

Following is a novel theorem of Robertson [RB] using a "left set," i.e., the set A. Theorem 2 of [RB] is an antecedent to the "Suslin type" development of his student S. J. Leese (see § 6).

THEOREM 4.6 [RB, Thm. 4]. *Suppose F is measurable and closed-valued and X is a continuous image of a set $A \subset R$ with the property that* inf $D \in A$ *for $\varnothing \neq D \subset A$. Then $\mathcal{S}(F) \neq \varnothing$.*

The next theorem is a generalization by Leese (personal communication) of Kuratowski's [KU4, Thm. 5.2]. The latter has T and X metric Suslin and concludes $\mathcal{S}(F) \neq \varnothing$.

THEOREM 4.7. *Suppose T is topologized and let $\mathcal{L}$ be the Suslin family generated by the closed sets of T. Suppose $\mathcal{M} \supset \mathcal{L}$, X is regular and weakly Suslin, F is closed-valued, and $F^{-}(A) \in \mathcal{L}$ for closed $A \subset X$. Then F is of Suslin type (see § 6) and hence F has a Castaing representation.*

Proof. Note § 6 and follow the proof of [RB, Lemma 1]. □

Maitra and Rao have weakened the separability of X in Theorem 4.1, adding other restrictions, as follows.

THEOREM 4.8 [MR1, Cor. 4]. *Assume the Zermelo–Frankel axioms, the axiom of choice, and Martin's axiom. Suppose $T = R$, $\mathcal{M}$ is the set of Lebesgue measurable subsets of R or the set of subsets of R having the Baire property, X is complete metric with base of cardinality less than $2^{\aleph_0}$, and F is closed-valued and weakly measurable. Then $\mathcal{S}(F) \neq \varnothing$.*

Artstein [AR2, Prop. 4.12] has shown that under conditions resembling those of Theorem 4.9 given next, if for $i = 1, 2, \cdots$, $(F_1, F_2, \cdots)$ "converges weakly" to F, and $f \in \mathcal{S}(F)$, then there exists $f_i \in \mathcal{S}(F_i)$ for $i = 1, 2, \cdots$, such that $(f_1, f_2, \cdots)$ converges weakly to f.

THEOREM 4.9 [AR2, Thm. 2.7]. *Suppose $T = [0, 1]$, μ is Lebesgue measure, $X = R^n$, and F is closed-valued. Then there exists a closed-valued G: $T \to \mathcal{P}(X)$ such that* Gr *G is Borel, $G(t) \subset F(t)$ for a.e. $t \in T$, and the a.e. measurable selections of F coincide with those of G.*

Many selection results make the strong assumption that F is compact-valued. Following is such a result by Leese which has weak assumptions in other respects. Combined with Theorem 4.11, we have a rather general selection result for closed-valued F. Theorem 4.10 is given in [LE5] under kinds of generalizations mentioned in § 13 below. As observed by Leese, 4.10(i) implies [RB, Thm. 1], which assumes X is a Hausdorff continuous image of a separable metric space.

THEOREM 4.10 [LE5, Thms. 4.1 and 4.2]. *Suppose F is compact-valued and measurable. Then $\mathcal{S}(F) \neq \varnothing$ providing one of the following holds:*

(i) *there exist closed $K_1, K_2, \cdots \subset X$ such that for each distinct pair of points in X, some K_n contains one and not both (Leese's Condition (S)); or*

(ii) *$\mathcal{B}(X)$ is generated by a family of closed sets whose cardinality is at most the first uncountable cardinal (Leese's Condition (B)).*

THEOREM 4.11 [LE3, Thm. 8.6]. *Suppose $\mathcal{M}$ is a Suslin family, X is regular and analytic in the sense of being a continuous image of a countable intersection of countable unions of closed compact subsets of some topological space, and F is measurable and closed-valued. Then there exists a measurable compact-valued $G: T \to \mathcal{P}(X)$ such that $G(t) \subset F(t)$ for $t \in T$.*

5. Graph-conditioned theorems. In this section we recount the development of selection theorems based on properties of Gr F rather than on conditions on the values of F. The two topics are linked, as shown in Theorem 4.2, in the proof of Theorem 5.3, and, more extensively, in § 6. We again remind the reader (for the last time) that T, $\mathcal{M}$, μ, X, F and $\mathcal{S}(F)$ are fixed in § 2. (See addendum (i).)

The present topic begins with the 1949 selection result of von Neumann (also given with same proof in [PR1]).

THEOREM 5.1 [NE, Lem. 5]. *Suppose $T = R$, X is a Suslin subset of a Polish space, $f: X \to T$ is continuous and surjective, $F = f^{-1}$, and μ arises from a non-decreasing right-continuous bounded $g: R \to R$. Then $\mathcal{S}(F) \neq \varnothing$.*

Proof (outline). Represent X as a continuous image of ω^ω, topologized homeomorphic to the irrationals, where $\omega = \{1, 2, \cdots\}$. For each $t \in T$, select the lexicographic minimum in ω^ω of the counterimage of $F(t)$ and map this back to $F(t)$. □

This in effect is what von Neumann stated. His proof is still valid if the conditions on T and μ are replaced by the condition that T be Hausdorff and $\mathcal{M}$ contain the Suslin sets of T. In this generality we note a corollary of a form (Suslin graph) in which von Neumann's theorem is often given. Recall that when T is Hausdorff, $\mathcal{M}$ contains the Suslin sets of T if, in particular, $\mathcal{M} \supset \mathcal{B}(T)$ and μ is an outer measure.

COROLLARY 5.2. *Suppose T and X are Polish,* Gr F *is Suslin, and $\mathcal{M}$ contains the Suslin sets of T. Then $\mathcal{S}(F) \neq \varnothing$.*

Proof. Let $f = \pi_T$ and, replacing X by $T \times X$, apply Theorem 5.1 (generalized as noted). □

Von Neumann's result seems to have been little known until around 1965 when it surfaced separately in mathematical economics, notably in Aumann's [AU1, 2], and in control theory, although it was referenced and used by Mackey [MC2] in 1957, for example. (We know of three leaders in measure theory who were unaware of it in 1971.) Ironically, one suspects that its recognition suffered from submergence under the prolific output of a giant.

We now depart from chronology to note how graph-conditioned theorems, including 5.1 and 5.2 just given, can be derived from a closed-valued result such as Theorem 4.1. Castaing [CA4, p. 123] was the first to do this—one lifts a set-valued function with Suslin graph to a measurable closed-valued function into a Polish space. This idea was used by Himmelberg and Van Vleck in [HV3] to prove a version of Theorem 5.3 in a more direct way than in [CA4]. It has been exploited more extensively by Leese in his "Suslin type" approach—see § 6.

The following theorem and proof are largely given by Leese [LE5, Thm. 7.4]. Both Theorem 5.3 and Corollary 5.4 were for the most part contained in a personal communication we received from Castaing in 1972 under the stronger

assumption that T and X are Suslin spaces (see also [SB3, Thm. 2] and [HJ, Thm. III.9.6]). Here it is assumed what seems just enough to make the proof of Theorem 5.3 work—the method is closer to that of [HV3] than to Castaing's.

THEOREM 5.3. *Suppose T is topologized as a T_1 space,* Gr F *is weakly Suslin, and $\mathcal{M}$ contains each weakly Suslin subset of T. Then F has a Castaing representation.*

Proof. Take a Polish space P and a continuous surjective $\varphi: P \to \text{Gr } F$. Let $G = (\pi_T \circ \varphi)^{-1}$. Since $\pi_T \circ \varphi$ is continuous and T is a T_1 space, G is closed-valued. Also, G is measurable, because for closed $A \subset P$, $G^-(A) = \pi_T(\varphi(A))$ so $G^-(A)$ is a continuous image of the Polish space A, whence $G^-(A) \in \mathcal{M}$. By Theorem 4.2(e), G has a Castaing representation $\{g_1, g_2, \cdots\}$. Then for $i = 1, 2, \cdots$, $f_i \equiv \pi_X \circ \varphi \circ g_i \in \mathcal{S}(F)$, and $\{f_1(t), f_2(t), \cdots\}$ is dense in $F(t)$ for $t \in T$. □

COROLLARY 5.4. *Suppose T and X are Suslin spaces, $\mathcal{M}$ contains each Suslin subset of T, $f: X \to T$ is continuous and surjective, and $F = f^{-1}$. Then F has a Castaing representation.*

Proof. Since f is continuous, Gr F is closed in $T \times X$, and hence is Suslin [HV3, Lem.]. Thus, Theorem 5.3 applies. □

Christensen and Jayne have shown [CH, Thm. 4.3] that a continuous map on a Polish space onto a compact metric space need not have a Borel function inverse; of course, when $\mathcal{M} = \mathcal{B}(T)$, $\mathcal{M}$ will not ordinarily contain all Suslin sets of T.

Hoffman-Jørgensen has obtained a Borel function inverse of a Borel function, as follows (a somewhat related result in [CH, Thm. 4.3] is a specialization of Theorem 6.1 below).

THEOREM 5.5 [HJ Thms. III.11.B. 8–11]. *Suppose T and X are Suslin spaces, $f: X \to T$ is a Borel function, $F = f^{-1}$, and $\mathcal{M} = \mathcal{B}(T)$. Then $\mathcal{S}(F) \neq \varnothing$ providing one of the following holds:*

(i) *F is weakly measurable and either F is compact-valued or* Gr *F is Polish;*

(ii) Gr F is σ-compact;

(iii) Gr F *is Lusin and F is countable-valued.*

Mägerl's [MG, Kapitel III, Satze 2.6, 2.7] follow from Theorem 5.3 by letting T be Hausdorff and μ be an outer measure for Satz 2.6 and T be locally compact Hausdorff and μ be Radon for Satz 2.7.

Returning to chronology, the first generalization of von Neumann's result was the following by Sion in 1960. Sion's paper has been well known in uniformization theory, but belatedly known in measurable selection theory; it does not reference [NE]. Note that his condition on X is satisfied when X is Polish. He made a weaker assumption on Gr F than that given here, viz., that Gr F is "analytic" in $T \times X$, by which he means a continuous image of a countable intersection of countable unions of compact subsets of a Hausdorff space. In [SN, Cor. 4.4], the assumption on μ is omitted, but $\mathcal{M}$ is generated by the "analytic" subsets of T.

THEOREM 5.6 [SN, Cor. 4.5]. *Suppose T is Hausdorff, μ is an outer measure, $\mathcal{M} \supset \mathcal{B}(T)$, X is a regular Hausdorff Lindelöf space with a base of cardinality no greater than $\aleph_1$, and* Gr *F is Suslin. Then $\mathcal{S}(F) \neq \varnothing$.*

The next graph-conditioned theorem to appear was the following of Blackwell and Ryll-Nardzewski in 1962. It is unusual in imposing measure-theoretic

conditions on X. Its motivation was to prove that if μ is a probability measure, f is a real random variable on T, and range f is not Borel, then there does not exist an everywhere proper conditional distribution given f. It is applied again in [FU] and [BD].

THEOREM 5.7 [BRN, Thm. 2]. *Suppose T and X are Borel subsets of Polish spaces, $\mathcal{M} \subset \mathcal{B}(T)$, $\mathcal{M}$ is countably generated, and* Gr $F \in \mathcal{M} \otimes \mathcal{B}(X)$. *Suppose also that there exists $g: T \times \mathcal{B}(X) \to R$ such that $g(t, \cdot)$ is a probability measure on $\mathcal{B}(X)$ for $t \in T$, $g(\cdot, B)$ is a measurable function for $B \in \mathcal{B}(X)$, and $g(t, F(t)) > 0$ for $t \in T$. Then $\mathcal{S}(F) \neq \varnothing$.*

Aumann [AU3] in 1967 made a significant advance with the following graph-conditioned theorem which involves no topological assumption, although as observed by Sainte-Beuve in [SB3], one may just as well assume that X is a Lusin space and $\mathcal{A} = \mathcal{B}(X)$. Following [AU3], we say $(X, \mathcal{A})$ is a *standard space* if $\mathcal{A}$ is a σ-algebra of subsets of X and there is a one-one correspondence between X (not necessarily topologized) and R which induces a one-one correspondence between $\mathcal{A}$ and $\mathcal{B}(R)$.

THEOREM 5.8 [AU3]. *Suppose μ is σ-finite, $(X, \mathcal{A})$ is a standard space, and* Gr $F \in \mathcal{M} \otimes \mathcal{A}$. *Then there exist $S \in \mathcal{M}$ and a selection f of $F|S$ such that $\mu(T \backslash S) = 0$ and $f^{-1}(A) \in \mathcal{M}$ for $A \in \mathcal{A}$.*

An example in [AU3] due to Lindenstrauss shows that one may not let $\mathcal{A}$ be an arbitrary σ-algebra on X and Aumann shows that σ-finiteness may not be omitted. Also given in [AU3] is an interesting discussion of the question of whether a theorem such as 5.2 holds if Gr F is complementary Suslin.

Complementary Suslin sets also arise in the following result of Castaing.

THEOREM 5.9 [CA8, Prop. 1]. *Suppose T is a Suslin space, X is a metric Suslin space, F is complete-valued, $T \backslash F^{-}(A)$ is Suslin for every closed $A \subset X$, and $\mathcal{M} = \mathcal{B}(T)$. Then* Gr F *is a Suslin space iff F has a Castaing representation.*

Sainte-Beuve generalized Theorem 5.8 in [SB1, 2, 3]. She assumed $\mathcal{M}$ is complete (no assumption on μ) and X is Suslin instead of Lusin and obtained $\mathcal{S}(F) \neq \varnothing$. Following is a further generalization by Leese [LE2, Cor. to Thm. 7] yielded by Theorem 6.1 below.

THEOREM 5.10. *Suppose $\mathcal{M}$ is a Suslin family, X is a weakly Suslin space, and* Gr $F \in \mathcal{M} \otimes \mathcal{B}(X)$. *Then F has a Castaing representation.*

To see that this, and similarly [SB1, 2, 3], generalize Theorem 5.8, extend the μ of Theorem 5.8 to an outer measure μ^* so that each μ^* measurable set differs from a μ measurable set by a set contained in a set of μ measure zero; the σ-algebra of μ^* measurable sets is a Suslin family. Apply Theorem 5.10 and check the counterimages of a countable base of X.

Dauer and Van Vleck have shown how measurable selections of cl F can be approximated by those of F. A metric on X induces the essential supremum pseudometric in $\mathcal{S}(F)$ and in Theorem 5.11 we topologize $\mathcal{S}(F)$ with this. A similar conclusion is in [LE3, Thm. 8.7], assuming F is weakly measurable instead of Gr F is Suslin.

THEOREM 5.11 [DV, Thms. 1, 2]. *Suppose T is locally compact separable metric, μ is Radon, X is metric Suslin and* Gr F *is Suslin. Then $\mathcal{S}(\text{cl } F) = \text{cl } \mathcal{S}(F)$. If instead the hypothesis of Theorem* 5.8 *is satisfied, then this conclusion holds for a.e. selections.*

When μ is Borel regular and σ-finite, von Neumann's theorem yields an a.e. selection of F which is a Borel function, by application of Lusin's theorem. The following version proved by Federer [WS, Thm. 4.1] obtains a Borel selection on most of T without assuming Borel regularity of μ. This version may also be proved by von Neumann's argument in Theorem 5.1 or, as Castaing has pointed out, by observing that every Suslin space is a Radon space [BO2, Chap. IX, § 3.3] and applying Theorem 5.3.

THEOREM 5.12. *Suppose T is Hausdorff, $\mathcal{M} \supset \mathcal{B}(T)$, μ is a bounded outer measure, X is a Suslin subset of a Polish space, $h: X \to T$ is continuous and surjective, $F = h^{-1}$, and $\varepsilon > 0$. Then there exist a compact $C \subset T$ and $f \in \mathcal{S}(F|C)$ such that $\mu(T \backslash C) < \varepsilon$ and f is a Borel function.*

We close this section with what seem to be the main graph-conditioned results of Leese's [LE5]. He generalizes in the following directions not shown here: (a) a "partial uniformization," i.e., existence of a well-behaved compact-valued subfunction of F, as in Theorem 4.11, is given, (b) properties are given of the T-projection of Gr F (F not necessarily defined on all of T), (c) X weakly Suslin as defined here, is sometimes weakened to X "analytic" and (d) non-σ-algebras are sometimes employed in place of $\mathcal{M}$ (see § 13).

THEOREM 5.13 [LE5, Thm. 5.5]. *Suppose $\mathcal{M}$ is a Suslin family,* Gr F *is in the Suslin family generated by $\{S \times K: S \in M$ and $K \subset X$ is closed$\}$, and X is weakly Suslin. Then $\mathcal{S}(F) \neq \varnothing$.*

THEOREM 5.14 [LE5, Thm. 6.2 or 6.3]. *Suppose T is topologized, $\mathcal{M}$ contains the Suslin family generated by the closed sets of T, X is weakly Suslin and* Gr F *is in the Suslin family generated by the closed sets of $T \times X$. Then $\mathcal{S}(F) \neq \varnothing$.*

6. Set-valued functions of Suslin type. We summarize in this section Leese's Suslin type approach, given in [LE2]. Theorem 6.1 is a succinct statement from which a great deal of the above results and of those in § 7 may be readily deduced. The theme of lifting F to a well-behaved map into the subsets of a Polish space, which underlies this development, has antecedents in work of Castaing [CA4], Himmelberg and Van Vleck [HV3], and Robertson [RB], as noted in § 4 and § 5.

We follow Leese [LE2], and add the bracketed "weak" version in the definition and associated theorems, in saying F *is of* {*weak*} *Suslin type* if there exist a Polish space P, a continuous $\varphi: P \to X$, and a {weakly} measurable closed-valued $G: T \to \mathcal{P}(P)$ such that $F(t) = \varphi(G(t))$ for $t \in T$. (Note that the significance of the word "weak" here differs from its significance in the definition of weak Suslin space.) The F in Kaniewski's Example 2.4 is of weak Suslin type but not of Suslin type. When F is of Suslin type, it is of weak Suslin type.

By Theorem 4.2(e) and the proof of Theorem 5.3, one easily obtains the following.

THEOREM 6.1 [LE2, Thm. 7]. *If F is of weak Suslin type, then F has a Castaing representation, so $\mathcal{S}(F) \neq \varnothing$.*

By itself, Theorem 6.1 adds little to prior knowledge. The usefulness of Leese's [LE2], which is considerable, lies in showing that many kinds of F are of Suslin type, and in giving additional properties of such F; this development holds for weak Suslin type also.

The following result of Robertson shows that Suslin type and weak Suslin type are the same thing in an important case.

THEOREM 6.2 [RB, Thm. 3]. *Suppose $\mathcal{M}$ is a Suslin family, X is a metrizable Suslin space, and F is closed-valued. Then F is measurable iff F is weakly measurable.*

COROLLARY 6.3. *If $\mathcal{M}$ is a Suslin family, then F is of Suslin type iff F is of weak Suslin type.*

Proof. Apply Theorem 6.2 to the G of the above definition. □

If X is Hausdorff and F is of weak Suslin type, then $\operatorname{Gr} F$ is in the Suslin family generated by $\{A \times B: A \in \mathcal{M} \text{ and } B \subset X \text{ is closed}\}$. The proof in [LE2, Thm. 5] must be modified with "weak" by using an open sifting and, as Leese has pointed out, by adding details to show $x = \varphi(y)$ on page 405.

Suppose $\mathcal{M}$ is a Suslin family and X is Hausdorff. Then the class of set-valued functions of Suslin type is a Suslin family (operating pointwise on T), in particular it is closed under countable union and intersection. It is also closed under countable Cartesian product. If also X is a topological vector space, then this class is closed under vector addition, multiplication by a measurable scalar function, and formation of closed convex hulls. These properties are in [LE2].

In each of the following cases F is of {weak} Suslin type (largely in [LE2]–[LE4] corrects the proof of [LE2, Thm. 6]):

(i) X is Polish and F is closed-valued and {weakly} measurable;

(ii) $G: T \to \mathcal{P}(X)$ is compact-valued and {weakly} measurable, X is separable metric with completion $\hat{X}$, i embeds X in $\hat{X}$, and $F = i \circ G$;

(iii) $\mathcal{M}$ is a Suslin family, X is a regular space and a Suslin space, and F is {weakly} measurable and closed-valued;

(iv) $\mathcal{M}$ is a Suslin family, X is a weakly Suslin space, and $\operatorname{Gr} F \in \mathcal{M} \otimes \mathcal{B}(X)$;

(v) T is a T_1 space, $\mathcal{M}$ contains each weakly Suslin subset of T, and $\operatorname{Gr} F$ is weakly Suslin.

With these observations and other hints in [LE2], one may readily deduce from Theorem 6.1 all of 4.1 (the primary basis of 6.1), 4.2(f)(g) [(ii)⇒(ix)], 4.5, 4.7, 5.1, 5.2, 5.3, 5.4, 5.6, 5.8, 5.10, and also 7.1 and 7.2 of the next section.

7. Measurable implicit functions. In this section we fix a topological space Y, a function $g: \operatorname{Gr} F \to Y$, and a measurable function $h: T \to Y$ such that $h(t) \in g(\{t\} \times F(t))$ for $t \in T$. We are concerned with whether there exists $f \in \mathcal{S}(F)$ such that $h = g(\cdot, f(\cdot))$; such an f is a measurable implicit function pertaining to this structure. If we define

$$G(t) = X \cap \{x: g(t, x) = h(t)\} \quad \text{for } t \in T, \tag{7.1}$$

this becomes the question of whether $\mathcal{S}(F \cap G) \neq \varnothing$. Results on this question have been quite numerous, apparently because many applications, notably in control theory, arise naturally in this form. They are sometimes called Filippov type theorems, recalling the lemma of [FI], which was the first selection result of this kind.

Theorems 7.1 and 7.2, due to Leese, 7.3, due to Hoffman-Jørgensen, and 7.4, due largely to Castaing and Himmelberg, are rather general theorems of the sort sought. (Theorem 7.4(i) was given by Castaing [CA9, Corollaire] under Polish X and σ-finite complete μ.) They treat the respective cases where g is $\mathcal{M} \otimes \mathcal{B}(X)$ measurable (i.e., $g^{-1}(U) \in \mathcal{M} \otimes \mathcal{B}(X)$ for open $U \subset X$), continuous, Borel, and a Carathéodory map. Under the latter condition, Lemma 7.5 (which generalizes

[HM2, Thm. 6.1 and KU1, p. 378]) further facilitates application of Theorem 7.1. In Theorems 7.1 and 7.2, $\mathcal{M}$ must be a Suslin family (§ 2), which is weak enough for most applications. Set-valued functions of Suslin type are defined in § 6. Lusin measurability of a function is defined in § 14; it implies ordinary measurability, and when the range space is separable metric and μ is σ-finite, the converse holds.

The statement of Theorem 7.2 in [LE2, Thm. 9] also assumes that X is regular, but Leese has shown [LE3, p. 82] that this assumption may be omitted.

In [HM2, Thm. 7.1] separability of Y is omitted. However, Leese has pointed out, and Himmelberg concurs (both in personal correspondence), that the *argument* fails with this omission; if Y is not separable, $p\colon T \to Y$ and $q\colon T \to Y$ are measurable, and $r(t) = (p(t), q(t))$ for $t \in T$, then r need not be measurable. The same difficulty arises in [HM2, Thms. 7.2 and 7.4]. The validity of these three theorems of [HM2] without separability of Y is an open question.

THEOREM 7.1 [LE2, Thm. 8].[3] *Suppose $\mathcal{M}$ is a Suslin family, X is Hausdorff, Y is separable metric, F is of Suslin type, and g is $\mathcal{M}\otimes\mathcal{B}(X)$ measurable. Then there exists $f \in \mathcal{S}(F)$ such that $h = g(\cdot, f(\cdot))$.*

THEOREM 7.2 [LE2, Thm. 9] and [LE3, p. 82]. *Suppose T is locally compact Hausdorff, μ is Radon, X and Y are Hausdorff, F is of Suslin type, g is continuous, and h is Lusin measurable* (see § 14). *Then there exists $f \in \mathcal{S}(F)$ such that $h = g(\cdot, f(\cdot))$.*

THEOREM 7.3 [HJ, Thm. III.16.10].[3] *Suppose T, X, Y, and* Gr F *are Suslin, g is a Borel function, and either* (i) *$\mathcal{M}$ is generated by the Suslin subsets of T, or* (ii) *$\mathcal{M} \supset \mathcal{B}(T)$ and μ is σ-finite and complete. Then there exists $f \in \mathcal{S}(F)$ such that $h = g(\cdot, f(\cdot))$.*

THEOREM 7.4. *Suppose Y is separable metric, and g is a Carathéodory map. Then there exists $f \in \mathcal{S}(F)$ such that $h = g(\cdot, f(\cdot))$, providing one of the following holds:*

(i) *$\mathcal{M}$ is a Suslin family, X is weakly Suslin, and* Gr $F \in \mathcal{M}\otimes\mathcal{B}(X)$; *or*

(ii) *$\mathcal{M}$ is a Suslin family, X is Hausdorff, and F is of Suslin type; or*

(iii) [HM2, Thm. 7.1] *X is separable metric, F is measurable, and either F is compact-valued or F is closed-valued and X is σ-compact.*

Proof. Under (ii), the definition of Suslin type lets us confine to a Suslin subspace of X. By Lemma 7.5 given next, g is $\mathcal{M}\otimes\mathcal{B}(X)$ measurable. Somewhat as in [HM2, Thm. 7.4], let $\psi(t, x) = (g(t, x), h(t))$ for $t \in T$, $x \in X$. Since Y is separable metric, $\mathcal{B}(Y \times Y) = \mathcal{B}(Y)\otimes\mathcal{B}(Y)$. Hence ψ is $\mathcal{M}\otimes\mathcal{B}(X)$ measurable. With G as in (7.1), Gr $G = \psi^{-1}(\{(y, y)\colon y \in Y\}) \in \mathcal{M}\otimes\mathcal{B}(X)$, so G is of Suslin type (§ 6), hence so is $F \cap G$. By Theorem 6.1, $(F \cap G) \neq \varnothing$. The proof under (i) is similar, using (iv) at the end of § 6. □

---

[3] Leese has observed that "Y is separable metric" in Theorem 7.1 may be weakened to "Y satisfies Condition (S)" (see Theorem 4.10): Then

$$(T \times X)\backslash \text{Gr } G = \bigcup_{n=1}^{\infty} [g^{-1}(K_n) \cap (h^{-1}(Y \backslash K_n) \times X)] \in \mathcal{M}\otimes\mathcal{B}(X),$$

where $\{K_1, K_2, \cdots\}$ is the separating family and G is as in (7.1), whence $F \cap G$ is of Suslin type. From this, Theorem 7.3(ii) follows from Theorem 7.1.

LEMMA 7.5 [LE3, Lem. 14.1]. *If X has a countable base, Y is perfectly normal, i.e., if Y is normal and each open set of Y is an F_σ, and g is a Carathéodory map, then g is $\mathcal{M}\otimes\mathcal{B}(X)$ measurable.*

If in implicit function results of this form we specialize g so that each $g(\cdot, x)$ is constant, replacing it with $k: X \to Y$, we obtain a lifting theorem, i.e., assurance of existence of $f \in \mathcal{S}(F)$ such that $h = k \circ f$. Theorems 7.1–7.4 yield fairly general statements of this nature. An additional lifting result is the following, suggested by Leese.

THEOREM 7.6. *Suppose F is of weak Suslin type, $\mathcal{M}$ is a Suslin family, Y is a Hausdorff space, $k: X \to Y$ is continuous, and $h(t) \in k(F(t))$ for $t \in T$. Then there exists $f \in \mathcal{S}(F)$ such that $h = k \circ f$.*

Himmelberg and Van Vleck [HV2] give lifting results with measurability of F, h, and f defined to mean that inverse images of compact sets are measurable ((xi) of Theorem 4.2) and also with $\mathcal{M}$ being a σ-ring rather than a σ-algebra. McShane and Warfield [MW] (see also [YO]) gave early results in lifting form; these are generalized by [HV2].

Hoffman-Jørgensen [HJ, III. 11] has given such lifting results, not involving F, and also results on the symmetric problem: Given $p: Z \to X$ and $q: Z \to T$, find a "nice" $f: T \to X$ such that $f \circ q = p$.

Under measurability of inverse images of compact sets, Himmelberg and Van Vleck have given an implicit function result as [HV6, Thm. 4(ii)]. Part (i) of that theorem follows from Theorem 7.2, above.

We now review various other results on measurable implicit functions, all of which may be readily deduced from the foregoing, most from Theorem 7.1.

In Filippov's highly influential 1959 lemma [FI], T, X, and Y are in Euclidean spaces, g is continuous, and F is compact-valued and usc, among other restrictions. Another early result is Wazewski's [WZ] (1961), heavily conditioned by compactness. Aronszayn in 1964 permitted F to be G_δ-valued, but constant, reported in [SV]. Olech [OL] in 1965 had g a Carathéodory map with X compact—he obtained a selection by lexicographic minimization, which has componentwise recursiveness in common with Filippov's approach. All of these were directly motivated by control theory applications.

Castaing [CA1] in 1965 (proof in [CA2]) was somewhat more general with X Polish, T compact metric, Y Hausdorff, and F closed-valued with Suslin graph, but with g continuous and μ Radon. Generalizations in similar vein were given in [CA4, 5, § 5] (with weaker assumptions on T) and by Jacobs [JC1, Thm. 2.2; JC2, Thms. 2.5, 2.5′]. Himmelberg, Jacobs, and Van Vleck [HJV, Theorems 3, 3′] put completeness on the values of F instead of on X.

In [HV3, Thms. 2, 3, 4], Himmelberg and Van Vleck primarily assume Gr F is weakly Suslin; Theorems 2 and 3 are implicit function theorems and Theorem 4 is a lifting theorem.

Furukawa's [FU, Lem. 4.6] is a special case of Theorem 7.4 (iii) above with $X \subset R^n$ compact, $Y = R^m$, T a Borel subset of a Polish space, and $\mathcal{M} = B(T)$.

Dauer and Van Vleck [DV] apply Aumann's Theorem 5.8 above, assuming in part μ σ-finite, X Lusin, and g measurable, to obtain an a.e. measurable implicit function. This is generalized independently by Sainte-Beuve [SB3] in fashion similar to her generalization of Theorem 5.8.

Mägerl [MA, Kapitel III, Satz 3.1] has T σ-compact, T and X Hausdorff, μ Radon, Y separable metric, Gr F Suslin, and g a Borel function.

Götz [GZ] has given a schematic tabular summary of measurable implicit function results (and of general measurable selection results and bang–bang results).

8. Convex-valued functions. In this section we assume that X is a linear space and usually that F is convex-valued. Separate topics are discussed, not ordered by chronology or supersession.

We define X' to be the dual of X, $\langle \cdot , \cdot \rangle$ to be the pairing on $X' \times X$, and for $C \subset X$,

$$\varphi(x', C) = \sup \{\langle x', x\rangle : x \in C\} \quad \text{for } x' \in X';$$

thus $\varphi(\cdot, C)$ is the support function of C.

When F is compact-convex-valued, we say F is *scalarly measurable* if for $x' \in X$, $\varphi(x', F(\cdot))$ is a measurable function. A function $f: T \to X$ is *scalarly measurable* if for $x' \in X'$, $\langle x', f(\cdot)\rangle$ is measurable. Thus named by Valadier [VA1], the concept of scalar measurability was (see [VA3, pp. 270–271]) introduced by Kudō [KD] and subsequently used by Richter [RI] and then Kellerer [KE] and Olech [OL] to obtain measurability of the lexicographic maximum of F with $X = R^n$ (generalized by Leese [LE3, Thm. 16.15]). Debreu [DE, (5.10)] and Castaing [CA 4, 5, Chap. 6] gave early results relating measurability of F to scalar measurability of F. The following selection theorem was given by Valadier (for earlier versions see [VA1, 2]).

THEOREM 8.1 [VA3, Props. 7, 8]. *Suppose X is locally convex Hausdorff, F is compact-convex-valued and scalarly measurable, and either* (i) *X is separated by a countable subset of X' or* (ii) *$F(t) \subset g(t)Q$ for $t \in T$, for some convex compact metrizable $Q \subset X$ and measurable $g: T \to R$. Then F has a Castaing representation consisting of scalarly measurable selections.*

Castaing [CA15, 16, Thm. 2] obtained this conclusion assuming instead of (i) or (ii) that μ is a complete probability measure, X is a Lusin space and each $F(t)$ is weakly locally compact and line-free. Benemara [BN1, Lem. 2] also obtained this conclusion, collateral to characterizing extreme scalarly measurable selections of F. Castaing [CA17, 18] treats a scalarly measurable convex-compact-valued F, parameterized on [0, 1] in an absolutely continuous manner, and he obtains parameterized well-behaved selections. Additional results on existence of scalarly measurable selections have been given by Ekelund and Valadier [EV] (see § 10) and Valadier [VA6] (see § 16). In [CA20] Castaing shows that the set of scalarly measurable selections (identified under a.e. equality) is nonempty and compact, when the support functions of F belong to a Köthe space and X is Suslin, among other assumptions.

Suppose $X = R^n$ and h is a selection of co F, where co $F(t)$ is the convex hull of $F(t)$ for $t \in T$. Then by Carathéodory's theorem, for $t \in T$, there exist $\lambda_0(t), \cdots, \lambda_n(t) \geqq 0$, and $g_0(t), \cdots, g_n(t) \in F(t)$ such that $\sum_{i=0}^{n} \lambda_i(t) = 1$ and $h(t) = \sum_{i=0}^{n} \lambda_i(t) g_i(t)$. If such λ_i's and g_i's can be chosen a.e. as measurable functions, we say h has a *measurable Carathéodory representation*. Existence of such a representation is a key to proving various versions of the LaSalle bang–bang principle of

control theory. We have stated the desired choice of λ_i's and g_i's as a measurable selection problem. It is solved, of course, by applying more general selection theorems. Consider the following theorem given by Wagner. Under (ii) it is essentially [CA6, Thm. 3]; under (i) or (iii) one picks a natural $G\colon T \to \mathcal{P}(R^{(n+1)^2})$ somewhat as in [AU1, Thm. 3] and [CA5, Thm. 7.1], proves G is of Suslin type by remarks in § 6, and applies Theorem 6.1. Theorem 4.2(g) affords alternative hypotheses equivalent to (i). Still earlier versions were given by Sonneborn and Van Vleck [SV], who applied Aronszajn's generalization of Filippov's lemma, and in [CA3]. A related result for constant F is given as [HJ, Thm. III.16.14], credited to Hermes [HE1].

THEOREM 8.2 [WG1, Lem. 2.5(a)]. *Suppose μ is a σ-finite outer measure, $X = R^n$, $h \in \mathcal{S}(\text{co } F)$, and either* (i) *$F$ is measurable and closed-valued; or* (ii) co F *is measurable and compact-valued; or* (iii) $\text{Gr } F \in \mathcal{M} \otimes \mathcal{B}(X)$. *Then h has a measurable Carathéodory representation.*

In [CA22, Thms. 1, 2], Castaing obtains Carathéodory map selections of a suitably parametrized closed-convex-valued function into a separable Banach space or the weak dual of such.

In discussing Theorem 4.2, we have noted Rockafellar's [RC1–6] use of measurable convex-valued functions in the form of epigraph functions associated with convex normal integrands. In this work, explicit results on existence of measurable selections are mainly those referenced in Theorem 4.2 and its proof; however, in additional various ways he uses the equivalence (iii)$\Leftrightarrow$(ix) in Theorem 4.2(e) to obtain and apply Castaing representations. (See addendum (v).)

Let $\ddot{F}(t)$ be the set of extreme points of $F(t)$ (the profile of $F(t)$) for $t \in T$. Himmelberg and Van Vleck have treated measurability properties of $\ddot{F}$ in [HV5]. Their Theorem 4(a) is a finite-dimensional version of the first of the following two theorems of Leese, who notes that their methods may be used to prove it. The Suslin type conclusion of Theorem 8.3 affords a generalization of [HV5, Thm. 3], which includes implicit function results (note that $\ddot{F}$ need not be closed-valued).

THEOREM 8.3 [LE3, Thm. 16.10]. *Suppose X is a separable metrizable topological vector space and F is measurable and compact-convex-valued. Then,* $\text{Gr } \ddot{F} \in \mathcal{M} \otimes \mathcal{B}(X)$. *Hence if also $\mathcal{M}$ is a Suslin family and X is a Suslin space, then $\ddot{F}$ is of Suslin type.*

THEOREM 8.4 [LE3, Thms. 16.13, 16.16, 16.18]. *Suppose X is a Hausdorff locally convex real vector space, F is measurable and convex-valued, and one of the following holds:*

(i) *$\mathcal{M}$ is a Suslin family, X is Suslin, and F is compact-valued;*

(ii) *X is separated by a countable subset of X' and F is compact-valued; or*

(iii) *X is separable metric and F is weakly-compact-valued.*

Then there exist $f_1, f_2, \cdots \in \mathcal{S}(\ddot{F})$ such that for $t \in T$, $F(t)$ is the closed convex hull of $\{f_1(t), f_2(t), \cdots\}$. Hence under (i) *or* (iii), *F has a Castaing representation.*

Leese has given the following two theorems in [LE1, 6]. In [LE6] the σ-algebra $\mathcal{M}$ is replaced by more general structures (see § 13). Related results on conjugate Banach spaces and some unsolved problems are also given in [LE6, § 4]. Under the hypothesis of Theorem 8.5, each compact convex set has a unique element closest to the origin [LE3, p. 54], and under the hypothesis of Theorem 8.6 this is true for closed convex sets [LE3, Lem. 9.5].

THEOREM 8.5 [LE1, Thm. 1; LE6, Thm. 2.3]. *Suppose X has a strictly convex norm and F is compact-convex-valued and weakly measurable. Then $\mathcal{S}(F) \neq \emptyset$.*

THEOREM 8.6 [LE1, Thm. 2; LE6, Thm. 3.3]. *Suppose X is a Banach space and has a uniform norm $\|\cdot\|$ (i.e., $\|x_n\| \leqq 1$, $\|y_n\| \leqq 1$, and $\|x_n + y_n\| \to 2$ implies $\|x_n - y_n\| \to 0$), and F is closed-convex-valued and weakly measurable. Then $\mathcal{S}(F) \neq \emptyset$.*

Cole [CL1, 2] has shown that if X is a separable reflexive Banach space, and F is convex-closed-bounded-valued on $T = [0, 1]$ and obeys a condition like Cesari's Q (e.g., [CE]), then F has a strongly measurable selection (pointwise limit of simple functions), and the set of such selections is weakly compact in itself. An earlier result of Himmelberg, Jacobs, and Van Vleck [HJV, Thm. 4] has some hypotheses in common with [CL1]. In [CA7, Cor. 4], Castaing obtains a Lusin measurable selection of F (see § 14), without separability of X.

9. Pointwise optimal measurable selections. Here we consider the existence of a measurable selection of F such that a real-valued function on $\operatorname{Gr} F$ is maximized pointwise: We suppose $u: \operatorname{Gr} F \to R$ is $\mathcal{M} \otimes \mathcal{B}(X)$ measurable and $u(t, \cdot)$ is usc on $F(t)$ for $t \in T$, and we let $v(t) = \sup\{u(t, x): x \in F(t)\}$ for $t \in T$. Our concern is whether there exists $f \in \mathcal{S}(F)$ such that $u(\cdot, f(\cdot)) = v$, and to this end, whether v is measurable. These results are sometimes called Dubins–Savage type theorems, after [DS, Lem. 6] (1965). (See addendum (vii).)

The strongest result to date appears to be the following, which combines Leese's [LE3, Prop. 14.8], [HPV, Thm. 2] of Himmelberg, Parthasarathy, and Van Vleck, and Schäl [SC1, Thm. 2; SC2, Prop. 9.4 and Thm. 12.1].

THEOREM 9.1. *Suppose F is compact-valued, and either*

(i) *$\mathcal{M}$ is a Suslin family, X is Hausdorff and F is of Suslin type; or*

(ii) *T and X are Borel subsets of Polish spaces, $\mathcal{M} = \mathcal{B}(T)$, and F is measurable; or*

(iii) *X is separable metric, F is measurable, and u is the limit of a decreasing sequence of Carathéodory maps.*

Then v is measurable and there exists $f \in \mathcal{S}(F)$ such that $u(\cdot, f(\cdot)) = v$.

Under (i), this is proved in [LE3] by showing, without assuming that F is compact-valued or that each $u(t, \cdot)$ is usc, that $G|S$ is of Suslin type, where $G(t) = F(t) \cap \{x: u(t, x) = v(t)\}$ for $t \in T$ and $S = T \cap \{t: G(t) \neq \emptyset\}$. Under (ii), it is proved in [HPV] via the "Kunugui–Novikov" theorem. Under (iii) one puts together the cited statements of Schäl (brought to our attention by Robert Kertz).

Various facts related to the condition on u given in Theorem 9.1 (iii) are given in [SC2, § 11]. If this condition were implied by the hypothesis of Theorem 9.1 (ii) (which includes that u is $\mathcal{M} \otimes \mathcal{B}(X)$ measurable and each $u(t, \cdot)$ is usc), then 9.1 (ii) would follow from 9.1 (iii); this appears to be an open question.

Castaing [CA17, Lem.] gave a version of Theorem 9.1 (i) with X a Lusin space and μ complete. Furukawa [FU, Thm. 4.1] obtained Theorem 9.1 (ii) with the added assumptions that X is compact, $X \subset R^n$, and u is a bounded Carathéodory map. Darst [DR, Thm. 1] obtained a Borel selection as in (ii), assuming X is compact metric, T is Polish, and u (and not just each $u(t, \cdot)$) is usc. Dubins and Savage made the stronger assumption that F is usc, as did Maitra [MT], and Hinderer [HD1, 2] in separate generalizations of [DS]. Debreu [DE, (4.5)] (1965) obtained measurability of v and G mentioned above.

Brown and Purves have given a related result when F is σ-compact-valued.

THEOREM 9.2 [BP, Cor. 1]. *Suppose T is a Borel subset of a Polish space, $\mathcal{M} = \mathcal{B}(T)$, X is Polish, F is σ-compact-valued, $\operatorname{Gr} F$ is Borel, $I = \{t$: for some $x \in F(t)$, $u(t, x) = v(t)\}$, and $\varepsilon > 0$. Then I is a Borel set and there exists $f \in \mathcal{S}(F)$ such that*

$$u(t, f(t)) = v(t) \quad \text{when } t \in I,$$

$$u(t, f(t)) \geqq v(t) - \varepsilon \quad \text{when } t \notin I \text{ and } v(t) \neq \infty,$$

$$u(t, f(t)) \geqq \frac{1}{\varepsilon} \quad \text{when } t \notin I \text{ and } v(t) = \infty.$$

The problem of finding $f \in \mathcal{S}(F)$ such that $u(\cdot, f(\cdot)) \geqq v(\cdot) - \varepsilon(\cdot)$ is treated by Schäl [SC1], Strauch [ST], and Furukawa [FU], for example.

10. Decomposition of Gr F into measurable selections. For the problem of decomposing $\operatorname{Gr} F$ into measurable selections, we cite principally a 1930 result of Lusin [LS] on countably-valued F, from the early beginnings of measurable selection theory, and a theorem from Wesley's thesis [WE1] which is probably the most profound result to date in measurable selections.

THEOREM 10.1 [LS, p. 244]. *Suppose $T = R^m$, $X = R$, $F(t)$ is countable for $t \in T$, and $\operatorname{Gr} F$ is Borel. Then there exists a Borel map $f_i: T \to X$ for $i = 1, 2, \cdots$, such that $\operatorname{Gr} F \subset \bigcup_{i=1}^{\infty} f_i$ and for $i, j = 1, 2, \cdots$, we have $f_i(t) < f_j(t)$ for $t \in T$ or $f_j(t) < f_i(t)$ for $t \in T$.*

COROLLARY 10.2. *Under the hypothesis of Theorem 10.1 with $\mathcal{M} = \mathcal{B}(T)$, there exist $g_1, g_2, \cdots \in \mathcal{S}(F)$ such that $\operatorname{Gr} F = \bigcup_{i=1}^{\infty} g_i$. If each $F(t)$ is infinite, the g_i's may be taken to be distinct.*

Wesley [WE1, Thm. 1] obtained a version of Corollary 10.2 (wherein the selections are Lebesgue measurable), having belatedly learned of Lusin's results as indicated by his footnote. We conjecture, but have not verified, that Corollary 10.2 holds for arbitrary $(T, \mathcal{M})$ and separable metrix X. Himmelberg has shown this when F is finite-valued [HM2, Thm. 5.4].

For certain F having uncountable values, Wesley has given a nice partitioning of $\operatorname{Gr} F$ into measurable selections, as stated next. In [WE2] he has applied his methods to mathematical economics, i.e., to showing existence of a well-behaved representation of continuous preference orders parameterized in Borel fashion over $2^{\aleph_0}$ traders; he avoids connectedness assumptions made by Aumann [AU3].

THEOREM 10.3 [WE1, Thm. 2]. *Suppose T and X are Lusin spaces, μ is the completion of a σ-finite measure on $\mathcal{B}(T)$, $\mu(T) > 0$, $\operatorname{Gr} F$ is Borel, and $F(t)$ is uncountable for $t \in T$. Let $\mathcal{L}$ be the σ-algebra of Lebesgue measurable subsets of $[0, 1]$. Then there exists $h: T \times [0, 1] \to \operatorname{Gr} F$ such that:*

(a) *for $t \in T$, $h(t, \cdot)$ is a one-to-one Borel function on $[0, 1]$ onto $F(t)$;*

(b) *for $y \in [0, 1]$, $h(\cdot, y) \in \mathcal{S}(F)$;*

(c) *h is an $\mathcal{M} \otimes \mathcal{L}$ measurable function.*

Wesley's statement of Theorem 10.3 has $T = X = [0, 1]$ and $\mathcal{M} = \mathcal{L}$. The generalization to Lusin spaces is straightforward, as pointed out to us by Aumann, by taking isomorphisms between the measurable spaces $([0, 1], \mathcal{L})$ and $(T, \mathcal{M})$ via, e.g., [AS, Lem. 6.2], and between $([0, 1], \mathcal{B}([0, 1]))$ and $(X, \mathcal{B}(X))$ (one also needs $\mathcal{B}(T \times X) = \mathcal{B}(T) \otimes \mathcal{B}(X)$, e.g., via [HJ, Props. I.6.A.4 and I.5.B.7]).

Wesley's proofs of [WE1, Thms. 1, 2] are based on the Cohen forcing methods of mathematical logic, which, without implying doubt, we do not understand. He recommends (personal communication) explanations in [WE2] for better understanding of [WE1]. He also states that by modifying his proof, conclusions (b) and (c) may be strengthened to assert universal measurability, i.e., measurability with respect to any σ-finite complete measure whose set of measurable sets includes the Borel sets. (See addendum (viii).)

It would be desirable to prove Theorem 10.3 without the use of metamathematics. This problem appears to be quite difficult. Wesley poses the problem of proving his result without the Zermelo–Frankel replacement axiom.

Ekeland and Valadier [EV] have given decomposition results in the vein of this section, in the form of representing a compact-convex-valued function which is a Carathéodory map (see § 2). The following is taken from their Corollary 5 and Theorem 2.

THEOREM 10.4. *Let X be a compact metrizable subset of a locally convex topological vector space, Z be a topological space, and $G: T \times Z \to \mathcal{P}(X)$ be compact-convex-valued and a Carathéodory map (with respect to the Hausdorff metric on the set of compact subsets of X). Then there exists a Carathéodory map $f: T \times (Z \times X) \to X$ such that*

$$G(t, z) = \{f(t, z, x): x \in X\} \quad \text{for } t \in T, \quad z \in Z,$$

and such that if $g: T \to Z$ is strongly measurable, and $h: T \to X$ is scalarly measurable and with $h(t) \in G(t, g(t))$ for $t \in T$, then there exists a measurable $u: T \to X$ for which

$$h(t) = f(t, g(t), u(t)) \quad \text{for } t \in T.$$

Of course, the decomposition of Gr G provided by f in this theorem need not be a partitioning of Gr G, i.e., we might have $f(\cdot, \cdot, x)$ and $f(\cdot, \cdot, x')$ overlapping and unequal. Included here is a measurable implicit function result. In [EV], these results are given for G more general than being a Carathéodory map.

Larman's result [LA1, 2] noted in § 12 below provides an uncountable disjoint family of selections of F which are Borel sets—it is not asserted that these exhaust Gr F.

11. Selections of partitions. In this section we suppose that $\mathcal{Q}$ is a partition of T. A *selection* of $\mathcal{Q}$ is a set $S \subset T$ such that $S \cap E$ is singletonic whenever $E \in \mathcal{Q}$. Here we let $T = X$ and F be given by the requirement that $t \in F(t) \in \mathcal{Q}$ for $t \in T$. We see that a selection of $\mathcal{Q}$ is the range of a selection f of F; however, f must also be constant on each $F(t)$. Note that the members of $\mathcal{Q}$ are closed iff F is closed-valued, and that this situation is associated naturally with the inverse of a continuous map. Also, to any $G: T \to \mathcal{P}(X)$ (without $T = X$) corresponds a natural partition of Gr G, viz., $\{\{t\} \times G(t): t \in T\}$, so that the results of this section are also relevant to the next section on uniformization. We let $\mathcal{L}$ be a family of subsets of T.

Early results on Borel selections of partitions were obtained by Mackey [MC1] in 1952 (Theorem 11.6 below) and Dixmier [DI] in 1962 (Corollary 11.2(ii) below—see also remarks in § 3 and following 4.1 and 11.6).

We begin with 1970 results of Hoffman-Jørgensen [HJ]. Although the hypothesis of the following rather general theorem seems complicated, all selection results of [HJ] cited in this survey are derived from it. Conditions somewhat similar to those of Theorem 11.1 are given at the end of [LE5, § 3], in a selection statement (not referring to partitions per se).

THEOREM 11.1 [HJ, Thm. II.6.1; or CH, Thm. 4.1]. *Suppose $\mathscr{L} \supset \mathscr{M}$, $\mathscr{L}$ is closed under countable union and countable intersection, and there exists $A: \mathscr{V}^* \to \mathscr{M}$ (see § 2) such that:*

(i) $T = \bigcup_{n=1}^{\infty} A_{(n)}$;

(ii) $A_{(\sigma_1;\cdots,\sigma_k)} = \bigcup_{n=1}^{\infty} A_{(\sigma_1,\cdots,\sigma_k,n)}$ *for* $\sigma \in \mathscr{V}$, $k = 1, 2, \cdots$;

(iii) *for* $\sigma \in \mathscr{V}$ *and* $t \in T$, *letting* $D_k = F(t) \cap A_{\sigma|k}$ *for* $k = 1, 2, \cdots$, *we have* $\bigcap_{k=1}^{\infty} D_k$ *is singletonic or for some* k, $D_k = \varnothing$;

(iv) $F^-(A_{\sigma|k}) \in \mathscr{L}$ *for* $\sigma \in \mathscr{V}$, $k = 1, 2, \cdots$.

Then $\mathscr{Q}$ has a selection S such that $T \backslash S \in \mathscr{L}$.

The following two corollaries are given in [HJ]; Corollary 11.2 (ii) was previously given by Dixmier [DI] (cited in [HJ]). Dixmier applied his results to show the Borel nature of equivalence classes of factorial representations of a separable involutive Banach algebra and to give a converse of a result of Mackey [MC2].

COROLLARY 11.2 [HJ, Thms. III. 8.3–8.6]. *If T is topologized, $\mathscr{M} = \mathscr{B}(t)$, and F is closed-valued, then $\mathscr{Q}$ has a selection $S \in \mathscr{M}$ providing one of the following holds:*

(i) *distinct points of T are separable by continuous functions into* [0, 1], *T is Suslin, and F is weakly measurable and compact-valued;*

(ii) *T is Polish and F is weakly measurable;*

(iii) *F is measurable, T is a countable union of closed Polish subspaces, and $\bigcup_{S \in \mathscr{Q}} (S \times S) \in \mathscr{B}(T \times T)$; or*

(iv) *T is a Lusin space and $F^-(B) \in \mathscr{B}(T)$ for $B \in \mathscr{B}(T)$.*

COROLLARY 11.3 [HJ, Thm. III. 8.7]. *Suppose F is closed-valued, T is Suslin, $\mathscr{L}$ is closed under countable union and countable intersection, and $\mathscr{L}$ contains A and $F^-(A)$ whenever $A \subset T$ is Suslin. Then $\mathscr{Q}$ has a selection S such that $T \backslash S \in \mathscr{L}$.*

Christensen has further applied Theorem 11.1 to the Effros σ-algebra over the set of closed subsets of T, when T is metric Suslin [CH, Thm. 4.2]. As noted in § 4, Theorem 11.2 (ii) constitutes a special case of Theorem 4.1 above, with $\mathscr{M} = \mathscr{B}(T)$.

Turning next to work of Kuratowski, Maitra, and Rao, following [KMT] we say $\mathscr{Q}$ is an $\mathscr{L}^-$ partition {an $\mathscr{L}^+$ partition} of T if $F^-(A) \in \mathscr{L}$ for each open $A \subset T$ {$T \backslash F^-(A) \in \mathscr{L}$ for each closed $A \subset T$}, with F as above. Kuratowski and Maitra have given the following.

THEOREM 11.4 [KMT, § 3]. *Suppose $\mathscr{L}$ is a Boolean algebra (i.e., field), T is Polish, the open sets of T belong to $\mathscr{L}_\sigma$ (see § 2), each member of $\mathscr{Q}$ is closed, and $\mathscr{Q}$ is an $\mathscr{L}^+$ or $\mathscr{L}^-$ partition. Then there is a selection S of $\mathscr{Q}$ such that $T \backslash S \in \mathscr{L}_\sigma$.*

One application of this in [KMT] is to find a Borel set selection of $\mathscr{Q}$ which intersects each member of an analytic set of compact sets of T.

Special cases of Theorem 11.4, when $\mathscr{L}$ is a σ-algebra, have been given in [KU4, Thm. B] and [KU5, Thm. 7.1].

Maitra and Rao [MR2] have taken a different approach, utilizing a linear order on T induced by a continuous open map from the irrationals with lexicographic ordering. Their main result follows.

THEOREM 11.5 [MR2, Thm. 4.1]. *Suppose T is Polish, $\mathscr{L}$ is a σ-lattice containing the closed subsets of T, each member of $\mathscr{Q}$ is closed, and $\mathscr{Q}$ is an $\mathscr{L}^-$ partition. Then there is a selection S of $\mathscr{Q}$ such that $T\backslash S \in \mathscr{L}$.*

Both [KMT] and [MR2] apply their results to the case where $\mathscr{L}$ is the σ-additive lattice of subsets of T of additive class α, with [MR2] having stronger results. Also given in [MR2] are several examples showing that the latter results cannot be improved in certain ways. They cite a 1927 antecedent by Mazurkiewicz [MK].

We conclude this section with results on topological groups.

THEOREM 11.6. *Suppose T is a locally compact topological group. H is a metrizable closed subgroup of T, and $\mathscr{Q} = \{Ht: t \in T\}$. Then $\mathscr{Q}$ has a Borel set selection.*

This result was obtained by Mackey [MC1, Lem. 1.1] in 1952 with metrizability of H strengthened to separability (the latter implies the former in this context), using in the proof [FM, Thm. 5.1] of Federer and Morse. It was obtained as given here by Feldman and Greenleaf [FG, Thm. 1]. Weaker versions were given as [HJ, Thm. III. 16.6] and earlier as [DI, Lem. 3].

The selection of $\mathscr{Q}$ obtained in Theorem 11.6 determines a selection f of p^{-1} where $p: T \to T/H$ is canonical; in [FG] it is added that $f^{-1}(C)$ is in the σ-algebra generated by the compact sets of T/M for compact $C \subset T$, and that if T has an open subgroup $U \supset H$ such that $p(U)$ is σ-compact, then f may be obtained to be measurable w.r.t. $\mathscr{B}(T/H)$.

Greenleaf has applied Theorem 11.6 in [GR] to prove that a closed subgroup of an amenable group is amenable (a locally compact group G is called amenable if there is a left invariant positive linear functional M on $L_\infty(G)$ such that $M(h) = 1$ if $h(g) = 1$ for $g \in G$).

In 1965, Baker [BA, Thm. 2] and Effros [EF, Thm. 2.9] independently showed that several conditions previously shown to be equivalent by Glimm [GL] were also equivalent to the existence of a Borel set selection of the partitioning of an "almost Hausdorff" space M (see Theorem 12.4 below) into orbits of a locally compact Hausdorff group G of transformations acting continuously on M, when G and M each have countable base; one of these conditions is merely that M/G is a T_0 space. This has recently been applied by Bondar [BR].

One cannot omit "H is closed" in Theorem 11.6 [HJ, p. 177]: Let T be the additive reals and H be the rationals. Then $\mathscr{Q}$ has no Lebesgue measurable selection (the selections of $\mathscr{Q}$ are the examples usually given of non-Lebesgue-measurable sets). This has been generalized by Kuratowski [KU6]. The following remarkable converse has been pointed out by Bondar (who brought Theorem 11.6 to our attention) as a consequence of [BA, Thm. 2] and [MC2, Thm. 7.2]: If T is Polish, and $\mathscr{Q}$ has a Borel set selection, then H is closed.

12. Uniformization. The term "uniformization" is a synonym for "selection." One usually refers to uniformizations of $\operatorname{Gr} F$ rather than of F, and with interest in properties of a selection as a subset of product space (such as being a Borel set) rather than properties of mappings (such as being a Borel function). It dates from the era of [LS] and [NO1], as noted in § 3, or perhaps earlier.

An early result is the following, proved independently by Lusin [LS2] and Sierpinski [SP1]. If $T = X = R$ and Gr F is Borel in R^2, then F has a selection (uniformization) f such that $(T \times X) \setminus f$ is Suslin, that is, f is a complementary Suslin (i.e., CA) subset of R^2. This was improved by Kondō [KN] in permitting Gr F to be complementary Suslin and in other ways (see Sampei [SM] or Suzuki [SZ] for a later proof). Kondō's results were further generalized by Rogers and Willmott [RW], [WI]. Related results are given by Kuratowski [KU6]. A variation is claimed by Hoffman-Jørgensen [HJ, Thm. III.9.5] under Suslin T, X, and Gr F; Leese finds the supporting argument incomplete. Jankov [JN] has shown that a Suslin subset of R^2 has a uniformization which is in the σ-algebra generated by the Suslin sets of R^2.

Results on G_δ uniformizations of F have been given by Braun [BR, Thm. 1] (she also showed that a closed subset of R^2 need not have an F_σ uniformization), Engelking [EN], and Michael [MI].

Larman's main theorem of [LA1, 2] yields an uncountable disjoint family of Borel set uniformizations of F, requiring that each $F(t)$ be an uncountable σ-compact G_δ, among other conditions. Brown and Purves [BP] show that if X and T are Polish, Gr F is Borel, F is σ-compact-valued, and $\mathcal{M} = \mathcal{B}(T)$, then there exists $f \in \mathcal{S}(F)$ (this much follows from Sion [SN]) such that f is a Borel subset of $T \times X$; they thereby generalize a result of Stschegolkow, given in [AL]. A similar result with different conditions on the values of F has been given by Sarbadhikari [SR].

To relate measurable selection results to uniformization results, one wishes to know when certain properties of $f\colon T \to X$ as a subset of $T \times X$ imply that f is a measurable function, and conversely. Following are some facts of this kind. See also [LE5, Appendix to § 6].

THEOREM 12.1 (Hoffman-Jørgensen [HJ, pp. 8–9]). *Suppose $\mathcal{N}$ is a σ-algebra over X (X not necessarily topologized), some countably generated sub-σ-algebra of $\mathcal{N}$ separates X (equivalently, $\{(x, x)\colon x \in X\} \in \mathcal{N} \otimes \mathcal{N}$), $f\colon T \to X$, and $f^{-1}(A) \in \mathcal{M}$ for $A \in \mathcal{N}$. Then $f \in \mathcal{M} \otimes \mathcal{N}$.*

This very general statement implies, in particular, [KU5, § 2, Thm. 8]. It is also given, essentially, as [SB3, Prop. 2].

THEOREM 12.2 (Leese—personal communication). *Suppose $\mathcal{M}$ is a Suslin family, X is weakly Suslin, $f\colon T \to X$, and $f \in \mathcal{M} \otimes \mathcal{B}(X)$. Then f is a measurable function.*

Proof. Note (iv) at the end of § 6. □

THEOREM 12.3 (Leese—personal communication). *Suppose T is topologized, $\mathcal{M}$ contains the Suslin family generated by the closed sets of T, X is analytic in the sense that there exist a Polish space P and a compact-valued u.s.c. $G\colon P \to \mathcal{P}(X)$ such that $X = G(P)$, $f\colon T \to X$, and f is in the Suslin family generated by the closed sets of $T \times X$. Then f is a measurable function.*

Proof. Apply [LE5, Thm. 8.2], originally due to Rogers and Willmott. □

THEOREM 12.4 (Baker [BA, Lem. 4]). *Suppose T is topologized and T and X each have a countable base and are almost Hausdorff, i.e., are locally compact T_0 spaces with every nonvoid locally compact subspace containing a nonvoid relatively open Hausdorff subspace. Suppose $B \in \mathcal{B}(T)$, $f\colon B \to X$, and $f \in \mathcal{B}(T \times X)$. Then f is a Borel function.*

THEOREM 12.5 (Hoffman-Jørgensen [HJ, Thm. III. 4.1]). *Suppose T and X are Suslin and $f: T \to X$. Then f is a Borel function iff f is a Borel subset of $T \times X$ iff f is a Suslin subset of $T \times X$.*

THEOREM 12.6 (Lehn [LN2]). *Suppose $(T, \mathcal{M}, \mu)$ is the completion of the measure space $(T, \mathcal{M}_0, \nu)$, $(T, \mathcal{M}_0)$ and $(X, \mathcal{N})$ are countably separated Blackwell spaces (see* [HJ]*), $f: T \to X$, and $f \in \mathcal{M} \otimes \mathcal{N}$. Then $f^{-1}(A) \in \mathcal{M}$ for $A \in \mathcal{N}$.*

Valadier [VA4, Cor.] relates scalar measurability of $f: T \to X$ (with X locally convex) to $f \in \mathcal{M} \otimes \mathcal{B}(X)$.

In [HJ, III. 16.3, 5] examples are given where (a) X is R^2, T is topologized but not Suslin, $g: X \to T$ is continuous and bijective, g^{-1} is a Borel subset of $T \times X$, and g^{-1} is not a Borel function; and (b) $f: R \to R$ is not a Lebesgue measurable function but f is a complementary Suslin subset of $R \times R$ and hence a Lebesgue measurable set ((b) assumes axiom of constructibility—see also [AU3]).

13. Measurability with other structures. In this section we replace the role of $\mathcal{M}$ with a family $\mathcal{L}$ of subsets of T not necessarily a σ-algebra and we define $\mathcal{Q}$ as a similar family of subsets of X. Our interest is in selections f of F which are $(\mathcal{L}, \mathcal{Q})$ measurable in the sense that $f^{-1}(A) \in \mathcal{L}$ for $A \in \mathcal{Q}$. Measurability of F is defined similarly. No role is played by μ in this section.

Let $\mathcal{K}$ and $\mathcal{G}$ be the respective families of closed and open subsets of X, and, when T is topologized, let $\mathcal{F}$ be the family of closed subsets of T.

The most important case where $\mathcal{L}$ is not a σ-algebra is when $\mathcal{L}$ and $\mathcal{Q}$ are both topologies. This is the subject of continuous selections, i.e., $(\mathcal{F}, \mathcal{K})$ measurable selections in the above terminology. This topic has extensive literature which is essentially topological, rather than measure-theoretical, in character, and which we do not review here. We merely cite three general references, [MI1], [FL], and the first half of [PR1], where numerous additional references may be found.

We have noted that Kuratowski and Ryll-Nardzewski [KRN] have shown that Theorem 4.1 holds with $\mathcal{M} = \mathcal{L}_\sigma$, where $\mathcal{L}$ is a Boolean algebra. This generality enables them to obtain selections which are continuous, continuous modulo first category sets, or of additive class α (i.e., $f^{-1}(U)$ is Borel of additive class α for open $U \subset X$). Leese [LE5, Thm. 3.2] has sharpened this slightly: Let $\mathcal{L}$ be closed under finite union and intersection,[4] $\varnothing \in \mathcal{L}$, and $\mathcal{D} = \{A \setminus B: A, B \in \mathcal{L}\}$; then if X is Polish and F is closed-valued and $(\mathcal{L}_\sigma, \mathcal{G})$ measurable, there exists a $(\mathcal{D}_\sigma, \mathcal{G})$ measurable selection of F. In fact several of Leese's results given above have been stated by him in this kind of generality, generalizing the σ-algebra $\mathcal{M}$ differently in the hypothesis and the conclusion, i.e., Theorems 4.10 [LE5, Thms. 4.1 and 4.2], 5.13 [LE5, Thm. 5.5], 5.14 [LE5, Thm. 6.2 or 6.3], 8.5 [LE6, Thm. 2.3], and 8.6 [LE6, Thm. 3.3]. Here is a companion result (when $\mathcal{L}$ is a σ-algebra and a Suslin family this is included in Theorem 5.13).

THEOREM 13.1 [LE5, Thm. 5.2]. *Suppose $\varnothing \in \mathcal{L}$, X is Polish, $\mathcal{R} = \{S \times K: S \in \mathcal{L}, K \in \mathcal{K}\}$, $\operatorname{Gr} F$ is in the Suslin family generated by $\mathcal{R}$, $\mathcal{A} = \{A: A \subset T$ is in the Suslin family generated by $\mathcal{L}\}$, and $\mathcal{D} = \{A \setminus A': A, A' \in \mathcal{A}\}$. Then there exists a selection f of F such that $f^{-1}(U) \in \mathcal{D}_\sigma$ for $U \in \mathcal{G}$.*

[4] Leese has pointed out that [LE5, Thm. 3.2] should include the requirement that $\mathcal{L}$ be closed under finite intersection.

Engelking [EN, Thm. 1] obtains an $(\mathcal{F}_\sigma, \mathcal{G})$ measurable selection of F; here T is compact and perfectly normal, X is metrized, and F is usc and complete-separable-valued. If "usc" is replaced by "lsc," then "separability" may be omitted, as proved independently by Čoban. Čoban [CB1, 2] gives numerous theorems on $(\mathcal{F}_\sigma, \mathcal{G})$ measurable selections, and related selection results. In [CB3] he gives a variation on Theorem 4.1 with $F^-(U)$ a kind of complementary Suslin set for open $U \subset X$ and with the selection obtained a kind of Borel function. Rogers and Willmott [RGW, Thm. 20] find a selection f of F such that for open $U \subset X, f^{-1}(U)$ is the T projection of a complementary Suslin subset of $T \times X$; here Gr F is complementary Suslin among other conditions.

Maitra and Rao give the following result.

THEOREM 13.2 [MR2, Thm. 2]. *Suppose $\varnothing \in \mathcal{L}$, $T \in \mathcal{L}$, and $\mathcal{L}$ is closed under countable union and finite intersection. Let $\mathcal{L}' = \{T \setminus D : D \in \mathcal{L}\}$. Then the following are equivalent:*

(a) *Whenever $A, B \in \mathcal{L}$ and $A \cap B = \varnothing$, there exists $D \in \mathcal{L} \cap \mathcal{L}'$ such that $A \subset D$ and $B \subset T \setminus D$ (i.e., $\mathcal{L}'$ satisfies the first principle of separation, equivalently the weak reduction principle).*

(b) *If X is compact metric, then any $(\mathcal{L}, \mathcal{G})$ measurable closed-valued $G: T \to \mathcal{P}(X)$ has an $(\mathcal{L} \cap \mathcal{L}')_\sigma$ measurable selection.*

Their Theorem 1 generalizes this statement to the use of higher ordinals and cardinals in the union closedness condition and in the weak reduction principle and to avoiding compactness in (b), thereby extending Theorem 4.1. From this Theorem 1, [MR2] further deduces, in addition to some known results, Theorem 4.8 above and a selection result (Theorem 6) which assumes that F is a countable union of weakly measurable closed-valued functions and that an "$\aleph_1$ weak reduction principle" holds for the "measurable" sets of T.

Kaniewski and Pol give the following result, which does not assume separability of X. They also present some related examples and pose some unsolved problems.

THEOREM 13.3 [KP, Thm. 2]. *Suppose T is an absolutely analytic* [HN] *and F is compact-valued and $(\mathcal{L}, \mathcal{G})$ measurable, where $\mathcal{L} = \{S : S \subset T$ is a Borel of additive class $\alpha\}$ with $0 < \alpha < \omega_1$. Then there exists an $(\mathcal{L}, \mathcal{G})$ measurable selection of F.*

Whitt [WH] gives conclusions in terms of $(\mathcal{F}_\sigma, \mathcal{G})$ measurable selections and of selections of third Baire class.

14. Lusin measurable set-valued functions and selections. Let us recall Lusin's theorem as given in [FE, § 2.3.4 and § 2.3.6].

THEOREM 14.1. *Suppose μ is an outer measure. Suppose also μ is Borel regular and T is metric $\{\mu$ is Radon and T is locally compact Hausdorff$\}$, X is separable metric, $f: T \to X$ is measurable, $\mu(T) < \infty$, and $\varepsilon > 0$. Then there is a closed $\{$compact$\}$ $C \subset T$ such that $\mu(T \setminus C) < \varepsilon$ and $f|C$ is continuous. If also μ is σ-finite, f is a.e. equal to a Borel function.*

We note three related directions in which Lusin's theorem has been generalized.

First, there are formulations of Lusin's theorem for set-valued maps. Pliś [PL1] (1961) and Castaing [CA1, 2, 4, 5] have given such for compact-valued maps, in which case the Hausdorff metric is a natural tool. Extensions to

closed-valued maps have been given by Jacobs [JC2], Himmelberg, Jacobs and Van Vleck [HJV], and Castaing [CA8]. In [HJV] and [CA8] the restricted maps obtained are semi-continuous; Castaing uses the term "approximately semi-continuous" maps.

Second, one may formulate Lusin type theorems for $g: T \times Y \to X$, where Y is a topological space and g is a Carathéodory map. These are called Scorza-Dragoni theorems, after [SD] (1948), the first result of this type. Van Vleck has pointed out to us that Krasnosel'skii's [KR] (first edition 1956) also gave such a result as Lemma 3.2. Subsequent generalizations have been given by Castaing [CA4, 5, 8], Goodman [GD], and Jacobs [JC1].

Third, we have Scorza-Dragoni type results for set-valued maps. Results of this kind have been given by Jacobs [JC2], Castaing [CA8, 13], Himmelberg, Jacobs, and Van Vleck [HJV], Brunovský [BV], Himmelberg [HM1], and Himmelberg and Van Vleck [HV4, 9], usually having the restricted set-valued map semi-continuous.

For the rest of this section, we assume T is topologized as a Hausdorff space and μ is an outer measure and is σ-finite and Radon. Castaing has defined F to be *Lusin measurable* if for some partition $\{S, C_1, C_2, \cdots\}$ of T, $\mu(S) = 0$ and for $i = 1, 2, \cdots$, C_i is compact and $F|C_i$ is usc. If $f: T \to X$ and $F(t) = \{f(t)\}$ for $t \in T$, Lusin measurability of F coincides with "μ measurability" of f, here called *Lusin measurability* of f, as defined in [BO2, Chap. IV, § 5.1], since F is then usc iff f is continuous. If $f: T \to X$ is Lusin measurable, it is measurable as defined in § 2. If the hypothesis of Theorem 14.1 holds, then f is Lusin measurable.

As Théorème 8.4 of [CA4], Castaing has given the following selection result and a corollary (there not restricted to positive measure).

THEOREM 14.2. *Suppose T is locally compact, X is separated by a sequence of continuous real-valued functions, and F is Lusin measurable and compact-valued. Then $\mathcal{S}(F) \neq \varnothing$.*

Castaing has given results on existence of Lusin measurable selections in [CA7, Cors. 1–4], with X a reflexive Banach space, not necessarily separable. Leese's Theorem 7.2 above uses a Lusin measurability hypothesis. In general, the main usefulness of Lusin measurability seems to be in dealing with nonseparable spaces.

15. Set-valued measures. Loosely speaking, one calls Φ a set-valued measure if X is (at least) an Abelian topological group and $\Phi: \mathcal{M} \to \mathcal{P}(X)$ is suitably countably additive. Central to the approaches that have been taken appear to be the definitions of convergence of an infinite sum in $\mathcal{P}(X)$. Our interest here is in the existence of a selection of such a Φ which is a measure on T.

Set-valued measures appear to have originated with Brooks' work [BK] on a finitely additive function Φ on $\mathcal{M}$ into the set of bounded convex sets of a real Banach space. From this point of departure, Godet-Thobie has developed the subject extensively during 1970–75 in a series of papers [GT1–4] and, with Pham The Lai, [GTP], culminating in her thesis [GT5]. She has X a Frechet space in [GT1], X a Banach space in [GT2], and Φ closed-bounded-convex-valued in both. In [GT3, 4], X is a locally convex Hausdorff real vector space; here a convex-compact-valued $\Phi: \mathcal{M} \to \mathcal{P}(X)$ is called a set-valued measure ("multimes-

ure," in [PB2] "multi-mesure faible") if for each point in the dual of X, the associated support function of Φ is a (not necessarily positive) measure. Apparently, [GT4] supersedes [GT1]. Her results are further generalized and unified in [GT5], where, with substantially more abstraction and embedding, X is an Abelian topological group.

Artstein [AR1] (1972) deals with $X \subset R^n$, and his results seem more accessible. In [AR1], $\Phi: \mathcal{M} \to \mathcal{P}(X)$ is a *set-valued measure* if $\Phi(\cup_{j=1}^{\infty} S_j) = \sum_{j=1}^{\infty} \Phi(S_j)$ whenever $S_1, S_2, \cdots \in \mathcal{M}$ are mutually disjoint; the sum of a sequence of subsets of R^n is the set of absolutely convergent sums of selections of the sequence. His main selection result follows.

THEOREM 15.1 [AR1, Theorem 8.1]. *Suppose* $\mu(T) < \infty$, $X \subset R^n$, Φ *is a set-valued measure with convex values,* $\Phi \ll \mu$, *i.e.,* $\mu(A) = 0$ *implies* $\Phi(A) = \{0\}$, $S \in \mathcal{M}$, *and* $x \in \Phi(S)$. *Then there exists a selection* θ *of* Φ *such that* θ *is a (vector-valued) measure on* $\mathcal{M}$ *and* $\theta(S) = x$.

Neither the convexity condition nor the condition $\Phi \ll \mu$ may be omitted, as shown in [AR1]. However, the conditions on μ and the convexity condition may be replaced by $\Phi(T)$ being bounded [AR1, Theorem 8.3]. The boundary condition $\theta(S) = x$ in this type of selection result originated in [AR1].

Pallu de la Barrière [PB1, Théorème 3] considers a compact-convex-valued Φ with X a reflexive vector space topologized compatibly with its dual; he uses the Hausdorff metric to define the above summation. With no further assumption he obtains the conclusion of Theorem 15.1.

Costé's [CS1, Thm. 1.2] is in the vein of [GT3, 4] with less assumption on X, but with locally compact line-free values of Φ. In [CS1, Thm. 2.1], [CS3, Thms. 1, 3], and [CS7, Thm.], X is a Banach space and results in the vein of Theorem 15.1 are given; Φ is closed-bounded-valued (and convex-valued in [CS7]) and the conclusion is of the form $x \in \text{cl}\{\theta(T): \theta \text{ is a selection measure of } \Phi\}$. A similar conclusion is attained in [CS6, Thm. 2–1] with "σ-additive" $\mathcal{M}$ and Φ. In [CS6, Thm. 1–3], he generalizes [PB1, Thm. 3] to finitely additive Φ and selections of Φ, with $\mathcal{M}$ a Boolean algebra. He further obtains in [CS2, Prop. 1] a Radon selection of a compact-valued Φ with X a complete Hausdorff locally convex space.

Thiam [TH1] requires Φ to have positive values as determined by a fixed cone in X, a vector space. For a minimal extremal point x of $\Phi(T)$, by methods of [PB1], he finds a selection measure θ such that $\theta(T) = x$ and $\theta(A)$ is minimal extremal in $\Phi(A)$ for $A \in \mathcal{M}$. When X is locally convex Hausdorff and Φ is weakly-compact-valued such that $\sup \Phi(A)$ exists for $A \in \mathcal{M}$, applying [CS 6], for $x \in \Phi(T)$ he finds a selection measure θ such that $\theta(T) = x$. In [TH2], he treats an additive function on a clan of subsets of T into a semi-group of subsets of X, assumed locally convex Hausdorff; additive selections are obtained.

In [GT-4, 5, 6], Godel-Thobie considers set-valued transition measures, i.e., set-valued measures measurably parameterized with respect to a second measure space. Selections are found in the form of transition measures analogous to those of Markov processes.

Selection results for set-valued measures are applied in [AR1, § 9], [GT2, 5], [CS1], and [CP1] to obtain Radon–Nikodym type results, extending earlier results of Debreu and Schmeidler [DES]. A counterexample to [AR1, Thm. 9.1] is asserted in [CP2].

16. Special topics. We note a few treatments of existence of measurable selections which do not come directly under our above topic headings.

Theorem 4.1, for example, may be used as follows to find a measurable extension of a measurable $f: S \to X$ where $S \in \mathcal{M}$ (e.g., see Maitra and Rao [MR2, Cor. 6]). For extension results without assuming $S \in \mathcal{M}$, see Himmelberg [HM2, § 8].

THEOREM 16.1. *Suppose $S \in \mathcal{M}$, $f: S \to X$ is measurable and X is a Lusin space. Then there exists a measurable $g: T \to X$ such that $g|S = f$.*

Proof. Let $F(t) = \{f(t)\}$ for $t \in S$ and $F(t) = X$ for $t \in T \setminus S$. Take a Polish space P, a continuous bijective $\varphi: P \to X$, and, by Theorem 4.1, $h \in \mathcal{S}(\varphi^{-1} \circ F)$. Let $g = \varphi \circ h$. □

Garnir and Garnir-Monjoie [GGM; GM] treat $T = R^m$, $X = R^n$, and F such that for some $S \in \mathcal{M}$, $\mu(S) = 0$ and $\mathrm{Gr}[F|(T \setminus S)]$ is Suslin. Measurable selections are readily found from known results.

Maritz' thesis [MZ] gives, first of all, an excellent history of the theory of set-valued functions, with an extensive bibliography. He develops a comprehensive treatment of the subject under F and μ having values in Banach spaces, including generalizations in this context of known selection results.

Nürnberger's thesis [NU] treats $T = X$ and F of the form

$$F(t) = \{x: d(t, A) = d(x, A)\} \quad \text{for } t \in T,$$

where d is a metric on X and $A \subset X$ is fixed. For such F, called a projection, he finds Borel function selections in Theorems 4, 5, 6, and 8.

In [VA6, Lem. 3 and Thm. 2], as a tool to generalizing Strassen's theorem, Valadier finds a "pseudo-selection" of F, i.e., a scalarly measurable (see § 8) $\sigma: T \to X'^*$ such that

$$\langle x', \sigma(t) \rangle \leqq \sup \{\langle x', z \rangle: z \in F(t)\} \quad \text{for a.e. } t \in T,$$

for $x' \in X'$. In fact, existence of σ for which equality holds is shown. Here X is a locally convex Hausdorff vector space, X' is its topological dual, X'^* is the algebraic dual of X', and F is convex-compact-valued with all of its support functions finitely integrable. Theorems 3 and 4 relate pseudo-selections to selections.

Blackwell and Dubins obtain the following result related to Theorem 5.7 (from [BRN]). When $\mathcal{M} = \mathcal{B}(X)$, the selection obtained is trivially the identity map of X, so the interest arises when $\mathcal{M}$ is a coarser σ-algebra than $\mathcal{B}(X)$.

THEOREM 16.2 [BD, Thm. 4]. *Suppose $T = X$, X is a Borel subset of a Polish space, $\mathcal{M} \subset \mathcal{B}(X)$, and there exists $g: X \times \mathcal{B}(X) \to R$ such that $g(x, \cdot)$ is a probability measure on $\mathcal{B}(X)$ for $x \in T$, and $g(\cdot, B)$ is a measurable function for $B \in \mathcal{B}(X)$. Then there exists a measurable $f: X \to X$ such that $f(x) \in S$ whenever $x \in S \in \mathcal{M}$.*

A structure more general than ours has been treated very recently by Delode [DL], using as foundation slightly earlier work (which generalizes on separable metrix X) by Delode, Arino, and Penot [DAP1, 2]. Suppose $p: E \to T$ is surjective, $p^{-1}(t)$ is topologized for $t \in T$ (E as a whole need not be topologized), $\mathcal{E}$ is a σ-algebra on E which induces $\mathcal{B}(p^{-1}(t))$ on $p^{-1}(t)$ for $t \in T$, and $p^{-1}(S) \in \mathcal{E}$ for $S \in \mathcal{M}$. Then $(E, \mathcal{E}, T, \mathcal{M}, p)$ is called a *measurable field of topological spaces*. It is *Suslin* if there exists another such object $(E', \mathcal{E}', T, \mathcal{M}, p')$ such that $p^{-1}(t)$ is a

Suslin space for $t \in T$ and $f: E' \to E$ such that $p' = p \circ f$, $f|p'^{-1}(t)$ is continuous for $t \in T$ and $f^{-1}(B) \in \mathscr{E}'$ for $B \in \mathscr{E}$. This structure specializes to ours by letting $E = \operatorname{Gr} F$ and $p = \pi_T|\operatorname{Gr} F$. (Beyond this specialization [DL] and [DAP2] give examples in various spaces of interest in functional analysis.) In this specialization, a Suslin field (as a subfield of $T \times X$) is the graph of a set-valued function of Suslin type (§ 6). In [DAP1, 2], each $p^{-1}(t)$ is metric (usually separable) and existence of a subset of $\mathscr{S}(p^{-1})$ satisfying certain axioms is assumed. Relevance of [DAP2] and [DL] to Theorem 4.2(g) above is noted following Theorem 4.2.

17. Recommended introductory reading. We briefly outline a recommended sequence of reading for someone who is fairly new to the subject of measurable selections and who would like to acquire at least a moderately general knowledge.

The best starting point is Rockafellar's [RC2]. This has $X = R^n$ and takes one through several important fundamentals in an easily readable way. A comprehensive exposition of closed-valued F: $T \to \mathscr{P}(R^n)$ is given in his forthcoming [RC6, § 1], also easily readable.

We recommend next Himmelberg's [HM2]. This gives the principal fundamental results on measurable selections and related properties of measurable sive exposition of closed-valued F: $T \to \mathscr{P}(R^n)$ is given in his [RC6, § 1], also easily readable.

Recommended next are Kuratowski's and Ryll-Nardzewski's [KRN], whose main theorem and proof have not been greatly improved upon, and the main published portion of Castaing's widely referenced thesis [CA5]. The latter is the first comprehensive treatment of measurable set-valued functions and is still worthy of careful review. We emphasize that is is more easily read if preceded by [RC2] and [HM2]. (A comment in [RC2] to the effect that [CA5] primarily treats compact-valued functions is not applicable to the measurable selection portion of [CA5]). One expects that [CA5] will be superseded by the forthcoming Castaing–Valadier text [CV2]. (See addendum (iii).)

We consider that Leese's [LE2] on set-valued functions of Suslin type has considerable unifying effect and we recommend it next accordingly.

This much should give the reader a rather good general knowledge. A graduation piece for an ambitious reader is Wesley's [WE1] profound proof of his easily stated result, Theorem 10.3 above (see also [WE2]). (See addendum (viii).)

Needless to say, a very considerable amount of excellent work on measurable selections is not included in this short list. A general knowledge afforded by these recommended papers can be substantially illuminated in terms of historical development and of specialization in several directions, as may be surmised from the diversity of topics addressed in this survey. It is hoped that the survey itself will give guidance to such further reading, for which the survey is certainly no substitute.

Acknowledgment. Since this survey is by a user of, rather than a significant contributor to, measurable selection results, we have leaned to an unusual degree on help from others. It is a pleasure to acknowledge first of all our particularly heavy debt to Charles Himmelberg, Fred Van Vleck, Charles Castaing, Stephen Leese, Kasimir Kuratowski, Robert Aumann, R. Tyrrell Rockafellar, and James Bondar. Dr. Leese gave a most useful critique of a recent preliminary version of

this survey, correcting several errors. Each of the above has provided especially helpful comments and source material. We further appreciate similar forms of help from (mentioned alphabetically) Zvi Artstein, David Blackwell, Herbert Federer, Robert Kertz, Dietrich Kölzow, John Oxtoby, Michel Valadier, and Eugene Wesley. Assistance from Roman Pol and Pawel Szeptycki is noted in § 4. Numerous others have provided helpful reprints. Among those acknowledged above are several leaders of schools of activity in measurable selection theory.

BIBLIOGRAPHY

The bibliography is composed of three categories, according to whether the bracketed coding has no prime, a single prime, or a double prime. References in the unprimed category contain one or more results on existence of measurable selections which contributed something new at the time presented. An effort at completeness has been made in this category. The single primed references are papers (not texts) which do not appear to belong in the preceding category but which contain properties of set-valued functions of a measurability nature. Moderate inclusiveness has been attempted here. The double primed category consists of (a) useful texts and (b) papers whose only connection with measurable selections is in applications—no attempt at completeness is made here. Several references in the single and double primed categories are not cited in the text.

[AL] W. J. ARSENIN AND A. A. LYAPUNOV, *Die Theorie der A-Mengen*, Arbeiten zur Deskriptiven Mengenlehre, H. Grell, ed., VEB Deutscher Verlag der Wissenschaften, Berlin, 1955, pp. 35–93.

[AR1] Z. ARTSTEIN, *Set-valued measures*, Trans. Amer. Math. Soc., 165 (1972), pp. 103–125.

[AR2] ———, *Weak convergence of set-valued functions and control*, this Journal, 13 (1975), pp. 865–878.

[AR3]′ ———, *On the calculus of closed set-valued functions*, Indiana Univ. Math. J., 24 (1974), pp. 433–441.

[AR4]″ ———, *On a variational problem*, J. Math. Anal. Appl., 45 (1974), pp. 404–415.

[AR5]″ ———, *A lemma and some bang-bang applications*, Brown Univ., Providence, RI, 1975.

[AV]′ Z. ARTSTEIN AND R. A. VITALE, *A strong law of large numbers for random compact sets*, Ann. Probability, 3 (1975), pp. 879–882.

[AU1] R. J. AUMANN, *Integrals of set-valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1–12.

[AU2]″ ———, *Existence of competitive equilibria in markets with a continuum of traders*, Econometrica, 34 (1966), pp. 1–17.

[AU3] ———, *Measurable utility and the measurable choice theorem*, La Décision Actes Colloq. Internat. du Centre Nat. Recherche Sci., Aix-en-Provence, 1967, Paris, 1969, pp. 15–26.

[AP]″ R. J. AUMANN AND M. PERLES, *A variational problem arising in economics*, J. Math. Anal. Appl., 11 (1965), pp. 488–503.

[AS]″ R. J. AUMANN AND L. S. SHAPLEY, *Values of Non-Atomic Games*, Princeton University Press, Princeton, 1974.

[AG]″ E. A. AZOFF AND F. GILFEATHER, *Measurable choice and the invariant subspace problem*, Bull. Amer. Math. Soc., 80 (1974), pp. 893–895.

[BA] K. A. BAKER, *Borel functions for transformation group orbits*, J. Math. Anal. Appl., 11 (1965), pp. 217–225.

[BJ]′ H. T. BANKS AND M. Q. JACOBS, *A differential calculus for multifunctions*, Ibid., 29 (1970), pp. 246–272.

[BM1] M. BENAMARA, *Sections mesurables extrémales d'une multi-application*, C. R. Acad. Sci. Paris Sér. A, 278 (1974), pp. 1249–1252.

[BM2]′ M. BENAMARA, *Points Extrémaux, Multi-applications et Functionelles Intégrales*, Thèse, Grenoble, 1975.

[BG]″ C. BERGE, *Espaces Topologiques—Fonctions Multivoque*, Dunod, Paris, 1959.

[BI]′ J. M. BISMUT, *Intégrales convexes et probabilités*, J. Math. Anal. Appl., 42 (1973), pp. 639–673.

[BL] D. BLACKWELL, *A Borel set not containing a graph*, Ann. Math. Statist., 39 (1968), pp. 1345–1347.

[BD] D. BLACKWELL AND L. E. DUBINS, *On existence and non-existence of proper, regular, conditional distributions*, Ann. Probability, 3 (1975), pp. 741–752.

[BRN] D. BLACKWELL AND C. RYLL-NARDZEWSKI, *Non-existence of everywhere proper conditional distributions*, Ann. Math. Statist., 34 (1963), pp. 223–225.

[BR] J. V. BONDAR, *Borel cross-sections and maximal invariants*, Ann. Statist., 4 (1976), pp. 866–877.

[BO1]″ N. BOURBAKI, *General Topology*, Part 2, Hermann, Paris, 1966.

[BO2]″ ———, *Intégration*, Hermann, Paris, 1974.

[BN] S. BRAUN, *Sur l'uniformisation des ensembles fermés*, Fund. Math., 28 (1937), pp. 214–218.

[BK]′ J. K. BROOKS, *An integration theory for set-valued measures I, II*, Bull. Soc. Roy. Sci. Liège, 37 (1968), pp. 312–319 and 375–380.

[BP] L. D. BROWN AND R. PURVES, *Measurable selections of extrema*, Ann. Statist., 1 (1973), pp. 902–912.

[BV]′ P. BRUNOVSKÝ, *Scorza-Dragoni's theorem for unbounded set-valued functions and its applications to control problems*, Mat. Casopis Sloven, Akad. Vied., 20 (1970), pp. 205–213.

[BU]′ J. J. BUCKLEY, *Graphs of measurable functions*, Proc. Amer. Math. Soc., 44 (1974), pp. 78–80.

[CA1] C. CASTAING, *Multi-applications mesurables, généralisation du principe de bang–bang*, Proc. Colloq. on Convexity (Copenhagen, 1975), W. Fenchel, ed., Copenhagen University, 1967.

[CA2] ———, *Quelques problèmes de mesurabilité liés à la théorie de la commande*, C. R. Acad. Sci. Paris Sér. A, 262 (1966), pp. 409–411.

[CA3] ———, *Sur une nouvelle extension du théorème de Ljapunov*, Ibid., 264 (1967), pp. 333–336.

[CA4] ———, *Sur les multi-applications mesurables*, Thèses, Caen, 1967.

[CA5] ———, *Sur les multi-applications mesurables*, Rev. Française Inf. Rech. Opéra., 1 (1967), pp. 91–126.

[CA6] ———, *Sur la mesurabilité du profil d'un ensemble convexe compact variant de façon mesurable*, Multilithed article, Université de Perpignan, 1968.

[CA7] ———, *Proximité et mesurabilité. Un théorème de compacité faible*, Colloque sur la Théorie Mathématique du Contrôle Optimal, Brussels, 1969, pp. 25–33.

[CA8] ———, *Sur le graphe d'une multi-application souslinienne (mesurable)*, Secrétariat des Math. de la Faculté des Sciences de Montpellier, Publication No. 55, 1969–1970.

[CA9] ———, *Le théorème de Dunford–Pettis généralisé*, Université de Montpellier, Secrétariat des Mathématiques, Publication No. 43, 1969.

[CA10] ———, *Le théorème de Dunford–Pettis généralisé*, C. R. Acad. Sci. Paris Sér. A, 268 (1969), pp. 327–329.

[CA11]′ ———, *Quelques compléments sur le graphe d'une multi-application mesurable*, Travaux du Séminaire d'Analyse Unilatérale, Faculté des Sciences, Université de Montpellier, 1969.

[CA12] ———, *Quelques applications du théorème de Banach Dieudonné a l'intégration*, Université de Montpellier, Secrétariat des Mathématiques, Publication No. 67, 1969–1970.

[CA13] ———, *Une nouvelle extension du théorème de Scorza-Dragoni*, C. R. Acad. Sci. Paris Sér. A, 271 (1970), pp. 396–398.

[CA14]′ ———, *Application d'un théorème de compacité à la désintégration des mesures*, Travaux de Séminaire d'Analyse Convexe, vol. 1, Exp. No. 12, Secrétariat des Mathematiques, U.E.R. de Math., Univ. Sci. Tech. Languedoc, Montpellier, 1971.

[CA15] ———, *Intégrales convexes duales*, C. R. Acad. Sci. Paris Sér. A, 275 (1972), pp. 1331–1334.

[CA16] ———, *Intégrales convexes duales*, Travaux du Séminaire d'Analyse Convexe, vol. 3, Exp. No. 6, Secrétariat des Mathématiques, Publication No. 125, U.E.R. de Math., Univ. Sci. Tech. Languedoc, Montpellier, 1973.

[CA17] ———, *Un théorème d'existence de sections séparément mesurables et séparément absolument continues*, Travaux du Séminaire d'Analyse Convexe, vol. 3, Exp. No. 3, Secrétariat des. Mathématiques, Publication No. 125, U.E.R. de Math., Univ Sci. Tech. Languedoc, Montpellier, 1973.

[CA18] ———, *Un théorème d'existence de sections séparément mesurables et séparément absolument continues d'une multiapplication séparément mesurable et séparément absolument continue*, C. R. Acad. Sci. Paris Sér. A, 276 (1973), pp. 367–370.

[CA19]′ ———, *Quelques proprietés du profil d'un convexe compact variable*, Travaux du Séminaire d'Analyse Convexe, vol. 15, Exp. No. 4, Secrétariat des Mathématiques, U.E.R. de Math., Univ. Sci. Tech. Languedoc, Montpellier, 1975.

[CA20] ———, *A compactness theorem of measurable selections of a measurable multifunction and its applications*, Actes du Colloque Intégration Vectorielle et Multivoque, Alain Costé, ed., Université de Caen, 1975.

[CA21]′ ———, *Quelques propriétés du profil d'un convexe compact variable*, Trauvaux du Séminaire d'Analyse Convexe, vol. 5, Exp. No. 4, Secrétariat des Mathématiques, U.E.R. de Math., Univ. Sci. Tech. Languedoc, Montpellier, 1975.

[CA22] ———, *Sur l'existence des sections séparément mesurables et séparément continues d'une multi-application*, Trauvaux du Séminaire d'Analyse Convexe, vol. 5, Exp. No. 14, Secrétariat des Mathématiques, U.E.R., des Math., Univ. Sci. Tech., Languedoc, Montpellier, 1975.

[CV1]′ C. CASTAING AND M. VALADIER, *Equations differentielles multivoques dans les espaces vectoriels localement convexes*, Rev. Française Automat. Inf. Rech. Opér., 3 (1969), pp. 3–16.

[CV2]″ ———, *Measurable Multifunctions and Applications*, Springer-Verlag, to appear.

[CV3]″ ———, *Convex Analysis and Measurable Multifunctions*, to appear.

[CH] J. P. R. CHRISTENSEN, *Topology and Borel Structure*, North-Holland, Amsterdam, 1974.

[CB1] M. M. ČOBAN, *Many-valued mappings and Borel sets I*, Trudy Moskov. Mat. Obšč., 22 (1970) = Trans. Moscow Math. Soc. 22 (1970) 258–280.

[CB2] ———, *Many-valued mappings and Borel sets II*, Trudy Moskov. Mat. Obšč., 23 (1970) = Trans. Moscow Math. Soc. 23 (1970), pp. 286–310.

[CB3] ———, *On B-measurable sections*, Dokl. Akad. Nauk. SSSR, 207 (1972), pp. 48–51 = Soviet Math. Dokl. 13 (1972), pp. 1473–1477.

[CL1] J. K. COLE, *A selector theorem in Banach spaces*, J. Optimization Theory Appl., 7 (1971), pp. 170–172.

[CL2]′ ———, *A note on a selector theorem in Banach spaces*, Ibid., 9 (1972), pp. 214–215.

[CR]′ R. R. CORNWALL, *Conditions for the graph and the integral of a correspondence to be open*, J. Math. Anal. Appl., 39 (1972), pp. 771–792.

[CS1] A. COSTÉ, *Set-valued measures*, Topology and Measure Theory, Zinnowitz (D.D.R), to appear.

[CS2] ———, *Sur l'intégration par rapport à une multimesure de Radon*, C. R. Acad. Sci. Paris Sér. A., 278 (1974), pp. 545–548.

[CS3] ———, *Sur les multimesures à valeurs fermées bornées d'un espace de Banach*, Ibid., 280 (1975), pp. 567–570.

[CS4]′ ———, *La propriété de Radon–Nikodym en intégration multivoque*, Ibid., 280 (1975), pp. 1515–1518.

[CS5]″ ———, *Applications de la théorie des probabilités cylindriques et des opérateurs radonifiants à l'étude des fonctions aléatoires sous-linéaires et des multimesures*, Ibid., 282 (1976), pp. 103–106.

[CS6] ———, *On multivalued additive set functions*, Ark. Mat., to appear.

[CS7] ———, *Densité des sélecteurs d'une multimesure à valeurs convexes fermées bornées d'un espace de Banach séparable*, C. R. Acad. Sci. Paris. Sér. A, 282 (1976), pp. 967–969.

[CP1]′ A. COSTÉ AND R. PALLU DE LA BARRIÈRE, *Un théorème de Radon–Nikodym pour les multimesures à valeurs convexes fermées localement compactes sans droite*, Ibid., 280 (1975), pp. 225–258.

[CP2]′ ———, *Radon–Nikodyms theorems for set-valued measures whose values are convex and closed*, to appear.

[CP3] ———, *Sur L'ensemble des sections d'une multimesure à valeurs convexes fermées*, C. R. Acad. Sci. Paris Sér. A-B, 282 (1976), no. 106, Ai, A 828–832.

[DR] R. B. DARST, *Remarks on measurable selections*, Ann. Statist., 2 (1974), pp. 845–847.

[DT]′ R. DATKO, *Measurability properties of set-valued mappings in a Banach space*, this Journal, 8 (1970), pp. 226–238.

[DV] J. D. DAUER AND F. S. VAN VLECK, *Measurable selectors of multifunctions and applications*, Math. Systems Theory, 7 (1974), pp. 367–376.

[DE]′ G. DEBREU, *Integration of Correspondences*, Proc. Fifth Berkeley Symp. Math. Statist. and Probability (1965/66), vol. II, University of California Press, Berkeley, CA, 1967, pp. 351–372.

[DES]′ G. DEBREU AND D. SCHMEIDLER, *The Radon–Nikodym derivative of a correspondence*, Proc. Sixth Berkeley Symp. Math. Statist. and Probability, vol. II, Univ. of California Press, Berkeley, CA, 1971, pp. 41–56.

[DL] C. DELODE, *Champs mesurables d'espaces sousliniens*, Département de Mathématiques, Université de Pau, 1976.

[DAP1] C. DELODE, O. ARINO AND J. P. PENOT, *Champs mesurables d'espaces polonais*, C. R. Acad. Sci. Paris Sér. A., 281 (1975), pp. 617–620.

[DAP2] ———, *Champs mesurables et multisections*, Ann. Inst. H. Poincaré Sect. B, 12 (1976), pp. 11–42.

[DW]′ M. DEWILDE, *A note on the bang–bang principle*, J. London Math. Soc. (2), 1 (1969), pp. 753–759.

[DI] J. DIXMIER, *Dual et quasi-dual d'une algebre de Banach involutive*, Trans. Amer. Math. Soc., 104 (1962), pp. 278–283.

[DS] L. E. DUBINS AND L. J. SAVAGE, *How to Gamble If You Must*, McGraw-Hill, New York, 1965.

[EV] I. EKELAND AND M. VALADIER, *Representation of set-valued mappings*, J. Math. Anal. Appl., 35 (1971), pp. 621–629.

[EN] R. ENGELKING, *Selectors of the first Baire class for semicontinuous set-valued functions*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 16 (1968), pp. 277–282.

[FE]″ H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[FM]′ H. FEDERER AND A. MORSE, *Some properties of measurable functions*, Bull. Amer. Math. Soc., 49 (1943), pp. 270–277.

[FG] J. FELDMAN AND F. P. GREENLEAF, *Existence of Borel transversals in groups*, Pacific J. Math., 25 (1968), pp. 455–461.

[FI] A. FILIPPOV, *On certain questions in the theory of optimal control*, Vestnik Moskov. Univ. Ser. Mat. Meh. Astronom. Fiz. Him. (1959), no. 2, pp. 25–32; English transl., this Journal, 1 (1962), pp. 76–84.

[FL] W. M. FLEISCHMAN, ed., *Set-valued Mappings, Selections, and Topological Properties of* 2^X, Springer-Verlag, New York, 1970.

[FU] N. FURUKAWA, *Markovian decision processes with compact action spaces*, Ann. Math. Statist., 43 (1972), pp. 1612–1622.

[GM] F. S. GARNIR-MONJOIE, *Sur certaines multifonctions μ-mesurables*, Bull. Soc. Roy. Sci. Liège, 42 (1973), pp. 183–194.

[GGM] H. G. GARNIR AND F. S. GARNIR-MONJOIE, *Multifonctions dans l'espace Euclidien*, Université de Liège, 1973.

[GL]″ J. GLIMM, *Locally compact transformation groups*, Trans. Amer. Math. Soc., 101 (1961), pp. 124–138.

[GT1] C. GODET-THOBIE, *Sur les sélections de mesures généralisées*, C. R. Acad. Sci. Paris Sér. A, 271 (1970), pp. 153–156.

[GT2] ———, *Sur les intégrales de fonctions réelles par rapport à une mesure généralisée et ses sélections*, Ibid., 271 (1970), pp. 497–500.

[GT3] ———, *Sur les multimesures de transition*, Ibid., 278 (1974), pp. 1367–1369.

[GT4] ———, *Sélections de multimesures*, Ibid., 279 (1974), pp. 603–606.

[GT5] ———, *Multimesures et Multimesures de Transition*, Thèse, Montpellier, 1975.

[GT6] ———, *On transition multimeasures*, Actes de Colloque Intégration Vectorielle et Multiŏque, Alain Costé, ed., Université de Caen, 1975.

[GD]′ G. S. GOODMAN, *On a theorem of Scorza-Dragoni and its application to optimal control*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, 1967, pp. 222–233.

[GHJ]′ G. S. GOODMAN AND J. HOFFMANN-JØRGENSEN, *Support functions and the integration of convex sets in infinite-dimensional spaces*, Colloque sur la Théorie Mathématique du Côntrole Optimal, Brussels, 1969, pp. 83–97.

[GZ]′ L. GÖTZ, *Bang–Bang Prinzipien*, Zulassungsarbeit zum Staatsexamen, Erlangen, 1972/73.

[GR]″ F. P. GREENLEAF, *Invariant Means on Topological Groups*, Van Nostrand Mathematics Study Series No. 16, Van Nostrand–Reinhold, New York, 1969.

[HL]″ P. R. HALMOS, *Measure Theory*, Van Nostrand, Princeton, 1954.

[HN]′ R. W. HANSELL, *Borel measurable mappings for nonseparable metric spaces*, Trans. Amer. Math. Soc., 161 (1971), pp. 145–169.

[HF]″ F. HAUSDORFF, *Set Theory*, Chelsea, New York, 1962 (second edition of translation from German).

[HE1] H. HERMES, *A note on the range of a vector measure, Application to the theory of optimal control*, J. Math. Anal. Appl., 8 (1964), pp. 78–83.

[HE2]′ ———, *Calculus of set-valued functions and control*, J. Math. Mech., 18 (1968), pp. 47–60.

[HE3]′ ———, *The generalized differential equations* $\dot{x} \in R(t, x)$, Advances in Math., 4 (1970), pp. 149–169.

[HI]″ W. HILDEBRAND, *Core and Equilibrium of a Large Economy*, Princeton University Press, Princeton, 1974.

[HM1]′ C. J. HIMMELBERG, *Precompact contraction of metric uniformities and the continuity of* $F(t, x)$, Rend. Sem. Mat. Univ. Padova, 50 (1973), pp. 185–188.

[HM2] ———, *Measurable relations*, Fund. Math., 87 (1975), pp. 53–72.

[HJV] C. J. HIMMELBERG, M. Q. JACOBS AND F. S. VAN VLECK, *Measurable multifunctions, selectors and Filippov's implicit functions lemma*, J. Math. Anal. Appl., 25 (1969), pp. 276–284.

[HPRV]″ C. J. HIMMELBERG, T. PARTASARATHY, T. E. S. RAGHAVAN AND F. S. VAN VLECK, *Existence of p-equilibrium and optimal stationary strategies in stochastic games*, Proc. Amer. Math. Soc., 60 (1976), pp. 245–251.

[HPV] C. J. HIMMELBERG, T. PARTHASARATHY AND F. S. VAN VLECK, *Optimal plans for dynamic programming problems*, Mathematics of Operations Research, to appear.

[HV1] C. J. HIMMELBERG AND F. S. VAN VLECK, *Some remarks on Filippov's lemma*, Proc. U.S.-Japan Seminar on Differential Functional Equations, W. A. Benjamin, New York, 1967, pp. 455–462.

[HV2] ———, *Some selection theorems for measurable functions*, Canad. J. Math., 21 (1969), pp. 394–399.

[HV3] ———, *Selections and implicit function theorems for multifunctions with Souslin graph*, Bull. Acad. Polon. Sci., 19 (1971), pp. 911–916.

[HV4]′ ———, *Lipschitzian generalized differential equations*, Rend. Sem. Mat. Univ. Padova, 48 (1972), pp. 159–169.

[HV5] ———, *Extreme points of multifunctions*, Indiana Univ. Math. J., 22 (1973), pp. 719–729.

[HV6] ———, *Multifunctions on abstract measurable spaces and application to stochastic decision theory*, Ann. Mat. Pura. Appl., 101 (1974), pp. 229–236.

[HV7]′ ———, *Multifunctions with values in a space of probability measures*, J. Math. Anal. Appl., 50 (1975), pp. 108–112.

[HV8]″ ———, *Two existence theorems for generalized differential equations*, submitted for publication.

[HV9]′ ———, *An extension of Brunovský's Scorza-Dragoni type theorem for unbounded set-valued functions*, Mat. Časopis Sloven. Akad. Vied., 26 (1976), pp. 179–188.

[HD1] K. HINDERER, *Zur Theorie Stochastischer Entscheidungsmodelle*, Habilitationsschrift, University of Stuttgart, 1967.

[HD2] ———, *Foundations of non-stationary dynamic programming with discrete time parameter*, Lecture Notes in Operations Research and Mathematical Systems, vol. 33, Springer-Verlag, New York, 1970.

[HJ] J. HOFFMANN-JØRGENSEN, *The Theory of Analytic Spaces*, Various Publication Series, No. 10, Århus Universitet, Denmark, 1970.

[HU]′ M. HUKAHARA, *Intégration des applications mesurables dont la valeur est un compact convexe*, Funkcial. Ekvac., 10 (1967), pp. 205–223.

[JC1] M. Q. JACOBS, *Remarks on some recent extensions of Filippov's implicit functions lemma*, this Journal, 5 (1967), pp. 622–627.

[JC2] ———, *Measurable multivalued mappings and Lusin's theorem*, Trans. Amer. Math. Soc., 134 (1968), pp. 471–481.

[JC3]′ ———, *On the approximation of integrals of multivalued functions*, this Journal, 7 (1969), pp. 158–177.

[JN] V. JANKOV, *Sur l'uniformisation des ensembles A*, C. R. Acad. Sci. URSS, 30 (1941), pp. 597–598.

[JO]′ J. A. JOHNSON, *Extreme measurable selections*, Proc. Amer. Math. Soc., 44 (1974), pp. 107–112.

[KP] J. KANIEWSKI AND R. POL, *Borel-measurable selectors for compact-valued mappings in the non-separable case*, Bull. Acad. Polon. Sci. Sér. Math. Astronom. Phys., 23 (1975), pp. 1043–1050.

[KE] H. G. KELLERER, *Bemerkung zu einem Satz von H. Richter*, Arch. Math., 15 (1964), pp. 204–207.

[KO] M. KONDŌ, *Sur l'uniformisation des complémentaires analytiques et les ensembles projectifs de la seconde classe*, Japan. J. Math., 15 (1938), pp. 197–230.

[KR]″ M. A. KRASNOSEL'SKII, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, New York, 1964.

[KD]′ H. KUDŌ, *Dependent experiments and sufficient statistics*, Natur. Sci. Rep. Ochanomizu Univ., 4 (1954), pp. 151–163.

[KU1]″ K. KURATOWSKI, *Topology*, vol. 1, Fifth ed., Academic Press, New York, 1966.

[KU2]″ ———, *Topology*, vol. 2, Fifth ed., Academic Press, New York, 1968.

[KU3]′ ———, *A general approach to the theory of set-valued mappings*, Proc. Third Prague Topological Symp., 1971, pp. 271–280.

[KU4] ———, *On the selector problems for the partitions of Polish spaces and for the compact-valued mappings*, Ann. Polon. Math., 29 (1974), pp. 421–427.

[KU5] ———, *The σ-algebra generated by Souslin sets and its applications to set-valued mappings and to selector problems*, Boll. Un. Mat. Ital., 11 (1975), pp. 285–298.

[KU6] ———, *On partitions of complete spaces which do not admit analytic selectors and on some consequences of a theorem of Gödel*, Instituto Naxionale di Alta Matematica, Symposia Mathematica, 16 (1975), pp. 67–74.

[KMT] K. KURATOWSKI AND A. MAITRA, *Some theorems on selectors and their applications to semi-continuous decompositions*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 22 (1974), pp. 877–881.

[KMS]″ K. KURATOWSKI AND A. MOSTOWSKI, *Set Theory*, second ed., North-Holland, Amsterdam, 1976.

[KRN] K. KURATOWSKI AND C. RYLL-NARDZEWSKI, *A general theorem on selectors*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 13 (1965), pp. 397–403.

[LA1] D. G. LARMAN, *Projecting and uniformising Borel sets with $\mathcal{K}_\sigma$ sections I*, Mathematika, 19 (1972), pp. 231–244.

[LA2] ———, *Projecting and uniformising Borel sets with $\mathcal{K}_\sigma$ sections II*, Ibid., 20 (1973), pp. 233–246.

[LE1] S. J. LEESE, *Measurable selections in normed spaces*, Proc. Edinburgh Math. Soc., 19 (1974), pp. 147–150.

[LE2] ———, *Multifunctions of Souslin type*, Bull. Austral. Math. Soc., 11 (1974), pp. 395–411.

[LE3] ———, *Set-valued functions and selectors*, Thesis, Keele, England, 1974.

[LE4] ———, *Multifunctions of Souslin type: Corrigendum*, Bull. Austral. Math. Soc., 13 (1975), pp. 159–160.

[LE5] ———, *Measurable selections and the uniformisation of Souslin sets*, Amer. J. Math., to appear.

[LE6] ———, *Continuous and Borel selectors in normed spaces*, Loughborough University Report No. 68, July 1975.

[LN1] J. LEHN, *Massfortsetzungen und Aumann's Selektionstheorem*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 35 (1976), pp. 265–268.

[LN2]′ ———, *Remark on measurable graph theorems*, Proc. Amer. Math. Soc., to appear.

[LM]″ J. LEHN AND G. MÄGERL, *On the uniqueness of preimage measures*, submitted for publication.

[LB]″ A. LUBIN, *Extensions of measures and the von Neumann selection theorem*, Proc. Amer. Math. Soc., 43 (1974), pp. 118–122.

[LS] N. LUSIN, *Leçons sur les Ensembles Analytiques et Leurs Applications*, Chelsea, New York (second edition), 1972. (First edition was 1930.)

[MC1] G. W. MACKEY, *Induced representations of locally compact groups I*, Ann. of Math., 55 (1952), pp. 101–139.

[MC2]″ ———, *Borel structure in groups and their duals*, Trans. Amer. Math. Soc., 85 (1957), pp. 134–165.

[MG] G. MÄGERL, *Stetige und messbare Schnitte fur Korrespondenzen*, Diplomarbeit, Erlangen, 1973.

[MT] A. MAITRA, *Discounted dynamic programming on compact metric spaces*, Sankhyā Ser. A, 27 (1968), pp. 241–248.

[MR1] A. MAITRA AND B. V. RAO, *Selection theorems and the reduction principle*, Trans. Amer. Math. Soc., 202 (1975), pp. 57–66.

[MR2] ———, *Selection theorems for partitions of Polish spaces*, Fund. Math., 93 (1976), pp. 47–56.

[MZ] P. MARITZ, *Integration of set-valued functions*, Thesis, Rijksuniversiteit te Leiden, 1975.
[MK] S. MAZURKIEWICZ, *Sur une propriété des ensembles $C(A)$*, Fund. Math., 10 (1927), pp. 172–174.
[MW] E. J. MCSHANE AND R. B. WARFIELD, JR., *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.
[MI1] E. A. MICHAEL, *Selected selection theorems*, Amer. Math. Monthly, 4 (1956), pp. 233–236.
[MI2] E. A. MICHAEL, *G_δ sections and compact-covering maps*, Duke. Math. J., 36 (1969), pp. 125–128.
[MU] D. R. MUNOZ, *Integration of Correspondences*, Thesis, Univ. of California, San Diego, 1971.
[NE] J. VON NEUMANN, *On rings of operators, reduction theory*, Ann. of Math., 30 (1949), pp. 401–485.
[NO1] P. NOVIKOV, *Sur les fonctions implicites mesurables B*, Fund. Math., 17 (1931), pp. 8–25.
[NO2] ———, *Sur les projections de certain ensembles mesurables*, C. R. (Doklady) Acad. Sci. URSS, 23 (1939), pp. 864–865.
[NU] G. NÜRNBERGER, *Dualität von Schnitten für die metrische Projektion und von Fortsetzungen kompakter Operatoren*, Dissertation, Erlangen, 1975.
[OL] C. OLECH, *A note concerning set-valued measurable functions*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 13 (1965), pp. 317–321.
[PB1] R. PALLU DE LA BARRIÈRE, *Queques propriétés des multimesures*, Travaux du Séminaire d'Analyse Convexe, vol. 3, Exp. No. 11, Secrétariat des Mathématiques U.E.R. de Math., Univ. Sci. Tech. Languedoc, Montpellier, 1973.
[PB2] ———, *Multimesures à valeurs convexes fermées*, Actes du Colloque Intégration Vectorielle et Multivoque, Alain Costé, ed., Université de Caen, 1975.
[PR1] T. PARTHASARATHY, *Selection Theorems and Their Applications*, Springer-Verlag, New York, 1972.
[PR2] ———, *Discounted, positive, and noncooperative stochastic games*, Internat. J. Game Theory, 2 (1973), pp. 25–37.
[PL1] A. PLIŚ, *Remark on measurable set-valued functions*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 9 (1961), pp. 857–859.
[PL2] ———, *Measurable orientor fields*, Ibid., 13 (1965), pp. 565–569.
[RI] H. RICHTER, *Verallgemeinerung eines in der Statistik benötigten Satzes der Masstheorie*, Math. Ann., 150 (1963), pp. 85–90.
[RB] A. P. ROBERTSON, *On measurable selections*, Proc. Royal Soc. Edinburgh, 72 (1972/73), pp. 1–7.
[RC1] R. T. ROCKAFELLAR, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.
[RC2] ———, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.
[RC3] ———, *Convex integral functionals and duality*, Contributions to Nonlinear Functional Analysis, Academic Press, New York, 1971.
[RC4] ———, *Weak compactness of level sets of integral functionals*, Proc. Troisième Colloque d'Analyse Fonctionelle, H. G. Garnir, ed., Liège, 1971.
[RC5] ———, *Integrals which are convex functionals, II*, Pacific J. Math., 39 (1971), pp. 439–469.
[RC6] ———, *Integral functionals, normal integrands, and measurable selections*, Proc. Colloq. Nonlinear Operators and the Calculus of Variations, Lecture Notes in Mathematics, 543, G. P. Gossez et al., ed., Springer-Verlag, New York, 1976, pp. 157–207.
[RCW] R. T. ROCKAFELLAR AND J. B. WETS, *Continuous versus measurable recourse in N-stage stochastic programming*, J. Math. Anal. Appl., 48 (1974), pp. 836–859.
[RG] C. A. ROGERS, *Hausdorff Measures*, Cambridge University Press, Cambridge, England, 1970.
[RGW] C. A. ROGERS AND R. C. WILLMOTT, *On the uniformization of sets in topological spaces*, Acta. Math., 120 (1968), pp. 1–52.
[RK1] V. A. ROKHLIN, *On decomposition of a dynamical system into transitive components*, Mat. Sb., 25 (1949), pp. 235–249.
[RK2] ———, *Selected topics from the metric theory of dynamic systems*, Uspekhi Mat. Nauk. 4 (1949), pp. 57–128; English transl., Amer. Math. Soc. Transl., 49 (1966), pp. 171–240.
[RN] C. RYLL-NARDZEWSKI, *On Borel measurability of orbits*, Fund. Math., 56 (1964), pp. 129–130.

[SB1] M. F. SAINTE-BEUVE, *Sur la généralisation d'un théorème de section mesurable de von Neumann–Aumann*, Travaux du Séminaire d'Analyse Convexe, vol. 3, Exp. No. 7, Secrétariat des Mathématiques, U.E.R. de Math., Univ. Sci. Tech. Languedoc, Montpellier, 1973.

[SB2] ———, *Sur la généralisation d'un théorème de section mesurable de von Neumann–Aumann et applications à un théorème de fonctions implicites mesurables et à un théorème de représentation intégrale*, C. R. Acad. Sci. Paris Sér. A, 276 (1973), pp. 1297–1300.

[SB3] ———, *On the extension of von Neumann–Aumann's theorem*, J. Funct. Anal., 17 (1974), pp. 112–129.

[SB4]′ ———, *Mesures vectorielles à variation bornée et représentation intégrale*, Actes du Colloque Intégration Vectorielle et Multivoque, Alain Costé, ed., Université de Caen, 1975.

[SK] S. SAKS, *Theory of the Integral*, Dover, New York, 1968.

[SM] Y. SAMPEI, *On the uniformization of the complement of an analytic set*, Comment. Math. Univ. St. Paul., 10 (1961), pp. 57–62.

[SR] H. SARBADAIKARI, *Some uniformization results*, Fund. Math., to appear.

[SC1] M. SCHÄL, *A selection theorem for optimization problems*, Arch. Math., 25 (1974), pp. 219–224.

[SC2] ———, *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 32 (1975), pp. 179–196.

[SD]′ G. SCORZA-DRAGONI, *Un teorema sulla funzioni continue rispetto ad una e misurabili rispetto ad un' altra variabile*, Rend. Sem. Math. Univ. Padova, 17 (1948), pp. 102–106.

[SP1] W. SIERPINSKI, *Sur l'uniformisation des ensembles mesurables (B)*, Fund. Math., 16 (1930), pp. 136–139.

[SP2] ———, *Sur deux complémentaires analytiques non separable B*, Ibid., 17 (1931), pp. 296–297.

[SN] M. SION, *On uniformization of sets in topological spaces*, Trans. Amer. Math. Soc., 96 (1960), pp. 237–245.

[SV] L. M. SONNEBORN AND F. S. VAN VLECK, *The bang–bang principle for linear control systems*, this Journal, 2 (1965), pp. 151–159.

[ST]″ R. E. STRAUCH, *Negative dynamic programming*, Ann. Math. Statist., 37 (1966), pp. 871–890.

[SY]′ C. SUNYACH, *Sur le théorème du graphe borélien*, C. R. Acad. Sci. Paris Sér. A, 269 (1969), pp. 233–234.

[SZ] T. SUZUKI, *On the uniformization principle*, Proc. Symposium on the Foundations of Mathematics, Katadaj, Japan, 1962, pp. 137–144.

[TH1] D. S. THIAM, *Multimesures positive*, C. R. Acad. Sci. Paris Sér. A, 280 (1975), pp. 993–995.

[TH2] ———, *Applications à l'intégration multivoque de l'intégrale de Daniell à valeurs dans un monoide ordonné*, Actes du Colloque Intégration Vectorielle et Multivoque, Alain Costé, ed., Université de Caen, 1975.

[VA1] M. VALADIER, *Sur l'intégration d'ensembles convexes compacts en dimension infinie*, C. R. Acad. Sci. Paris Sér. A, 266 (1968), pp. 14–16.

[VA2] ———, *Contributions à l'analyse convexe*, Thèse, Paris, 1970.

[VA3] ———, *Multi-applications mesurables à valeurs convexes compactes*, J. Math. Pures Appl., 50 (1971), pp. 265–297.

[VA4] ———, *Convex integrands on Souslin locally convex spaces*, Travaux du Séminaire d'Analyse Convexe, vol. 3, Exp. No. 2, Secrétariat des Math., Publication No. 125, U.E.R. de Math., Univ. Sci. Tech. Languedoc, Montpellier, 1973.

[VA5] ———, *Intégrandes sur des localement convexes sousliniens*, C. R. Acad. Sci. Paris Sér. A, 276 (1973), pp. 693–695.

[VA6] ———, *On the Strassen theorem*, Travaux du Séminaire d'Analyse Convexe, vol. 4, Exp. No. 4, Secrétariat des Math., U.E.R. de Math., Univ. Sci. Tech. Languedoc, Montpellier, 1974.

[VA7] ———, *Sur le théorème de Strassen*, C. R. Acad. Sci. Paris Sér, A. 278 (1974), pp. 1021–1024.

[VA8]″ ———, *About the Strassen theorem for vector valued sublinear functions*, Actes du Colloque Intégration Vectorielle et Multivoque, Alain Costé, ed., Université de Caen, 1975.

[WG1] D. H. WAGNER, *Integral of a convex-hull-valued function*, J. Math. Anal. Appl., 50 (1975), pp. 548–559.

[WG2]″ ———, *Integral of a set-valued function with semi-closed values*, Ibid., 55 (1976), pp. 616–633.

[WS] 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.

[WZ] T. WAZEWSKI, *Sur une condition d'existence des fonctions implicites mesurables*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 9 (1961), pp. 861–863.

[WE1] E. WESLEY, *Extensions of the measurable choice theorem by means of forcing*, Israel J. Math., 14 (1973), pp. 104–114.

[WE2]″ ———, *Borel preference orders in markets with a continuum of traders*, J. Math. Economics, 3 (1976), pp. 155–165.

[WH] W. WHITT, *Baire classification of measured selections of extrema*, School of Organization and Management, Yale Univ., 1976.

[WI] R. C. WILLMOTT, *On the uniformization of Souslin $\mathcal{F}$ sets*, Proc. Amer. Math. Soc., 22 (1969), pp. 148–155.

[YO]″ L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

Addenda in proof. In the above, coverage of Russian contributions to measurable selection theory is inadequate. Items (i), (ii), and (iii) below were brought to our attention very recently by A. D. Ioffe via Rockafellar. We had just previously learned of item (i) from E. B. Dynkin via Aumann. It is hoped that Russian contributions will be further surveyed in subsequent publications by Ioffe and perhaps others.

Items (iv) through (xi), given in order of the sections to which they relate, note various additional matters of which we have recently learned. Of these, we consider the announcement by Cenzer and Mauldin in (viii) most important.

(i) Our comment in § 12 on Jankov's [JN] is particularly inadequate. (Henceforth, we transliterate "Yankov.") Statement (3) in the proof of his theorem is the main content of what has been widely called the "von Neumann selection theorem" (5.1 and 5.2 above). In our usages it says: if $T = X = R$ and Gr F is Suslin, then F has a selection which is a Lebesgue measurable function. (This does not follow from his theorem statement, which is given in § 12, by reason of the last sentence in § 12.) To understand the proof in [JN] (also given in [AL, Satz 32]) recourse must be made to usages of [LS], to which he refers, as follows (we are grateful to R. D. Mauldin, A. A. Yuškevič, and J. C. Oxtoby for clarifying these points): (a) all real domains are identified with the irrationals in (0, 1), further identified with ω^ω as usual (e.g., see 5.1 above); (b) an "elementary G_δ" is (the graph of) a continuous map on ω^ω; and (c) "inferior point" means lexicographic minimum. He should probably have stated that $\sum_k \delta_{(n+1)k} \subset \sum_k \delta_{nk}$ (easily obtained), although that appears to be implicit in the definitions from [LS]. Reference in [JN] to the Baire property is redundant since the σ-algebra generated by the Suslin subsets of R^n is contained in the family of Baire property sets.

Yankov's [JN] was published in 1941 and was presented in 1940. Von Neumann's [NE] appeared in 1949, having been submitted in 1948; it states at the outset that the paper was written in 1937–38 and publication was delayed to make certain changes which are itemized and which do not pertain to the selection result, Lemma 5. Both authors obtained the same selection (lexicographic minimum), by different constructions. We have no doubt that these two works were independent of each other, having moreover consulted two former collaborators of von Neumann's, F. J. Murray and H. H. Goldstine. Murray observes

that Lemma 5 is of central importance to [NE] ("without it there is no paper"). He recalls a prewar conversation in which von Neumann spoke with pride over solving this selection problem (although it is not spotlighted in [NE] and was little known for several years). Of course, Yankov was the first to publish.

We conclude that a statement of the form of 5.1 or 5.2 above is appropriately called a "Yankov–von Neumann theorem." Subsequent improvements by Aumann, Sainte-Beuve, and Leese have resulted in 5.10 above.

In Russian literature (e.g., [AL, § 11], [NA, § 40.3 and App. IV], [IT1], [IT2]) statements such as 5.2 have been referred to as the "Lusin–Yankov theorem"; [RK1] credits Yankov. Having reviewed [LS2], which evidently inspired [JN], we do not conclude that Lusin should be credited with this result, despite his eminent pioneering contributions to the foundations of the subject (e.g., see § 3 above). It does appear that the construction on page 57 of [LS2] (which differs from those of [JN] and [NE]) if specialized in the most natural way, yields the Yankov–von Neumann selection. However, [LS2] does not prove that his selection is a Lebesgue *measurable function*, in fact, in contrast to [NE], neither he nor Yankov appeared to seek that kind of result. Again, Yankov did state and prove a measurable function result during his proof of his theorem.

(ii) A second important omission, pointed out by Ioffe, is Novikov's [NO2, Cor. 2] (1939), also quoted in [AL, § 14], which we render: if $T = X = R$, F is closed-valued, and $\operatorname{Gr} F$ is Borel, then F has what has now been termed a Castaing representation. Contrary to the end of § 3 above, this is the first result on existence of measurable selections without assuming countable or compact values.

(iii) Ioffe points out that Rokhlin's argument in [RK2] (also given in [RK1]), discussed in § 4, becomes a valid proof if the following changes are made (we concur): (a) replace 2^{-n} by 2^{-n+2} in (10_n), and (b) redefine A_i to be $B_i / \bigsqcup_{j=1}^{i-1} B_j$, where

$$B_i = \{x : r(Y_i, \Psi(x)) < 2^{-n} \text{ and } r(Y_i, \psi_{n-1}(x)) < 2^{-n+2}\}.$$

Moreover, this argument suffices for Theorem 4.1 as given above, without the Lebesgue space assumption made by Rokhlin.

Ioffe feels that the error in [RK2] was "insignificant and easily correctible." Were it only (a), we would agree. However, (b) is a substantive change, e.g., the new A_i involves the approximating function ψ_{n-1} and in [RK2] it did not. Therefore, we feel the argument in [RK2] should be regarded as incomplete. Thanks to Ioffe, we now know that it is completable within the main ideas of Rokhlin's reasoning. Thus, Rokhlin gave in 1949 a statement of the essence of Theorem 4.1 and the principal ideas of its proof.

From the facts on the origin of Theorem 4.1 given after its statement and from the observations just made, it appears that the credit for this result is somewhat diffuse among, chronologically, Rokhlin [RK2], Kuratowski and Ryll-Nardzewski [KRN], and Castaing [CA, 1, 2, 4]. Moreover, Novikov contributed a significant special case (see (ii)) in 1939, albeit with the strong assumption that $\operatorname{Gr} F$ is Borel. We propose that Theorem 4.1 be given the impersonal name "Fundamental Measurable Selection Theorem," which we believe is commensurate with its importance.

(iv) A new and fairly general exposition of measurable selections and continuous selections is given by Kuratowski and Mostowski [KMS, Chap. XIV]. A briefer discussion in similar vein (in Polish) is given in [KU7].

(v) In § 4 and § 8 we have noted Rockafellar's use of F such that $F(t)$ is the epigraph of a convex $f(t, \cdot)$ for $t \in T$, where $f: T \times X \to R \cup \{\infty, -\infty\}$ and μ is complete and σ-finite. He points out (personal communication) that for the most part, "convex" is weakened to "lsc" and "complete" is avoided in [RC6], which supersedes most of the finite-dimensional parts of [RC1–5]. In [RC6], the key condition for such an f to be normal, by definition, is that F be measurable, and the latter property is the focus of his manipulations, via Theorem 4.2(e) ((ii)$\Leftrightarrow$(ix)).

With this approach, Rockafellar obtains in a relatively easy way, within $X = R^n$, variants of several results reviewed above, e.g., his result in (vii) below and an implicit function result [RC6, Thm. 2J] in which the g constraint in § 7 is generalized to an infinite sequence of inequalities.

Evstigneev [ES] has applied Theorem 4.2(e) ((iii)$\Leftrightarrow$(ix)) to dynamic programming problems, generalizing certain results of Rockafellar and West [RCW] and of Dynkin.

(vi) Cenzer and Mauldin [CM1] give variants on Theorem 5.7 above from [BRN]. In one they replace $\mathcal{B}(X)$ by its completion. In another they assume Gr F is complementary Suslin and obtain $2^{\aleph_0}$ distinct Borel function selections of F (Larman [LA1, 2] obtained $\aleph_1$ with F σ-compact-valued—see § 12).

(vii) Schäl [SC3] has answered affirmatively the open question in § 9. Shreve and Bertsekas [SHB] have given a variant of a result of Brown and Purves [BP]. They assume Gr F and, for $a \in R$, $\{(t, x): u(t, x) > a\}$ are Suslin (u as in § 9). Rockafellar [RC6, Thm. 2K] gave the following variant on 9.1 with u and v as in § 9: If $X = R^n$, $-u$ is normal (see (v)), F is measurable and closed-valued, and $G(t) \equiv F(t) \cap \{x: u(t, x) = v(t)\}$ for $t \in T$, then v and G are measurable, and since also G is closed-valued, $\mathcal{S}(G) \neq \varnothing$.

(viii) In [WE3], Wesley proves his universal measurability assertion in § 10. Cenzer and Mauldin announce (personal communication) an extension [CM2] of this result and moreover a proof that uses only standard techniques of descriptive set theory, not requiring forcing or other metamathematical methods: In Theorem 10.3 they replace $\mathcal{L}$, $\mathcal{M}$, and $\mathcal{M} \otimes \mathcal{L}$ by the smaller σ-algebras $S([0, 1])$, $S(T)$, and $S(T \times [0, 1])$, where for a topological space Y, $S(Y)$ is the smallest σ-algebra which is a Suslin family and contains $\mathcal{B}(Y)$.

(ix) Kallman and Mauldin [KAM] have extended Corollary 11.2(ii) (due to Dixmier) as follows (under partition usages of § 11): If $X (= T)$ is a Borel subset of a Polish space, each $F(t)$ is an F_δ and a G_δ in X, $\mathcal{M} = \mathcal{B}(T)$, and F is weakly measurable, then $\mathcal{S}(F) \neq \varnothing$. Kaniewski [KA2] has obtained a Borel set selection of a partition into compact sets of a Borel subset of a metric Suslin space; he also generalizes Kunugui–Novikov [NO2].

(x) Kaniewski [KA1] has generalized Kondō's theorem (§ 12). Mauldin points out that ZFC + MA + not CH (without the constructibility axiom) denies (b) at the end of § 12 (this (b) is included in Aumann's discussion [AU3] of Gr F complementary Suslin.

(xi) Further contributions to selections of set-valued measures (§ 15) have been given by M. Rao [RA], Vincent-Smith [VS], and Talagrand [TA]. These relate to Choquet theory.

BIBLIOGRAPHY ADDED IN PROOF

[AN]″ V. I. ARKIN AND E. L. LEVIN, *The convexity of values of vector integrals, Theorems on measurable selections and variational problems*, Uspehi Mat. Nauk, 27 (1972), pp. 22–77, English transl. Russian Math. Surveys, London Math. Soc., 1972, pp. 21–85.

[BS] D. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, to appear.

[DY]″ E. B. DYNKIN AND A. A. YUŠKEVIČ, *Markov Control Processes and their Applications*, Nauka, Moscow, 1975, English ed. Springer-Verlag, New York, to appear.

[CM1] D. CENZER AND R. D. MAULDIN, *Inductive definitions: Measure and category*, in preparation.

[CM2] ———, *Measurable parametrizations and selections*, submitted for publication.

[ES]′ I. V. EVSTIGNEEV, *Measurable selection and dynamic programming*, Math. Operations Res., 1 (1976), pp. 267–272.

[IT1]″ A. D. IOFFE AND V. M. TIKHOMIROV, *Duality of convex functions and extremal problems*, Uspehi Mat. Nauk, 22 (1968), pp. 51–116; English transl. Russian Math. Surveys, 23 (1968), pp. 53–125.

[IT2]″ ———, *On minimization of integral functionals*, Funkcional. Anal. i Priložen, 3 (1969), pp. 61–70; English transl. Plenum, New York.

[KLM] R. R. KALLMAN AND R. D. MAULDIN, *A cross section theorem and application to C^*-algebras*, Pure Appl. Math. Sci., to appear.

[KA1] J. KANIEWSKI, *A generalization of Kondo's uniformization theorem*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astron. Phys., 24 (1976), pp. 393–398.

[KA2] ———, *A selection theorem for partitions of Borel sets into compact subsets*, Dept. of Math. and Mech., Warsaw Univ., 1976–77.

[KU7]″ K. KURATOWSKI, *O selektorach w topologii i teorii miary* (On selectors in topology and measure theory), Wiadom. Mat., 19 (1975), pp. 3–9.

[LS2] N. LUSIN, *Sur le problème de M. Jacques Hadamard d'uniformisation des ensembles*, Mathematica, 4 (1930), pp. 54–66.

[NA]″ M. NAIMARK, *Normed Rings*, Nauka, Moscow, 1968.

[RA] M. RAO, *Measurable selection of representing measures*, Quart. J. Math. Ser. 2, 22 (1971), pp. 571–572.

[SC3] M. SCHÄL, *Addendum to* [SC1] *and* [SC2], Tech. Rpt., University of Bonn, 1977.

[TA] M. TALAGRAND, *Sélection measurable de mesures maximales*, Séminaire Choquet, 15th year, 1975/76, Communications No. 3, Secrétariat mathématique, Équipe d'Analyse, Université Paris.

[VS] G. F. VINCENT-SMITH, *Measurable selection of simplicial maximal measures*, J. London Math. Soc. Ser. 2, 7 (1973), pp. 427–428.

[WE3] E. WESLEY, *On the existence of absolutely measurable selection functions*, submitted for publication; see Abstract 76T-B208, Notices Amer. Math. Soc., 23 (1977), p. A-648.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 5, August 1977

ERRATUM: STRUCTURAL STABILITY FOR THE RICCATI EQUATION*

R. S. BUCY†

It was pointed out by Shohie Fujita [1] that the condition $\det(\bar{F}_* \otimes I + I \otimes \bar{F}_*) \neq 0$ is not sufficient in Theorem 1, p. 750, as the condition does not imply index $(\bar{F}_* \otimes I + I \otimes \bar{F}_*) = (n_1^*, 0, n_{-1}^*)$. Replacing the condition by the following:

$$\det(\bar{F}_* \otimes I + I \otimes \bar{F}_* + iw) \neq 0 \quad \text{for all real } w$$

and leaving the proof as it stands correct the situation.

REFERENCE

[1] S. FUJITA, *Structural stability for the algebraic Riccati equation*, Internat. Conf. on Information Sciences and Systems, Patras, Greece, August 1976.

* This Journal, 13 (1975), pp. 749–753. Received by the editors September 17, 1976.

† Aerospace Engineering Department, University of Southern California, Los Angeles, California 90007 and Laboratoire d'Automatique, Toulouse, France.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

LEGENDRE DUALITY IN NONCONVEX OPTIMIZATION AND CALCULUS OF VARIATIONS*

IVAR EKELAND†

Abstract. A general duality theory is given for smooth nonconvex optimization problems, covering both the finite-dimensional case and the calculus of variations. The results are quite similar to the convex case; in particular, with every problem $(\mathcal{P})$ is associated a dual problem $(\mathcal{P}^*)$ having opposite value. This is done at the expense of broadening the framework from smooth functions $\mathbb{R}^n \to \mathbb{R}$ to Lagrangian submanifolds of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$.

Introduction. Duality methods are nowadays an important tool in the study of convex optimization problems. A systematic treatment within the framework of convex analysis can be found in the books of R. T. Rockafellar [14] and I. Ekeland and R. Temam [8]. However, it is easily forgotten that duality methods have been in use for quite a long time in classical mechanics, where people are used to stating a problem either in terms of x-phase variables, or of p-momentum variables, the mapping $x \to p$ being the Legendre transformation. A major difficulty lies in the fact that the Legendre transformation need not be one-to-one, except of course in the convex case.

This paper aims to provide people used to convex optimization problems with a systematic and updated treatment of duality theory for the smooth nonconvex case. The first two sections set up the general framework. It turns out that the framework of functions is not broad enough to cover our needs, because the Legendre transform of a smooth nonconvex function need not be a function. So we define Lagrangian submanifolds of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ as a better concept to work with, because the Legendre transform of a Lagrangian submanifold is still a Lagrangian submanifold, and because a Lagrangian submanifold comes very close to being a function from $\mathbb{R}^n$ to $\mathbb{R}$. Section 1 investigates the local properties of Lagrangian submanifolds, and § 2 studies the Legendre transform in this framework.

The duality theorems then follow quite easily, either in § 3 for the finite-dimensional case, or in § 4 for the calculus of variations. They are exactly what one would expect from the convex case. References to the bibliography are relegated to § 5.

1. Lagrangian submanifolds. Let f be a C^∞ real-valued function on $\mathbb{R}^n$. We can associate with f the following n-dimensional submanifold of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$:

$$V_f = \{(x, f'(x), f(x)) | x \in \mathbb{R}^n\}. \tag{1.1}$$

* Received by the editors September 10, 1976.

† Mathematics Research Center, University of Wisconsin—Madison, Madison, Wisconsin. Now at Centre de Recherche de Mathématiques de la Division, Université de Paris IX Dauphine, Paris, France. This work was supported by the United States Army under Contract No. DAAG29-75-C-0024.

This submanifold has the property of annihilating the differential form ω defined at any point (x, p, z) of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ by the formula

$$\omega = dz - \sum_{i=1}^{n} p_i \, dx_i. \tag{1.2}$$

Indeed, the restriction of ω to V_f reduces to $df - \sum_{i=1}^{n} (\partial f/\partial x_i)\, dx_i$ which is identically zero. This motivates the following definitions.

DEFINITION 1.1. A *Lagrangian submanifold* of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ is a closed n-dimensional C^∞-submanifold V such that

$$i_V^* \omega = 0 \tag{1.3}$$

where i_V: $V \to \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ is the canonical injection and i_V^*: $T^*(\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}) \to T^* V$ the induced map of differential 1-forms. We shall say that $\bar{x} \in \mathbb{R}^n$ is a *critical point* of V and that $\bar{z} \in \mathbb{R}$ is a *critical value* whenever

$$(\bar{x}, 0, \bar{z}) \in V. \tag{1.4}$$

We shall associate with V a multivalued mapping F_V from $\mathbb{R}^n$ to $\mathbb{R}$:

$$F_V(x) = \{z \mid \exists\, p \in \mathbb{R}^n : (x, p, z) \in V\} \tag{1.5}$$

and call it the *characteristic map* of V.

In the following, we shall denote by π_x and π_{xz} respectively the restriction to V of the projections $(x, p, z) \to x$ and $(x, p, z) \to (x, z)$. The analogous notations π_p and π_{pz} will also be used. These maps send V into $\mathbb{R}^n$ and $\mathbb{R}^{n+1}$ respectively; note that:

$$\text{graph } F_V = \pi_{xz}(V). \tag{1.6}$$

Particularly simple situations arise when these projections are proper. Recall that a continuous map π: $V \to \mathbb{R}^k$ is proper at $\xi \in \mathbb{R}^k$ iff every sequence ω_n in V such that $\pi(\omega_n) \to \xi$ is bounded. It is proper iff it is proper at every point $\xi \in \mathbb{R}^k$; this amounts to saying that $\pi^{-1}(K)$ is compact in V whenever K is compact in $\mathbb{R}^k$.

As a fundamental example of a Lagrangian submanifold, take the set V_f associated with a C^∞ function f: $\mathbb{R}^n \to \mathbb{R}$ by formula (1.1). Note that in this case π_x is a diffeomorphism from V on $\mathbb{R}^n$, and hence proper.

As a variant, consider a C^∞ function f defined on an open subset Ω of $\mathbb{R}^n$, and assume that $|f(x)| \to \infty$ whenever x converges to some point in the boundary of Ω. Then the set V_f defined by

$$V_f = \{(x, f'(x), f(x)) \mid x \in \Omega\} \tag{1.7}$$

is a Lagrangian submanifold. Note that in this case π_x is a diffeomorphism from V on Ω, but no longer on $\mathbb{R}^n$. Hence π_x is no longer proper, but π_{xz} is.

In both cases, the critical points/values of V_f are the critical points/values of f, and the characteristic map F_V of V_f coincides with f:

$$\forall x \in \mathbb{R}^n, \qquad F_V(x) = \{f(x)\}. \tag{1.8}$$

We now seek a partial converse: describe, at least locally, a given Lagrangian submanifold V, in terms of a smooth function f: $\mathbb{R}^n \to \mathbb{R}$. For that purpose, we

introduce the set $\mathcal{R}$ of points $\bar{x} \in \mathbb{R}^n$ such that the 1-forms $i_V^* dx_1, \cdots, i_V^* dx_n$ are linearly independent at every point $(\bar{x}, \bar{p}, \bar{z})$ of V projecting on $\bar{x}$.

PROPOSITION 1.2. *The subset $\mathbb{R}^n \backslash \mathcal{R}$ has Lebesgue measure zero in $\mathbb{R}^n$. For every point $\bar{x} \in \mathcal{R}$ there exist a (possibly empty) countable set of indices A, a family $\mathcal{U}_\alpha$, $\alpha \in A$, of neighborhoods of $\bar{x}$ in $\mathbb{R}^n$, a family $f_\alpha : \mathcal{U}_\alpha \to \mathbb{R}$ of smooth functions, such that*

$$\pi_x^{-1}(\bar{x}) \subset \bigcup_{\alpha \in A} \mathcal{V}_\alpha \subset V \tag{1.9}$$

where

$$\mathcal{V}_\alpha = \{(x, f'_\alpha(x), f_\alpha(x)) | x \in \mathcal{U}_\alpha\}. \tag{1.10}$$

Note that (1.9) implies that $F_V(\bar{x}) = \{f_\alpha(\bar{x}) | \alpha \in A\}$. Intuitively, the part of F_V lying above $\bar{x}$ is decomposed into smooth branches f_α, $\alpha \in A$, with $z_\alpha = f_\alpha(\bar{x})$ and $p_\alpha = f'_\alpha(\bar{x})$. Two branches may intersect, but they must do so transversally: if $f_\alpha(\bar{x}) = f_\beta(\bar{x})$ with $\alpha \neq \beta$, then $f'_\alpha(\bar{x}) \neq f'_\beta(\bar{x})$.

Proof of Proposition 1.2. To say that the 1-forms $i_V^* dx_1, \cdots, i_V^* dx_n$ are linearly independent at $(\bar{x}, \bar{p}, \bar{z}) \in V$ means that $(\bar{x}, \bar{p}, \bar{z})$ is a regular point for the projection $\pi_x : V \to \mathbb{R}^n$. The set $\mathbb{R}^n \backslash \mathcal{R}$ is just the set of critical values for π_x, and it follows from Sard's theorem that it has measure zero.

Take $\bar{x} \in \mathcal{R}$, and let $\{(\bar{x}, \bar{p}_\alpha, \bar{z}_\alpha) | \alpha \in A\}$ be the (possibly empty) set of points of V projecting on $\bar{x}$. By the definition of $\mathcal{R}$, each $(\bar{x}, \bar{p}_\alpha, \bar{z}_\alpha)$, $\alpha \in A$, is a regular point for π_x. By the implicit function theorem, there are neighborhoods $\mathcal{U}_\alpha$ of $\bar{x}$ and $\mathcal{V}_\alpha$ of $(\bar{x}, \bar{p}_\alpha, \bar{z}_\alpha)$ such that $\pi_x : \mathcal{V}_\alpha \to \mathcal{U}_\alpha$ is a diffeomorphism. In other words, there are real-valued C^∞ functions f_α and $g_{\alpha i}$, $1 \leqq i \leqq n$, defined over $\mathcal{U}_\alpha$, such that

$$(x, p, z) \in \mathcal{V}_\alpha \Leftrightarrow \{x \in \mathcal{U}_\alpha, z = f_\alpha(x), p_i = g_{\alpha i}(x)\}. \tag{1.11}$$

The vanishing of $i_V^* \omega$ means that

$$df_\alpha - \sum_{i=1}^n g_{\alpha i}(x)\, dx_i = 0 \quad \text{over } \mathcal{U}_\alpha, \tag{1.12}$$

which yields

$$g_{\alpha i}(x) = \frac{\partial f_\alpha}{\partial x_i}(x) \quad \forall x \in \mathcal{U}_\alpha. \tag{1.13}$$

Writing (1.13) in (1.11), we get formula (1.10), with formula (1.9) being satisfied by construction. It only remains to prove that the set A is at most countable. For this, notice that

$$\pi^{-1}(\bar{x}) \cap \mathcal{V}_\alpha = \{(\bar{x}, \bar{p}_\alpha, \bar{z}_\alpha)\} \tag{1.14}$$

and hence that $\alpha \neq \beta \Rightarrow (\bar{x}, \bar{p}_\beta, \bar{z}_\beta) \notin \mathcal{V}_\alpha$. This shows that all points in $\pi^{-1}(\bar{x})$ are isolated; hence any compact subset of V can contain only a finite number of them. As V is a closed subset of $\mathbb{R}^{2n+1}$, it can be written as a countable union of compact subsets, and the result follows. □

In the special case where the map π_x is proper at $\bar{x}$, it is easily seen that the set A has to be finite. Setting $\mathcal{U} = \bigcap_{\alpha \in A} \mathcal{U}_\alpha$, we get the following corollary.

COROLLARY 1.3. *Assume moreover the map π_x is proper. Then $\mathcal{R}$ is open in $\mathbb{R}^n$, and for every point $\bar{x} \in \mathcal{R}$ there is a neighborhood $\mathcal{U}$ of $\bar{x}$ and a (possibly empty) finite family of smooth functions f_α: $\mathcal{U} \to \mathbb{R}$, $\alpha \in A$, such that*

$$\pi_x^{-1}(\mathcal{U}) = \bigcup_{\alpha \in A} \{(x, f'_\alpha(x), f_\alpha(x)) | x \in \mathcal{U}, \alpha \in A\}. \tag{1.15}$$

We now have a description of $\pi_x^{-1}(\bar{x})$ which is valid whenever $\bar{x} \in \mathcal{R}$, i.e. for almost every point $\bar{x} \in \mathbb{R}^n$. Points in $\mathbb{R}^n \setminus \mathcal{R}$ form a negligible subset, but they may nevertheless turn out to be important, so we will attempt a partial description in that case also.

PROPOSITION 1.4. *Let $t \mapsto (x(t), p(t), z(t))$ be a C^1 map from $]0, T]$ into V such that $x(t) \in \mathcal{R}$ $\forall t > 0$. Assume that, when $t \to 0$,*

$$x(t) \to \bar{x} \quad \text{and} \quad \frac{dx}{dt}(t) \to \xi, \tag{1.16}$$

$$z(t) \to \bar{z}, \tag{1.17}$$

$$\liminf \|p(t) - \bar{p}\| = 0, \tag{1.18}$$

with $(\bar{x}, \bar{p}, \bar{z})$ an isolated point of $\pi_{xz}^{-1}(\bar{x}, \bar{z})$. Then

$$p(t) \to \bar{p}, \tag{1.19}$$

$$\frac{dz}{dt}(t) \to \bar{p} \cdot \xi. \tag{1.20}$$

Proof. As $\bar{p}$ is an isolated point in $\pi_{xz}^{-1}(\bar{x}, \bar{z})$, there is a compact neighborhood $\mathcal{W}$ of $(\bar{x}, \bar{p}, \bar{z})$ in V such that

$$(\bar{x}, p, \bar{z}) \in \mathcal{W} \Rightarrow p = \bar{p}. \tag{1.21}$$

Assume $p(t)$ does not converge to $\bar{p}$. Then there is an open neighborhood $\mathcal{V}$ of $(\bar{x}, \bar{p}, \bar{z})$, contained in $\mathcal{W}$, and a sequence $t_n \to 0$ such that

$$(x(t_n), p(t_n), z(t_n)) \in \mathcal{W} \setminus \mathcal{V}. \tag{1.22}$$

Using (1.16) and (1.17), together with the fact that $\mathcal{W} \setminus \mathcal{V}$ is compact, we can extract a subsequence converging to some point

$$(\bar{x}, p', \bar{z}) \in \mathcal{W} \setminus \mathcal{V} \tag{1.23}$$

contradicting (1.21).

So $p(t)$ has to converge to $\bar{p}$, yielding (1.19). Setting $z(0) = \bar{z}$, we define a continuous real-valued function $t \mapsto z(t)$ on $[0, T]$. It follows from Proposition 1.2 and the fact that $x(t) \in \mathcal{R}$ for $t > 0$ that this function is derivable on $]0, T]$ with derivative:

$$\frac{dz}{dt}(t) = p(t)\frac{dx}{dt}(t). \tag{1.24}$$

When $t \to 0$, the right-hand side converges to $\bar{p} \cdot \xi$, and so does the left-hand side. □

Note that $(dp/dt)(t)$ need not converge. Note also that (1.16) and (1.20) imply that $(d^+x/dt)(0)=\xi$ and $(d^+z/dt)(0)=\bar{p}\cdot\xi$, with d^+/dt denoting the right-derivative. Equation (1.20) can be written

$$\frac{d^+z}{dt}(0)=\bar{p}\cdot\frac{d^+x}{dt}(0) \tag{1.25}$$

which expresses the vanishing of $dz-p\,dx$ above a point $\bar{x}$ not in $\mathcal{R}$.

Let us give a more accurate picture in a simple case:

PROPOSITION 1.5. *Assume π_x is proper and $\pi_x^{-1}(\bar{x})$ is finite. Let a simply connected subset Ω of $\mathcal{R}$ be given in the following way*:

$$\Omega=\{\bar{x}+t\xi \mid 0<t<a,\ \xi\in S\} \tag{1.26}$$

with S an open subset of the unit sphere $\xi_1^2+\cdots+\xi_n^2=1$. There is a (possibly empty) finite family of C^1 functions $f_\alpha\colon \Omega\cup\{\bar{x}\}\to\mathbb{R}$, $\alpha\in A$, such that

$$\pi^{-1}(\Omega\cup\{\bar{x}\})=\{(x, f'_\alpha(x), f_\alpha(x)) \mid x\in\Omega\cup\{\bar{x}\},\ \alpha\in A\}. \tag{1.27}$$

By a derivative of f_α at $\bar{x}$ we mean a linear functional $f'_\alpha(\bar{x})$ such that

$$\begin{aligned}&\forall\varepsilon>0,\ \exists\,\eta>0\colon \|x-\bar{x}\|\leqq\eta \quad\text{and}\quad x\in\Omega\\ &\Rightarrow |f_\alpha(x)-f_\alpha(\bar{x})-\langle f'_\alpha(\bar{x}),\, x-\bar{x}\rangle|\leqq\varepsilon\|x-\bar{x}\|.\end{aligned} \tag{1.28}$$

By a C^1 function on $\Omega\cup\{\bar{x}\}$ we mean a function f_α such that $f'_\alpha(x)$ is well-defined and continuous on $\{\bar{x}\}\cup\Omega$.

Proof of Proposition 1.5. The set $\pi_x^{-1}(x)$ has to be both compact (because π_x is proper) and discrete (because $x\in\mathcal{R}$), so it is finite. By Proposition 1.2, the map $\pi_x\colon \pi_x^{-1}(\Omega)\to\Omega$ is a covering. As Ω is simply connected, the restriction of π_x to each connected component of $\pi_x^{-1}(\Omega)$ is a diffeomorphism, hence the representation formula

$$\pi^{-1}(\Omega)=\{(x, f'_\alpha(x), f_\alpha(x)) \mid x\in\Omega,\ \alpha\in A\}. \tag{1.29}$$

Now fix $\alpha\in A$ and let x converge to $\bar{x}$ in Ω. As π_x is proper, $(x, f'_\alpha(x), f_\alpha(x))$ has cluster points $(\bar{x}, p, z)\in\pi_x^{-1}(\bar{x})$. As this set is finite, all its points are isolated. As in the preceding proof, we conclude that $f'_\alpha(x)\to p_\alpha$ and $f_\alpha(x)\to z_\alpha$. Setting $f_\alpha(\bar{x})=z_\alpha$ and $f'_\alpha(\bar{x})=p_\alpha$, we get a C^1 function as desired. □

Let us conclude this investigation of Lagrangian submanifolds by the following remark, which throws some light on the case where $\pi_x^{-1}(\bar{x})$ is not discrete. Let $t\to(x(t), p(t), z(t))$ be a C^1 path drawn on V along which $x(t)$ is constant: $x(t)=\bar{x}$, $0\leqq t\leqq T$. Then $z(t)$ has to be constant also: $z(t)=\bar{z}$, $0\leqq t\leqq T$, so in fact only $p(t)$ varies. This follows easily from the vanishing of $i_V^*\omega$, which yields in this case $(dz/dt)(t)=\sum_{i=1}^n p_i(dx_i/dt)(t)$. In particular, if $\mathcal{V}$ is an open path-connected subset of V projecting on $\bar{x}$, i.e. $\mathcal{V}\subset\pi_x^{-1}(\bar{x})$, then $\mathcal{V}$ is also contained in some hyperplane $H=\{(x, p, z) \mid x=\bar{x},\ z=\bar{z}\}$ as an open path-connected subset (openness follows from the fact that $\dim V=n=\dim H$).

2. The Legendre transformation. The mapping $\mathscr{L}$ of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ into itself defined by

$$\begin{aligned} &\mathscr{L}(x, p, z) = (x', p', z'), \\ &x' = p, \quad p' = x, \quad z' = px - z \end{aligned} \tag{2.1}$$

is called the *Legendre transformation*. Note the following.

PROPOSITION 2.1. *The Legendre transformation is a C^∞ involution*:

$$\mathscr{L}^2 = Id. \tag{2.2}$$

Proof. Using notations (2.1), we set $\mathscr{L}(x', p', z') = (x'', p'', z'')$, with

$$\begin{aligned} x'' &= p' = x, \\ p'' &= x' = p, \\ z'' &= p'x' - z' = px - (px - z) = z; \end{aligned}$$

hence we get the result. □

The fundamental fact about the Legendre transformation is that it preserves the 1-form ω, up to a change of sign.

THEOREM 2.2. $\mathscr{L}^*\omega = -\omega$.

Proof. Using notations (2.1), we get

$$\begin{aligned} \mathscr{L}^*\omega &= dz' - p'\,dx' \\ &= (x\,dp + p\,dx - dz) - x\,dp \\ &= p\,dx - dz \\ &= -\omega. \quad \square \end{aligned}$$

COROLLARY 2.3. *If V is a Lagrangian submanifold of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$, then so is $\mathscr{L}V$.*

Proof. It follows from Proposition 2.1 that $\mathscr{L}$ is a diffeomorphism of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ onto itself. Hence $\mathscr{L}V$ is a closed submanifold whenever V is. There only remains to check that $i^*_{\mathscr{L}V}\omega = 0$. To do that, we write the following diagram:

$$\begin{array}{ccc} V & \xrightarrow{i} & \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \\ {\scriptstyle l}\downarrow & & \downarrow{\scriptstyle \mathscr{L}} \\ \mathscr{L}V & \xrightarrow{j} & \mathbb{R}^n \times \mathbb{R} = \mathbb{R} \end{array} \tag{2.3}$$

where l is the restriction of $\mathscr{L}$ to V and j is the canonical injection. This diagram commutes, and gives rise to another commutative diagram relating 1-forms:

$$\begin{array}{ccc} T^*V & \xleftarrow{i^*} & T^*(\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}) \\ {\scriptstyle l^*}\uparrow & & \uparrow{\scriptstyle \mathscr{L}^*} \\ T^*(\mathscr{L}V) & \xleftarrow{j^*} & T^*(\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}). \end{array} \tag{2.4}$$

Taking ω in the lower right-hand corner, and using formula (1.3) and Theorem 2.2, we get

$$i^* \circ \mathscr{L}^*(\omega) = i^*(-\omega) = -i^*(\omega) = 0; \tag{2.5}$$

going the other way around the diagram, we get

$$0 = l^* \circ j^*(\omega). \tag{2.6}$$

As l is a diffeomorphism, l^* is an isomorphism, and (2.6) implies that $j^*\omega = 0$, i.e. $\mathscr{L}V$ is Lagrangian. □

We now introduce a slight misuse of notations. Let V and W be Lagrangian submanifolds of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$, with $W = \mathscr{L}V$, and let F_V and F_W be the associated characteristic maps. We shall write freely $F_W = \mathscr{L}F_V$, and call F_W the Legendre transform of F_V. For instance, if $f: \mathbb{R}^n \to \mathbb{R}$ is a C^∞ function, then $\mathscr{L}f$ is the multivalued map from $\mathbb{R}^n$ to $\mathbb{R}$ defined by

$$\mathscr{L}f(x') = \{z' | \exists\, p' \in \mathbb{R}^n : (x', p', z') \in \mathscr{L}V_f\}. \tag{2.7}$$

Using (1.1) and (2.1), we get

$$\mathscr{L}f(p) = \{px - f(x) | f'(x) = p\}. \tag{2.8}$$

Several remarks are now in order. First of all, if f, in addition to being smooth, is *convex*, then the function $x \mapsto px - f(x)$ is concave, and the equation $p = f'(x)$ simply means that this function attains its maximum at x. Equation (2.8) then becomes

$$\mathscr{L}f(p) = \max\{px - f(x) | x \in \mathbb{R}^n\}. \tag{2.9}$$

Formula (2.9) shows that $\mathscr{L}f$ is single- or possibly empty-valued. In other words, $\mathscr{L}f$ is a real-valued function defined on some subset of $\mathbb{R}^n$. It is to be compared with the classical Fenchel transform of convex analysis:

$$f^*(p) = \sup\{px - f(x) | x \in \mathbb{R}^n\}. \tag{2.10}$$

Formulas (2.9) and (2.10) coincide whenever the function $x \to px - f(x)$ attains its maximum over $\mathbb{R}^n$. Define the effective domain of f^* as the set of points where it is finite:

$$\operatorname{dom} f^* = \{p | f^*(p) < \infty\}. \tag{2.11}$$

PROPOSITION 2.4. *$\mathscr{L}f(p) = f^*(p)$ if and only if f^* is subdifferentiable at p, i.e. $\partial f^*(p) \neq \phi$. This is the case at every interior point p of* $\operatorname{dom} f^*$:

$$p \in \operatorname{int} \operatorname{dom} f^* \Rightarrow \mathscr{L}f(p) = f^*(p). \tag{2.12}$$

Proof. Let us write down the definition of the subdifferential of f^*:

$$\partial f^*(p) = \{\bar{x} \in \mathbb{R}^n | p\bar{x} - f^{**}(\bar{x}) = \max_x\} \tag{2.13}$$

where the notation $= \max_x$ means that the left-hand side attains its maximum at $\bar{x}$.

But, as f is continuous and convex, it coincides with its biconjugate f^{**}; hence

$$\partial f^*(p) = \{\bar{x} \in \mathbb{R}^n \mid p\bar{x} - f(\bar{x}) = \max_x\} \tag{2.14}$$

which proves the first part of the proposition.

It is a well-known fact from convex analysis that any convex function on $\mathbb{R}^n$ is continuous, and hence subdifferentiable, on the interior of its effective domain. Hence we have (2.12). □

In the general (smooth, nonconvex) case, formula (2.8) sets $\mathscr{L}f(p)$ in one-to-one correspondence with the sets of tangents to f having slope p.

PROPOSITION 2.5. *$z' \in \mathscr{L}f(p)$ if and only if $z = px - z'$ is a tangent hyperplane to graph f in $\mathbb{R}^n \times \mathbb{R}$.*

Proof. The hyperplane $z = px - z'$ in (x, z)-space is tangent to graph f if and only if there exists $\bar{x} \in \mathbb{R}^n$ such that $f'(\bar{x}) = p$ and $f(\bar{x}) = p\bar{x} - z'$. This reduces to $z' \in \mathscr{L}f(p)$ by (2.8). □

From Proposition 2.5 one sees instantly that $\mathscr{L}f$ can be multivalued. Indeed $\mathscr{L}f$ is a function, i.e. $\mathscr{L}f(p)$ is empty or a singleton for every p, if and only if f has only zero or one tangent of prescribed slope. In dimension $n = 1$, this means exactly that f is convex. In higher dimensions, this also happens in the nonconvex case: take for instance $f(x_1, x_2) = x_1^2 - x_2^2$; then f': $(x_1, x_2) \mapsto (2x_1, -2x_2)$ is one-to-one. But the fact remains that, in contrast with the convex case, in the general case we have to deal with multivalued Legendre transforms. So let us attempt a description of $\mathscr{L}f$. We denote by V the Lagrangian submanifold (1.1) of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ associated with f, and by $A(x)$ the matrix of second derivatives of f at x:

$$A(x) = \left(\left(\frac{\partial^2 f}{\partial x_i\, \partial x_j}(x)\right)\right), \qquad 1 \leqq i, j \leqq n. \tag{2.15}$$

PROPOSITION 2.6. *Assume $A(\bar{x})$ has full rank n. Then there exists a neighborhood $\mathscr{V}$ of $(f'(\bar{x}), \bar{x}, \bar{x}f'(\bar{x}) - f(\bar{x}))$ in $\mathscr{L}V$ projecting onto a neighborhood $\mathscr{U}$ of $f'(\bar{x})$ in $\mathbb{R}^n$, and a local inverse φ for f' such that*

$$\mathscr{V} = \{(p, [\mathscr{L}_{\mathscr{V}} f]'(p), [\mathscr{L}_{\mathscr{V}} f](p)) \mid p \in \mathscr{U}\} \tag{2.16}$$

with $[\mathscr{L}_{\mathscr{V}} f](p) = p\varphi(p) - f \circ \varphi(p)$. In particular, we have

$$[\mathscr{L}_{\mathscr{V}} f]'(\bar{p}) = \bar{x}. \tag{2.17}$$

Proof. It follows from the implicit function theorem that the map $x \mapsto f'(x)$ has a local inverse φ defined on some neighborhood $\mathscr{U}$ of $\bar{p}$. Setting

$$\mathscr{V} = \{(f'(x), x, xf'(x) - f(x)) \mid x \in \varphi(\mathscr{U})\} \tag{2.18}$$

and using the definition of φ, we get

$$\mathscr{V} = \{(p, \varphi(p), p\varphi(p) - f \circ \varphi(p)) \mid p \in \mathscr{U}\}. \tag{2.19}$$

Computing the derivative of $\mathscr{L}_\varphi f$, we get

$$\begin{aligned}[\mathscr{L}_\varphi f]'(p) &= \varphi(p) + {}^t\varphi'(p)p - {}^t\varphi'(p) f' \circ \varphi(p) \\ &= (p) + {}^t\varphi'(p)p - {}^t\varphi'(p)p \\ &= \varphi(p)\end{aligned} \tag{2.20}$$

and formula (2.19) reduces to (2.16). □

$\mathscr{L}_\varphi f$ is a smooth branch of $\mathscr{L}f$ lying above $\bar{p}$. Note that $\bar{p}$ is a regular value for $f'\colon \mathbb{R}^n \to \mathbb{R}^n$ if and only if it is a regular value for $\pi_x\colon \mathscr{L}V \to \mathbb{R}^n$. This is almost always the case, by Sard's theorem, and the part of $\mathscr{L}f$ lying above $\bar{p}$ then is a countable union of smooth branches such as $\mathscr{L}_\varphi f$ (this is a particular case of Proposition 1.2). If moreover f' is proper at $\bar{p}$, then so is π_x, and there are only a finite number of branches of $\mathscr{L}f$ lying above $\bar{p}$ (this is a particular case of Corollary 1.3).

We can of course apply Propositions 1.4 and 1.5 to get a description of $\mathscr{L}f$ above critical values of f'. But, in this particular case, we prefer another approach, which has the advantage of directly relating the shape of the Legendre transform above $f(\bar{x})$ to the degeneracy of the matrix of second derivatives at $\bar{x}$. We write the Taylor expansion of f at $\bar{x}$:

$$f(\bar{x}+\xi) = f(\bar{x}) + \bar{p}\xi + \tfrac{1}{2}\langle A(\bar{x})\xi, \xi\rangle + \tfrac{1}{6}P_3(\bar{x}; \xi_1, \cdots, \xi_n) + O(|\xi|^4) \tag{2.21}$$

where $P_3(\bar{x}; \cdot)$ is a homogeneous polynomial of degree 3 in n variables. Using the Euler formula, we may write

$$P_3(\bar{x}; \xi_1, \cdots, \xi_n) = \tfrac{1}{3}\sum_{i=1}^{n} \xi_i \frac{\partial P_3}{\partial \xi_i}(\bar{x}; \xi_1, \cdots, \xi_n) = \sum_{i=1}^{n} \xi_i \langle B_i(\bar{x})\xi, \xi\rangle$$

where $B_i(\bar{x})$ is the matrix with elements $\frac{1}{3}\,\partial^3 f/\partial x_i\,\partial x_j\,\partial x_k$, $1 \leqq j, k \leqq n$. Denote by $\langle B(\bar{x})\xi, \xi\rangle$ the n-vector with components $\langle B_i(\bar{x})\xi, \xi\rangle$.

PROPOSITION 2.7. *Assume that $A(\bar{x})$ has rank $(n-1)$ and that*

$$\xi \neq 0, \qquad \xi \in \operatorname{Ker} A(\bar{x}) \Rightarrow \begin{cases} P_3(x; \xi_1, \cdots, \xi_n) \neq 0, \\ \langle B(\bar{x})\xi, \xi\rangle \notin \operatorname{Im} A(\bar{x}). \end{cases} \tag{2.22}$$

Then (possibly after reordering the linear coordinates $(p_1, \cdots, p_n)$ in $\mathbb{R}^n$ and changing p_n to $-p_n$) there is a neighborhood $\mathscr{V}$ of $(f'(\bar{x}), \bar{x}, f'(\bar{x})\bar{x} - f(\bar{x}))$ in $\mathscr{L}V$, a neighborhood $\mathscr{U} = \mathscr{U}' \times \mathscr{U}_n$ of $(\bar{p}_1, \cdots, \bar{p}_{n-1}, \bar{p}_n)$ in $\mathbb{R}^n$, C^∞ functions $k_1, k_2\colon \mathscr{U}' \to \mathbb{R}$ and $h\colon \mathscr{U} \to \mathbb{R}$, such that $\pi_{xz}\mathscr{V}$ is completely described by the set of conditions

$$(p_1, \cdots, p_{n-1}, p_n) \in \mathscr{U}' \times \mathscr{U}_n \quad \text{and} \quad p_n \geqq k_1(p_1, \cdots, p_{n-1}), \tag{2.23}$$

$$z \in \{z_+(p), z_-(p)\}, \tag{2.24}$$

with

$$z_+(p) = k_2(p_1, \cdots, p_{n-1}) + (p_n - k_1)h(p_1, \cdots, p_{n-1}, \sqrt{p_n - k_1}),$$

$$z_-(p) = k_2(p_1, \cdots, p_{n-1}) + (p_n - k_1)h(p_1, \cdots, p_{n-1}, -\sqrt{p_n - k_1}).$$

Moreover $\partial z/\partial p_i = x_i$, $1 \leqq i \leqq n$, along the hypersurface

$$p_n = k_1(p_1, \cdots, p_{n-1}). \tag{2.25}$$

Proof. The $(x_1, \cdots, x_n)$ are a system of coordinates in $\mathscr{L}V$ with formula (2.8) yielding $(p_1, \cdots, p_n, z)$ in terms of $(x_1, \cdots, x_n)$. In particular,

$$\frac{\partial f}{\partial x_i}(x) = p_i \quad \text{for } 1 \leqq i \leqq n. \tag{2.26}$$

The rank assumption on the matrix $A(\bar{x})$ implies that one of its $(n-1)\times(n-1)$ minors is invertible, for instance the one defined by the $(n-1)$ first rows and the $(n-1)$ first columns. Moreover, the nth row then is a linear combination of the $(n-1)$ first rows.

It follows from the implicit function theorem that the $(n-1)$ first equations of system (2.26) can be solved locally for $(x_1, \cdots, x_{n-1})$. In other words, $(p_1, \cdots, p_{n-1}, x_n)$ can be used as coordinates in some neighborhood $\mathscr{V}_1$ of $(\bar{p}, \bar{x}, \bar{z})$ in $\mathscr{L}V$.[1] Now consider the path $w(t) = (p(t), x(t), z(t))$ in $\mathscr{V}_1$ such that $p_1(t) = \bar{p}_1, \cdots, p_{n-1}(t) = \bar{p}_{n-1}, x_n(t) = \bar{x}_n + t$. There is some $T > 0$ such that $w(t)$ is well-defined for $-T \leqq t \leqq T$. Obviously $w(0) = (\bar{p}, \bar{x}, \bar{p}\bar{x} - f(\bar{x}))$; we shall write ξ' for $(dx/dt)(0)$ and ξ'' for $(d^2x/dt^2)(0)$. Equations (2.26) are satisfied along $w(t)$:

$$p_i(t) = \frac{\partial f}{\partial x_i}(x_1(t), \cdots, x_n(t)) \quad \text{for } 1 - T \leqq t \leqq T. \tag{2.27}$$

Writing Taylor expansions into (2.27), we get

$$p(t) - \bar{p} = tA(\bar{x})\xi' + \frac{t^2}{2}[\langle B(\bar{x})\xi', \xi'\rangle + A(\bar{x})\xi''] + O(t^3). \tag{2.28}$$

But $p_i(t) - \bar{p}_i = 0$ for $1 \leqq i \leqq n-1$, so that both sides of the $(n-1)$ first equations of system (2.28) are identically zero on $(-T, T)$. It follows that the $(n-1)$ first components of $A(\bar{x})\xi$ are zero, and, by the rank assumption, so is the last one

$$A(\bar{x})\xi' = 0. \tag{2.29}$$

Assumption (2.22) then yields

$$\langle B(\bar{x})\xi', \xi'\rangle + A(\bar{x})\xi'' \neq 0. \tag{2.30}$$

But again, both sides of the $(n-1)$ first equations (2.28) being identically zero on $(-T, T)$, the $(n-1)$ first components of vector (2.30) must be zero. It follows that the nth component must be nonzero. We summarize our results so far by stating that the nth equation of system (2.28) can be written as

$$p_n(t) - \bar{p}_n = \tfrac{1}{2}a_n t^2 + O(t^3), \qquad a_n \neq 0. \tag{2.31}$$

Similarly, we compute the Taylor expansion of $z(t)$ at $t = 0$. By definition, we have

$$z(t) = f'[x(t)]x(t) - f[x(t)]. \tag{2.32}$$

[1] From now on we set $\bar{p} = f'(\bar{x})$ and $\bar{z} = \bar{p}\bar{x} - f(\bar{x})$.

Successive derivations yield

$$\frac{dz}{dt}(0) = \langle A(\bar{x})\xi', \xi' \rangle \tag{2.33}$$

$$\frac{d^2z}{dt^2}(0) = 2\langle A(\bar{x})\xi', \xi'' \rangle + P_3(\bar{x}; \xi_1', \cdots, \xi_n'). \tag{2.34}$$

But we have seen that $A(\bar{x})\xi' = 0$, so that $(dz/dt)(0) = 0$ and $(d^2z/dt^2)(0) = b_n \neq 0$ by assumption (2.22). Finally we get

$$z(t) - \bar{z} = \tfrac{1}{2}b_n t^2 + O(t^3), \qquad b_n \neq 0. \tag{2.35}$$

Now $w'(0)$ is just the tangent vector $(\partial/\partial x_n)(\bar{p}_1, \cdots, \bar{p}_{n-1}, x_n)$ associated with the new coordinate systems. In other words, p_n and z, considered as functions of $(p_1, \cdots, p_{n-1}, x_n)$ in $\mathcal{V}_1$, satisfy

$$\frac{\partial p_n}{\partial x_n}(\bar{p}_1, \cdots, \bar{p}_{n-1}, \bar{x}_n) = 0, \tag{2.36}$$

$$\frac{\partial^2 p_n}{\partial x_n^2}(\bar{p}_1, \cdots, \bar{p}_{n-1}, \bar{x}_n) \neq 0, \tag{2.37}$$

$$\frac{\partial z}{\partial x_n}(\bar{p}_1, \cdots, \bar{p}_{n-1}, \bar{x}_n) = 0, \tag{2.38}$$

$$\frac{\partial^2 z}{\partial x_n^2}(\bar{p}_1, \cdots, \bar{p}_{n-1}, \bar{x}_n) \neq 0. \tag{2.39}$$

But other points (p, x, z) in $\mathcal{V}_1$ enjoy the property that $A(x)$ is of rank $(n-1)$ and satisfies (2.22). Indeed, consider the Jacobian determinant

$$\begin{aligned} \Delta(p_1, \cdots, p_{n-1}, x_n) &= \frac{D(p_1, \cdots, p_{n-1}, p_n)^2}{D(p_1, \cdots, p_{n-1}, x_n)} \\ &= \frac{\partial p_n}{\partial x_n}(p_1, \cdots, p_{n-1}, x_n) \end{aligned} \tag{2.40}$$

by a simple computation. Clearly rank $A(x_1, \cdots, x_n) < n$ if and only if $\Delta(p_1, \cdots, p_{n-1}, x_n) = 0$. But $\Delta = 0$ and $(\partial\Delta/\partial x_n) = (\partial^2 p_n/\partial x_n^2) \neq 0$ at point $(\bar{p}_1, \cdots, \bar{p}_{n-1}, \bar{x}_n)$. By the implicit function theorem, there are neighborhoods $\mathcal{U}_1$ of $(\bar{p}_1, \cdots, \bar{p}_{n-1})$ and $\mathcal{W}_1$ of $\bar{x}_n$ and C^∞ map $g\colon \mathcal{U}_1 \to \mathcal{W}_1$ such that

$$\Delta(p_1, \cdots, p_{n-1}, x_n) = 0 \Leftrightarrow x_n = g(p_1, \cdots, p_{n-1}) \forall (p_1, \cdots, x_n) \in \mathcal{U}_1 \times \mathcal{W}_1. \tag{2.41}$$

Conversely, $x_n = g(p_1, \cdots, p_{n-1})$ implies rank $A(x_1, \cdots, x_n) < n$. By a continuity argument, we can shrink $\mathcal{U}_1$ and $\mathcal{W}_1$ to $\mathcal{U}_2$ and $\mathcal{W}_2$ so that rank $A(x_1, \cdots, x_n)$ is exactly $n-1$ and assumption (2.22) is satisfied whenever $x_n = g(p_1, \cdots, p_{n-1})$ in $\mathcal{U}_2 \times \mathcal{W}_2$. We may even include in the bargain the fact that the first minor of $A(x)$ is invertible, so that $(p_1, \cdots, p_{n-1}, x_n)$ enjoys all the properties of $(\bar{p}_1, \cdots, \bar{p}_{n-1}, \bar{x}_n)$. By (2.36) and (2.39), it follows that $\partial p_n/\partial x_n = 0$, $\partial^2 p_n/\partial x_n^2 \neq$

[2] Recall that $D(f_1, \cdots, f_n)/D(x_1, \cdots, x_n)$ denotes the determinant $\|\partial f_i/\partial x_j\|$.

0, $\partial z/\partial x_n = 0$, $\partial^2 z/\partial x_n^2 \neq 0$ at every point $(p_1, \cdots, p_{n-1}, x_n) \in \mathcal{U}_2 \times \mathcal{W}_2$ such that $x_n = g(p_1, \cdots, p_{n-1})$.

It follows that

$$p_n = k_1(p_1, \cdots, p_{n-1}) + [x_n - g(p_1, \cdots, p_{n-1})]^2 h_1(p_1, \cdots, p_{n-1}, x_n), \tag{2.42}$$

$$z = k_2(p_1, \cdots, p_{n-1}) + [x_n - g(p_1, \cdots, p_{n-1})]^2 h_2(p_1, \cdots, p_{n-1}, x_n) \tag{2.43}$$

with

$$x_n = g(p_1, \cdots, p_{n-1}) \Rightarrow h_1(p_1, \cdots, p_{n-1}, x_n) h_2(p_1, \cdots, p_{n-1}, x_n) \neq 0. \tag{2.44}$$

The point of V defined by $(p_1, \cdots, p_{n-1}, x_n = g(p_1, \cdots, p_n))$ yields $p_n = k_1(p_1, \cdots, p_{n-1})$ and $z = k_2(p_1, \cdots, p_{n-1})$, so that k_1 and k_2 are C^∞ functions. It follows from the C^∞ division theorem of Malgrange that h_1 and h_2 can be chosen to be C^∞ functions also.

Assume that $h_1(\bar{p}_1, \cdots, \bar{p}_{n-1}, \bar{x}_n) > 0$. Then we can define $y_n = [x_n - g]\sqrt{h_1}$ and use $(p_1, \cdots, p_{n-1}, y_n)$ as a new system of local coordinates in some smaller neighborhood $\mathcal{V}_2$ of $(\bar{p}, \bar{x}, \bar{z})$ corresponding to $(p_1, \cdots, p_{n-1}, y_n) \in \mathcal{U}_3 \times \mathcal{W}_3$. Equations (2.42) and (2.43) become

$$p_n - k_1(p_1, \cdots, p_{n-1}) = y_n^2, \tag{2.45}$$

$$z - k_2(p_1, \cdots, p_{n-1}) = y_n^2 h_3(p_1, \cdots, p_{n-1}, y_n) \tag{2.46}$$

with $(p_1, \cdots, p_{n-1}) \in \mathcal{U}_3$ and $y_n \in \mathcal{W}_3$. This implies that $p_n - k_1$ is nonnegative. Conversely, whenever $p_n \geqq k_1$, we can solve (2.45) by $y_n = \pm\sqrt{p_n - k_1}$, getting two distinct values whenever the inequality is strict; possibly shrinking $\mathcal{U}_3$ to $\mathcal{U}_4$, we can arrange that both those values are in $\mathcal{W}_3$, so that (2.46) becomes

$$z - k_2 = (p_n - k_1) h(p_1, \cdots, p_{n-1}, \pm\sqrt{p_n - k_1}) \tag{2.47}$$

which, together with $(p_1, \cdots, p_{n-1}) \in \mathcal{U}_4$, completely describes $\pi_{xz}\mathcal{V}_2$.

If $h_1(\bar{p}_1, \cdots, \bar{p}_{n-1}, \bar{x}_n)$ should be negative, then we simply reverse p_n to $-p_n$, and we are back to the preceding case. So formulae (2.23) and (2.24) are proved.

For the sake of convenience, denote by Ω the set of points $(p_1, \cdots, p_n)$ such that $p_n > k_1$, and by Σ its boundary, the equation of which is $p_n = k_1$. Formula (2.24) yields along Σ

$$\begin{aligned} &\frac{\partial z_+}{\partial p_i} = \frac{\partial z_-}{\partial p_i} = \frac{\partial k_2}{\partial p_i}, && 1 \leqq i \leqq n-1, \\ &\frac{\partial z_+}{\partial p_n} = \frac{\partial z_-}{\partial p_n} = h. \end{aligned} \tag{2.48}$$

It follows also from formula (2.24) that with any $p \in \Sigma$ and any vector $\pi' = (\pi'_1, \cdots, \pi'_n)$ pointing to the interior of Ω (i.e. $\pi'_n - \sum_{i=1}^{n-1} (\partial k_1/\partial p_i)\pi'_i > 0$) we can associate two continuous paths $t \to (p(t), x(t), z_+(t))$ and $t \to (p(t), x(t), z_-(t))$ in $\mathcal{L}V$ starting at (p, x, z) and satisfying $(dp/dt)(t) \to \pi'$ as $t \to 0$. From Proposition 1.4 (taking care that x- and p-coordinates are interchanged) it follows that, when $t \to 0$,

$$\frac{dz_+}{dt}(t) \to x \cdot \pi' \quad \text{and} \quad \frac{dz_-}{dt}(t) \to x \cdot \pi'. \tag{2.49}$$

But from formula (2.48) we get directly

$$\frac{dz_+}{dt}(t) \to \frac{\partial z}{\partial p} \cdot \pi' \quad \text{and} \quad \frac{\partial z_-}{\partial t}(t) \to \frac{\partial z}{\partial p} \cdot \pi' \tag{2.50}$$

where $\partial z/\partial p$ denotes the common value of the n-vectors (2.48). This yields $(\partial z/\partial p) \cdot \pi' = x \cdot \pi'$ for every vector π' in some half-space, and hence the desired formula $x = \partial z/\partial p$. □

In other words, $\mathscr{L}f$ is not defined locally for $p_n < k_1(p_1, \cdots, p_{n-1})$. In the region $p_n \geqq k_1(p_1, \cdots, p_{n-1})$, there are two well-defined branches for $\mathscr{L}f$. Along the boundary they coincide and have the same tangent hyperplane, and their shape away from the boundary is given by the following result.

COROLLARY 2.8. *We keep the assumptions and notations of Proposition* 2.7, *and we set* $q_n = p_n - k_1(p_1, \cdots, p_{n-1})$. *Then* $\mathscr{L}f$ *can be expanded near the boundary* $q_n = 0$ *as*

$$z = k_2(p_1, \cdots, p_{n-1}) + q_n[a_0(p_1, \cdots, p_{n-1}) \pm a_1(p_1, \cdots, p_{n-1})\sqrt{q_n}] + O(q_n^{3/2}) \tag{2.51}$$

where the functions k_2, a_0, a_1 *are* C^∞. *Moreover*

$$\frac{\partial k_2}{\partial p_i}(p_1, \cdots, p_{n-1}) = x_i \quad \text{for } 1 \leqq i \leqq n-1, \tag{2.52}$$

$$a_0(p_1, \cdots, p_{n-1}) = x_n. \tag{2.53}$$

The proof consists simply of replacing h by its Taylor expansion in formula (2.24). We see that the two branches only intersect at the boundary $p_n = k_1$ of the admissible domain $p \geqq k_1$ (this is true even in the special case where $a_1 = 0$, because then the third order term $\pm a_3 q_n^{3/2}$ takes precedence). This is the classical "cusp" situation, so that Proposition 2.7 can be loosely stated as follows: a simple inflection point of f gives rise to a simple cusp of $\mathscr{L}f$.

Of course, more degenerate inflection points of f give rise to more complicated situations in $\mathscr{L}f$. A classification can be attempted along the lines of Proposition 2.7, but we are not going to conduct it any further. Let us only point out that, for all functions $f \in \mathscr{F}$, where $\mathscr{F}$ is a dense G_δ subset of $C^\infty(\mathbb{R}^n)$ in the Whitney topology, the space $\mathbb{R}^n$ can be partitioned as $\Sigma_0 \cup \Sigma_1 \cup \Sigma_2$ where:

Σ_0 consists of all points x where $A(x)$ is nondegenerate; it is an open subset of $\mathbb{R}^n$.

Σ_1 consists of all points x where $A(x)$ has rank $(n-1)$ and satisfies (2.22); it is a codimension one submanifold.

Σ_2 consists of all other points; it is a stratified subset of codimension $\geqq 2$.

Without going into details, this follows from Thom's transversality theorems. So, for most functions, the analysis performed thus far describes everything up to codimension two. In the one-dimensional case, $n = 1$, that means precisely everything. Let us conclude by a simple example.

Define a function f on the real line by

$$f(x) = (x + x^2)^2. \tag{2.54}$$

We want to know what $\mathscr{L}f$ looks like. We need some data on f which are summarized in the following:

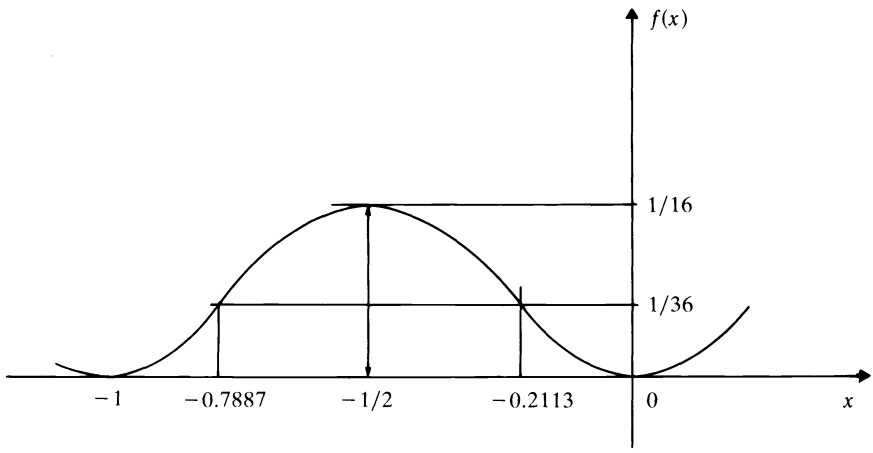
$$f'(x) = 4x(x+1)(x+\tfrac{1}{2}) = 4x^3+6x^2+2x,$$

$$f''(x) = 12x^2+12x+2,$$

x	$f(x)$	$p=f'(x)$	$f''(x)$	$z=f'(x)x-f(x)$
$-\infty$	$+\infty$	$-\infty$	*	$+\infty$
-1	0	0	*	0
-0.7887	$\frac{1}{36}$	0.19245	0	-0.1796
$-\frac{1}{2}$	$\frac{1}{16}$	0	*	$-\frac{1}{16}$
-0.2113	$\frac{1}{36}$	-0.19245	0	0.0129
0	0	0	*	0
$+\infty$	$+\infty$	$+\infty$	*	$+\infty$

We now can draw the graphs of f and $\mathscr{L}f$ (Figs. 1 and 2.) Note that the z-axis $p=0$ intersects $\mathscr{L}f$ at the simple point $z=-\frac{1}{16}$ and the double point $z=0$. This means that there are two distinct tangents to f with slope $p=0$: the first one is tangent to f at $x=-\frac{1}{2}$ only, the second one is tangent to f both at $x=-1$ and $x=0$. From formula (2.17), the tangent to $\mathscr{L}f$ at $(p=0, z=-\frac{1}{16})$ has slope $-\frac{1}{2}$, and the two branches of $\mathscr{L}f$ which intersect at $(p=0, z=0)$ have distinct tangents of slopes -1 and 0 respectively.

Moreover $\mathscr{L}f$ features two cusps at (0.1945, -0.1796) and (-0.1945, 0.0129). By Proposition 2.7, the tangents at those cusps are well-defined, and have slopes -0.7887 and -0.2113 respectively.



0.01

FIG. 1. $x \mapsto f(x)$. *Scale*: ↑ ⟶ 0.1

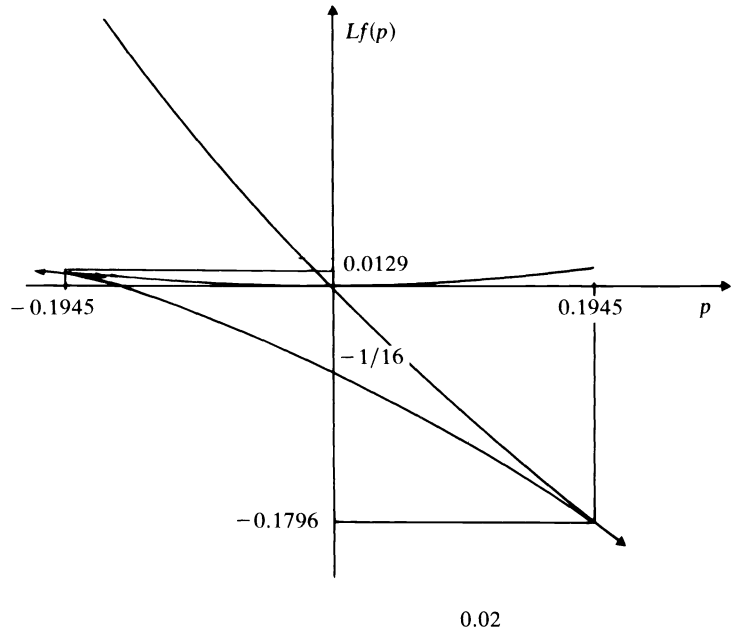


FIG 2. $p \mapsto \mathscr{L}f(p)$. *Scale*: $\uparrow$ 0.02, $\rightarrow$ 0.02

Note the parameteric equations for $\mathscr{L}f$:

$$\begin{aligned} p &= 2x(x+1)(2x+1), \\ z &= x(x+1)(3x^2+x). \end{aligned} \tag{2.55}$$

Thus the graph of $\mathscr{L}f$ is the semi-algebraic set obtained by writing that the two algebraic equations (2.55) have a common solution in x, i.e. by eliminating x between the two equations.

3. Extremization problems and duality. Whenever V is a subset of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$, we shall denote by

$$\text{ext } V \tag{P}$$

the problem of determining all couples $(x, z) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$(x, 0, z) \in V. \tag{3.1}$$

$(\mathscr{P})$ will be termed an *extremization problem*, and any couple (x, z) satisfying (3.1) will be called a *solution* of $(\mathscr{P})$. The value of $(\mathscr{P})$, denoted by $\{\text{ext } \mathscr{P}\}$, will be the set of all $z \in \mathbb{R}$ such that there is an $x \in \mathbb{R}^n$ with (3.1) satisfied.

An important special case occurs when V is the Lagrangian submanifold associated with some C^∞ function f: $\mathbb{R}^n \to R$

$$V = \{(x, f'(x), f(x)) | x \in \mathbb{R}^n\}. \tag{3.2}$$

In that case formula (3.1) becomes

$$f'(x) = 0, \qquad z = f(x) \tag{3.3}$$

so that $(\mathscr{P})$ is simply the problem of determining the critical points and values of f. We shall write it

$$(\mathscr{P}) \qquad \operatorname*{ext}_x f(x)$$

and call it an *unconstrained smooth extremization problem.*

Another important special case occurs when

$$V=\left\{\left(x, f'(x)-\sum_{j=1}^{k} \lambda_j g_j'(x), f(x)\right) \middle| g_j(x)=0, \lambda_j \in \mathbb{R}, 1 \leqq j \leqq k\right\} \tag{3.4}$$

where f and the g_j, $1 \leqq j \leqq k$, are C^∞ functions on $\mathbb{R}^n$. We set

$$S=\pi_x V=\{x \mid g_j(x)=0, 1 \leqq j \leqq k\}. \tag{3.5}$$

LEMMA 3.1. *If the $g_j'(x)$, $1 \leqq j \leqq k$, are linearly independent at every $x \in S$, $x \in S$, then S is a closed submanifold of $\mathbb{R}^n$ and V is a Lagrangian submanifold of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$.*

Proof. The fact that S and V are closed $(n-k)$- and n-dimensional submanifolds follows easily from the implicit function theorem. We check condition (1.3) for V:

$$\begin{aligned} i_V^* \omega &= df(x)-(f'(x)-\textstyle\sum \lambda_j g_j'(x))\, dx \\ &= (df(x)-f'(x)\, dx)+\textstyle\sum \lambda_j g_j'(x)\, dx. \end{aligned} \tag{3.6}$$

The first term vanishes identically, and along V we have $g_j'(x)\, dx = 0$ since $g_j(x)$ is a constant. □

The solutions of $(\mathscr{P})$ are all couples $(x, f(x))$ such that

$$x \in S \quad \text{and} \quad \exists \lambda_1, \cdots, \lambda_k : f'(x)-\sum_{j=1}^{k} \lambda_j g_j'(x)=0. \tag{3.7}$$

If the $g_j'(x)$, $1 \leqq j \leqq k$, are linearly independent at every point $x \in S$, condition (3.7) means that x is a critical point of $f|_S$, the restriction of f to S. For that reason, we shall write $(\mathscr{P})$ as

$$(\mathscr{P}) \qquad \begin{aligned} &\operatorname{ext} f(x), \\ &g_j(x)=0, \qquad 1 \leqq j \leqq k, \end{aligned}$$

and call it a *constrained smooth extremization problem.*

Any critical point of a smooth convex (or concave) function is a minimum (or a maximum). For that reason, the various extremization problems we stated reduce to optimization problems when f is convex (or concave) and the g_j linear. So extremization is a natural generalization of optimization to the nonconvex case. Now it is a well-known fact that there is a duality theory of convex optimization problems, and we want to extend it to nonconvex extremization problems.

From now on we are given a linear map $A: \mathbb{R}^n \to \mathbb{R}^m$. We shall denote by x, p, y, q the vectors of $\mathbb{R}$, $(\mathbb{R}^n)^*$, $\mathbb{R}^m$, $(\mathbb{R}^m)^*$ respectively. With any subset V of

$\mathbb{R}^{n+m} \times \mathbb{R}^{n+m} \times \mathbb{R}$ we associate the subset V_A of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ defined by

$$V_A = \{(x, p + A^*q, z) | (x, Ax; p, q; z) \in V\}. \tag{3.8}$$

Applying this definition to the transpose $A^*: (\mathbb{R}^m)^* \to (\mathbb{R}^n)^*$,[3] and to any subset V^* of $\mathbb{R}^{n+m} \times \mathbb{R}^{n+m} \times \mathbb{R}$, we get

$$V^*_{A^*} = \{(q, y + Ax, z) | (A^*q, q; x, y; z) \in V^*\} \subset \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}. \tag{3.9}$$

We now state the main result of this section.

THEOREM 3.2. *Let $A: \mathbb{R}^n \to \mathbb{R}^m$, be a linear map and V any subset of $\mathbb{R}^{n+m} \times \mathbb{R}^{n+m} \times \mathbb{R}$. Consider the extremization problems*

$$(\mathcal{P}) \qquad \text{ext } V_A, \tag{3.10}$$

$$(\mathcal{P}^*) \qquad \text{ext } (\mathcal{L}V)_{-A^*} \tag{3.11}$$

The formulae

$$(x, Ax; -A^*q, q; z) \in V, \qquad z' = -z, \tag{3.10}$$

$$(-A^*q, q; x, Ax; z') \in \mathcal{L}V, \qquad z = -z' \tag{3.11}$$

are equivalent. Whenever (x, z) is a solution of $(\mathcal{P})$, the set of (q, z') satisfying (3.11) *or* (3.10) *is nonempty, and all of them are solutions of $(\mathcal{P}^*)$. Whenever (q, z') is a solution of $(\mathcal{P}^*)$, the set of (x, z) satisfying* (3.11) *or* (3.10) *is nonempty, and all of them are solutions of $(\mathcal{P})$.*

Proof. To say that (x, z) is a solution of $(\mathcal{P})$ means that there exists (p, q) such that

$$(x, Ax; p, q; z) \in V, \qquad p + A^*q = 0 \tag{3.12}$$

which we may write in a more symmetric form:

$$(x, y; p, q; z) \in V, \quad y - Ax = 0, \quad p + A^*q = 0. \tag{3.13}$$

Applying the Legendre transformation, we obtain

$$(p, q; x, y; px + qy - z) \in \mathcal{L}V, \quad y - Ax = 0, \quad p + A^*q = 0. \tag{3.14}$$

The last two equations imply that

$$z' = px + qy - z = -A^*q \cdot x + q \cdot Ax - z = -z \tag{3.15}$$

and formula (3.14) becomes

$$(p, q; x, y; -z) \in \mathcal{L}V, \quad y - Ax = 0, \quad p + A^*q = 0. \tag{3.16}$$

Breaking the symmetry, we get

$$(-A^*q, q; x, y; -z) \in \mathcal{L}V, \qquad y - Ax = 0 \tag{3.17}$$

which means precisely that $(q, -z)$ is a solution of $(\mathcal{P}^*)$. Since the Legendre transformation is an involution, formulae (3.12) and (3.17) are equivalent, and set up a one-to-one pairing between solutions (x, z) of $(\mathcal{P})$ and $(q, -z)$ of $(\mathcal{P}^*)$. But (3.12) is just (3.10), and (3.17) is (3.11). □

[3] From now on we shall omit the star.

The following is an easy consequence of the fact that the Legendre transformation $\mathscr{L}$ and the operation $A \to -A^*$ are involutions.

COROLLARY 3.3. $(\mathscr{P}^{**}) = (\mathscr{P})$.

Problems $(\mathscr{P})$ and $(\mathscr{P}^*)$ will be said to be *dual* to each other. Another easy consequence of Theorem 3.2 is the following.

COROLLARY 3.4. $\{\text{ext } \mathscr{P}\} = -\{\text{ext } \mathscr{P}^*\}$.

Theorem 3.2 is more readily understandable in the case of unconstrained smooth extremization problems. It reads as follows.

PROPOSITION 3.5. *Let $A: \mathbb{R}^n \to \mathbb{R}^m$ be a linear map and $f: \mathbb{R}^{n+m} \to \mathbb{R}$ a C^∞ function. Consider the extremization problems*:

$$(\mathscr{P}) \qquad \operatorname{ext}_x f(x, Ax),$$

$$(\mathscr{P}^*) \qquad \operatorname{ext}_q \mathscr{L}f(-A^*q, q).$$

The formulae

$$-A^*q = f'_x(x, Ax), \quad q = f'_y(x, Ax), \quad z' = -f(x, Ax) \tag{3.18}$$

set up a one-to-one pairing between solutions $(x, f(x, Ax))$ of $(\mathscr{P})$ and (q, z') of $(\mathscr{P}^*)$. Whenever the matrix of second derivatives f'' has rank $(n+m)$ at (x, Ax), there is a neighborhood $\mathscr{U}$ of $(-A^*q, q)$ and a C^∞ selection $\mathscr{L}_{\mathscr{U}}f$ of $\mathscr{L}f$ over $\mathscr{U}$ such that

$$f(x, Ax) = -(\mathscr{L}_{\mathscr{U}}f)(-A^*q, q), \tag{3.19}$$

$$x = (\mathscr{L}_{\mathscr{U}}f)'_p(-A^*q, q), \qquad Ax = (\mathscr{L}_{\mathscr{U}}f)'_q(-A^*q, q). \tag{3.20}$$

This follows easily from taking $V = V_f$, the Lagrangian submanifold associated with f, in Theorem 3.2. The last part is a consequence of Proposition 2.6. Note that relations analogous to (3.20) hold whenever $(\mathscr{L}f)'$ can be defined in a consistent way at $(p, q; z')$; this would be the case for the cusp points described in Proposition 2.7.

Let us give an important special case.

COROLLARY 3.6. *Let $\varphi: \mathbb{R}^n \to \mathbb{R}$ and $\psi: \mathbb{R}^m \to \mathbb{R}$ be C^∞ functions, and consider the extremization problems*

$$(\mathscr{P}) \qquad \operatorname{ext}_x \varphi(x) \dot{+} \psi(Ax),$$

$$(\mathscr{P}^*) \qquad \operatorname{ext}_q \mathscr{L}\varphi(-A^*q) + \mathscr{L}\psi(q).$$

Then $\{\text{ext } \mathscr{P}\} = -\{\text{ext } \mathscr{P}^*\}$, *and there is a one-to-one pairing between solutions $(x, \varphi(x) + \psi(Ax))$ of $(\mathscr{P})$ and (q^*, z') of $(\mathscr{P}^*)$, described by the relation*

$$-A^*q = \varphi'(x), \quad q = \psi'(Ax), \quad -z' = \varphi(x) + \psi(Ax). \tag{3.21}$$

Whenever φ'' has rank n at x and ψ'' has rank m at Ax, there are neighborhoods $\mathscr{U}_1$ and $\mathscr{U}_2$ of $-A^*q$ and q, selections $\mathscr{L}_{\mathscr{U}}\varphi$ and $\mathscr{L}_{\mathscr{U}}\psi$ of $\mathscr{L}\varphi$ and $\mathscr{L}\psi$ over $\mathscr{U}_1$ and $\mathscr{U}_2$,

such that

$$\mathscr{L}_{\mathscr{U}}\varphi(-A^*q)+\mathscr{L}_{\mathscr{U}}\psi(q)=\varphi(x)+\psi(Ax), \tag{3.22}$$

$$x=(\mathscr{L}_{\mathscr{U}}\varphi)'(-A^*q), \qquad Ax=(\mathscr{L}_{\mathscr{U}}\psi)'(q). \tag{3.23}$$

We now give two examples of applications of Theorem 3.2. They are both related to the problem of finding the eigenvectors and eigenvalues of a self-adjoint operator: we write it as an extremization problem in two different ways, and dualize both of them.

Let us start with the constrained smooth extremization problem

$$(\mathscr{P}) \qquad \begin{aligned} &\operatorname{ext}\|Ax\|^2, \\ &\|x\|^2=1. \end{aligned}$$

A solution to $(\mathscr{P})$ is a couple (x, z) such that

$$\|x\|^2=1, \qquad \exists\,\lambda\in\mathbb{R}\colon A^*Ax-\lambda x=0, \tag{3.24}$$

$$z=\|Ax\|^2=\lambda, \tag{3.25}$$

i.e. x is an eigenvector of A^*A and z is the corresponding eigenvalue.

Consider the subset $V\subset\mathbb{R}^{n+m}\times\mathbb{R}^{n+m}\times\mathbb{R}$ defined by

$$V=\{(x, y; -2\lambda x, 2y; \|y\|^2)\,|\,\|x\|^2=1, \lambda\in\mathbb{R}\}. \tag{3.26}$$

By Lemma 3.1 it is a Lagrangian submanifold. It is clear that problem $(\mathscr{P})$ is simply ext V_A. For the sake of convenience we will cut out part of V; indeed, it is apparent from formula (3.25) that $\lambda\geqq 0$ for any solution (x, z) of $\mathscr{P}$. So we introduce the "Lagrangian submanifold with boundary"

$$V'=\{(x, y; -2\lambda x, 2y; \|y\|^2)\,|\,\|x\|^2=1, \lambda\geqq 0\} \tag{3.27}$$

and we state problem $(\mathscr{P})$ as

$$(\mathscr{P}) \qquad \operatorname{ext} V'_A.$$

The Legendre transform of V' is again a Lagrangian submanifold with boundary. Going through the computations, we write it as a disjoint union $\mathscr{L}V=\Omega\cup\Gamma$, where Γ is the boundary

$$\Omega=\{(p, q; -p/\|p\|, q/2; -\|p\|+\|q\|^2/4)\,|\,p\neq 0\}, \tag{3.28}$$

$$\Gamma=\{(0, q; \xi, q/2; \|q\|^2/4)\,|\,\|\xi\|^2=1\}. \tag{3.29}$$

$\mathscr{L}V$ is clearly associated with the function $(p, q)\to -\|p\|+\|q\|^2/4$. The function $p\to -\|p\|$ is not differentiable at the origin, but let us agree that

$$\frac{d}{dp}(-\|p\|)\Big|_{p=0}=\{\xi\in\mathbb{R}^n\,|\,\|\xi\|^2=1\}. \tag{3.30}$$

This being agreed upon, we can now state the dual problem $(\mathscr{P}^*)$ in the following way:

$$(\mathscr{P}^*) \qquad \operatorname*{ext}_q -\|A^*q\|+\|q\|^2/4.$$

Theorem 3.2 implies that whenever $(q, -\|A^*q\| + \|q\|^2/4)$ is a solution to $(\mathscr{P}^*)$, all couples $(x, \|Ax\|^2)$ given by

$$x = A^*q/\|A^*q\| \quad \text{if } A^*q \neq 0, \quad Ax = q/2, \quad \|x\|^2 = 1, \tag{3.31}$$

$$\|Ax\|^2 = \|A^*q\| - \|q\|^2/4 \tag{3.32}$$

are solutions to $(\mathscr{P})$; in other words x is an eigenvector of A^*A with norm one, and $\|A^*q\| - \|q\|^2/4$ is the corresponding eigenvalue. For instance, formula (3.30) shows us that (0, 0) is a solution to $(\mathscr{P}^*)$ provided there exist $\xi \in \mathbb{R}^n$ with $\|\xi\|^2 = 1$ and $A\xi = 0$. Formulae (3.31) and (3.32) then yield the trivial fact that every such ξ is an eigenvector of A^*A with eigenvalue 0. Note as a conclusion that $-\{\text{ext } \mathscr{P}^*\}$ is just the spectrum of A^*A.

We now treat the same problem in another way. We define a subset W of $\mathbb{R}^{n+m} \times \mathbb{R}^{n+m} \times \mathbb{R}$ by

$$W = \{(x, y; -2x\|y\|^2/\|x\|^4, 2y/\|x\|^2; \|y\|^2/\|x\|^2) | x \neq 0\} \cup \{(0, 0; 0, \eta; 0) | \eta \in \mathbb{R}^m\}. \tag{3.33}$$

It can be checked that W is a Lagrangian submanifold. We asociate with it the extremization problem

$$(\mathscr{P}) \qquad \text{ext } W_A$$

which we state somewhat loosely as

$$(\mathscr{P}) \qquad \text{ext } \|Ax\|^2/\|x\|^2.$$

Of course, solving $(\mathscr{P})$ is just looking for the eigenspaces of A^*A. We now construct the dual problem $(\mathscr{P}^*)$. A simple computation yields

$$\mathscr{L}W = \{(p, q; -2p/\|q\|^2, 2q\|p\|^2/\|q\|^4; -\|p\|^2/\|q\|^2) | q \neq 0\} \cup \{0, 0; \pi, 0; 0) | \pi \in \mathbb{R}^n\}. \tag{3.34}$$

The dual problem $(\mathscr{P}^*)$, which is ext W_{-A^*}, will be stated somewhat loosely as

$$\text{ext } -\|A^*q\|^2/\|q\|^2. \tag{3.35}$$

We leave it to the reader to see what becomes of formulae (3.10)–(3.11). They tell us essentially that the eigenvalues of A^*A and AA^* coincide—a trivial fact.

We conclude this section by pointing out a technicality: even if V is a Lagrangian submanifold of $\mathbb{R}^{n+m} \times \mathbb{R}^{n+m} \times \mathbb{R}$, the set V_A need not be a Lagrangian submanifold of $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}$. Indeed, it need neither be closed nor be a submanifold. As a simple example, take

$$V = \{(x, y; -y/x^2, 1/x; y/x) | \neq 0\} \tag{3.36}$$

a Lagrangian submanifold of $\mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}$. Setting $A: x \to mx$, we get

$$V_A = \{(x, 0, m) | x \neq 0\} \tag{3.37}$$

which is not closed in $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$.

However, we have the following.

LEMMA 3.7. *If V is a Lagrangian submanifold and if V_A is a closed submanifold, then V_A is Lagrangian.*

Proof. We check condition (1.3) for V_A:

$$\begin{aligned} i_V^* \omega &= dz - (p + A^* q)\, dx \\ &= dz - p\, dx - q d(Ax) \end{aligned} \tag{3.38}$$

which is zero since $(x, Ax; p, q; z) \in V$, and the restriction of ω to V vanishes. □

Note also that if V is the Lagrangian submanifold associated with a C^∞ function $F: \mathbb{R}^n = \mathbb{R}^m \to \mathbb{R}$, then V_A is the Lagrangian submanifold associated with the C^∞ function $x \mapsto f(x, Ax)$ from $\mathbb{R}^n$ to $\mathbb{R}$—a fact we have used repeatedly in this section.

4. Applications to the calculus of variations. From now on, $\bar{\Omega} \subset \mathbb{R}^n$ will be an n-dimensional C^∞-submanifold with boundary Γ. We set $\Omega = \bar{\Omega} - \Gamma$, an open subset of $\mathbb{R}^n$; we endow Ω with the Lebesgue measure $d\omega$ and Γ with the induced $(n-1)$-dimensional measure $d\gamma$.

We consider a continuous linear map $A: V \to E$, where $E = L^2(\Omega; \mathbb{R}^m)$ and V is some Hilbertian subspace of $H = L^2(\Omega; \mathbb{R}^k)$ (i.e. V is a linear subspace of H endowed with some Hilbertian structure such that the inclusion mapping $V \to H$ is continuous). We assume that there is some Hilbert space T and some continuous liner map τ: $V \to T$ such that τ is surjective and $V_0 = \tau^{-1}(0)$ is dense in H. In practical examples, A will be some differential operator, V_0 will be $\overline{\mathscr{D}(\Omega)}$, the closure in V of the set of C^∞ functions with compact support in Ω, and T will associate with every function in V its "trace" on the boundary Γ. We shall state an abstract Green's formula for later use.

THEOREM 4.1. *There exist a Hilbertian subspace V^* of E, and continuous linear maps A^*: $V^* \to H$ and τ^*: $V^* \to T'$, the topological dual of T, such that, for every $x \in V$ and $q \in V^*$, we have*

$$(q, Ax) - (A^* q, x) = \langle \tau^* q, \tau x \rangle \tag{4.1}$$

where $(\cdot, \cdot)$ denotes scalar product in L^2 and $\langle \cdot, \cdot \rangle$ denotes the duality pairing between T' and T.

We now turn to extremization problems in the calculus of variations. From now on, we are given a family W_ω, $\omega \in \Omega$, of Lagrangian submanifolds of $\mathbb{R}^{k+m} \times \mathbb{R}^{k+m} \times \mathbb{R}$, and we denote by $F_\omega(x, y)$ the associated characteristic maps. Moreover, we are given a convex lower semi-continuous function Φ: $T \to \mathbb{R} \cup \{+\infty\}$; as usual in convex analysis, its subdifferential will be denoted by $\partial\Phi$. We now state.

DEFINITION 4.2. The *calculus of variations problem*[4]

$$(\mathscr{P}) \qquad \operatorname*{ext}_{x \in V} \int_\Omega F_\omega(x(\omega), Ax(\omega))\, d\omega + \Phi(\tau x)$$

consists in looking for all mappings $\omega \to (x(\omega), q(\omega), z(\omega))$ from Ω to

[4] Henceforth denoted by C.V. problem.

$\mathbb{R}^{k+m} \times \mathbb{R}^{k+m} = \mathbb{R}$ such that

$$x \in V, \quad q \in V^*, \quad z \in L^1, \tag{4.2}$$

$$(x(\omega), Ax(\omega); -A^*q(\omega), q(\omega); z(\omega)) \in W_\omega \quad \text{for a.e. } \omega \in \Omega, \tag{4.3}$$

$$\tau^* q \in -\partial\Phi(\tau x). \tag{4.4}$$

Any pair $(x, z) \in V \times L^1$ such that there exists $q \in V^*$ satisfying (4.2)–(4.4) will be called an *extremal* of $(\mathcal{P})$. The number ζ defined by

$$\zeta = \int_\Omega z(\omega)\, d\omega + \Phi(\tau x) \tag{4.5}$$

will be the associated *value* of $(\mathcal{P})$. The set of values of problem $(\mathcal{P})$ will be denoted by $\{\text{ext } \mathcal{P}\}$.

The motivation for this definition is clear. In the case where $F_\omega(\xi, \eta) = f(\omega; \xi, \eta)$, a function which is C^∞ in (ξ, η) for almost every $\omega \in \Omega$, and measurable in ω for every $(\xi, \eta) \in \mathbb{R}^k \times \mathbb{R}^m$, then (4.2)–(4.6) become

$$f'_\xi(\omega; x(\omega), Ax(\omega)) + A^* f'_\eta(\omega;, x(\omega), Ax(\omega)) = 0 \quad \text{a.e.}, \tag{4.6}$$

$$\tau^*[f'_\eta(x, Ax)] \in -\partial\Phi(x). \tag{4.7}$$

Equation (4.6) is the Euler–Lagrange equation on Ω associated with the integral

$$\int_\Omega f(\omega; x(\omega), Ax(\omega))\, d\omega \tag{4.8}$$

and formula (4.7) yields the so-called transversality conditions on the boundary Γ. In the case where f is convex in (ξ, η) for every ω, those are necessary and sufficient conditions for optimality. If f is not convex, but satisfies some growth condition at infinity, we get the first-order conditions for stationarity.

We now state the duality theorem.

THEOREM 4.3. *Consider the C.V. problems*

$$(\mathcal{P}) \qquad \operatorname*{ext}_{x \in V} \int_\Omega F_\omega(x(\omega), Ax(\omega))\, d\omega + \Phi(\tau x),$$

$$(\mathcal{P}^*) \qquad \operatorname*{ext}_{q \in V^*} \int_\Omega \mathcal{L}F_\omega(-A^*q(\omega), q(\omega))\, d\omega - \Phi^*(-\tau^* q).^5$$

Let (x, z) be an extremal of $(\mathcal{P})$ with value ζ; then, for any q satisfying (4.2)–(4.4), $(q, -xA^*q + qAx - z)$ is an extremal of $(\mathcal{P}^*)$ with value $-\zeta$. Conversely, let $(q, z') \in V^* \times L^1$ be an extremal of $(\mathcal{P}^*)$ with value ζ'; then, for any $x \in V$ satisfying

$$(-A^*q(\omega), q(\omega); x(\omega), Ax(\omega); z'(\omega)) \in \mathcal{L}W_\omega \quad \text{for a.e. } \omega \in \Omega, \tag{4.9}$$

$$\tau x \in \partial\Phi^*(-\tau^* q), \tag{4.10}$$

---

[5] Φ^* is the Fenchel conjugate of Φ in the sense of convex analysis:

$$\Phi^*(\delta') = \sup\{\langle \delta, \delta' \rangle - \Phi(\delta) | \delta \in T\} \quad \forall \delta' \in T'.$$

$(x, qAx - xA^*q - z')$ is an extremal of $(\mathscr{P})$ with value $-\zeta'$. Hence

$$\{\text{ext } \mathscr{P}\} = -\{\text{ext } \mathscr{P}^*\}. \tag{4.11}$$

Proof. The pointwise equation

$$(x(\omega), Ax(\omega); -A^*q(\omega), q(\omega); z(\omega)) \in W_\omega \tag{4.12}$$

can be written

$$(-A^*q(\omega), q(\omega); x(\omega), Ax(\omega); -x(\omega)A^*q(\omega) + Ax(\omega)q(\omega) - z(\omega)) \in \mathscr{L}W_\omega. \tag{4.13}$$

Moreover, formula (4.4) an also be written

$$\tau x \in \partial\Phi^*(-\tau^* q). \tag{4.14}$$

But equations (4.13) and (4.14), together with $x \in V$, $q \in V^*$, $z \in L^1$, simply mean that $(q, -xA^*q + Axq - z)$ is an extremal of $(\mathscr{P}^*)$. The associated value is

$$\zeta' = \int_\Omega (-x(\omega)A^*q(\omega) + Ax(\omega)q(\omega) - z(\omega))\, d\omega - \Phi^*(-\tau^* q). \tag{4.15}$$

Using Green's formula we have

$$\zeta' = -\int_\Omega z(\omega)\, d\omega + \langle \tau^* q, \tau x\rangle - \Phi^*(-\tau^* q). \tag{4.16}$$

Making use of (4.14), this becomes

$$\zeta' = -\int_\Omega z(\omega)\, d\omega - \Phi(\tau x) = -\zeta. \tag{4.17}$$

Hence the first part of the theorem is proved. The converse is proved along the same lines. □

Typical instances of such a mapping $A : V \to E$ are

$$\overrightarrow{\text{grad}}: H^1(\Omega) \to L^2(\Omega; \mathbb{R}^n), \tag{4.18}$$

$$\Delta: H^2(\Omega) \to L^2(\Omega; \mathbb{R}). \tag{4.19}$$

In the first case, T is $H^{1/2}(\Gamma)$, and Green's formula reads

$$\int_\Omega (\overrightarrow{\text{grad}}\, x \cdot \vec{q} + x \cdot \text{div}\, \vec{q})\, d\omega = \int_\Gamma \vec{n} \cdot \vec{q} x\, d\gamma. \tag{4.20}$$

In the second case, T is $H^{3/2}(\Gamma)$, and Green's formula reads

$$\int_\Omega (\Delta x \cdot q + x \cdot \Delta q)\, d\omega = \int_\Omega (q\vec{n} \cdot \overrightarrow{\text{grad}}\, x + x\vec{n} \cdot \overrightarrow{\text{grad}}\, q)\, d\gamma. \tag{4.21}$$

In both cases, we could define Φ as

$$\Phi(\delta) = \begin{cases} 0 & \text{if } \delta = \delta_0, \\ +\infty & \text{otherwise} \end{cases} \tag{4.22}$$

which gives a Dirichlet condition (fixed boundary values). We could also define

$$\Phi(\delta)=\begin{cases}0 & \text{if } \int_\Gamma \delta = 0,\\ +\infty & \text{otherwise}\end{cases} \tag{4.23}$$

which is a kind of periodicity condition.

Let us give an example:

$$(\mathscr{P}) \qquad \begin{aligned} &\operatorname{ext} \int_\Omega f(\omega; x(\omega), \operatorname{grad} x(\omega))\, d\omega,\\ &x \in H^1(\Omega), \qquad \int_\Gamma x(\gamma)\, d\gamma = 0 \end{aligned}$$

has the following dual:

$$(\mathscr{P}^*) \qquad \begin{aligned} &\operatorname{ext} \int_\Omega \mathscr{L}f(\omega; -\operatorname{div} q(\omega), q(\omega))\, d\omega,\\ &q \in H(\Omega; \operatorname{div}), \qquad q = \text{const. on } \Gamma \end{aligned}$$

where $H(\Omega, \operatorname{div}) = \{u \in L^2(\Omega, \mathbb{R}^n) | \operatorname{div} u \in L^2(\Omega, \mathbb{R}^n)\}$. The task of rewriting (4.2)–(4.4) and (4.9)–(4.11) is left to the reader.

We are now going to show that we can get simultaneously the extremals (x, z) of $(\mathscr{P})$ and the extremals (q, z') of $(\mathscr{P}^*)$ from the extremals of a single C.V. problem.

PROPOSITION 4.4. *Consider the C.V. problems*

$$(\mathscr{Q}) \qquad \operatorname*{ext}_{\substack{(x,y,q)\in\\ V\times E\times V^*}} \int_\Omega [-A^*q(\omega)\cdot x(\omega) + q(\omega)y(\omega) - F_\omega(x(\omega), y(\omega))]d\omega - \Phi^*(-\tau^*q),$$

$$(\mathscr{Q}^*) \qquad \operatorname*{ext}_{\substack{(x,p,q)\in\\ V\times E\times V^*}} \int_\Omega [p(\omega)x(\omega) + q(\omega)\cdot Ax(\omega) - \mathscr{L}F_\omega(p(\omega), q(\omega))]\, d\omega + \Phi(\tau x).$$

The following are equivalent statements:

(a) (x, y, q, z') *is an extremal of* $(\mathscr{Q})$,

(b) (x, p, q, z) *is an extremal of* $(\mathscr{Q}^*)$,

(c) (x, q, z) *satisfy* (4.2)–(4.4),

(d) (q, x, z') *satisfy* (4.9)–(4.10) *and* $z' \in L^1$,

with $z + z' = -A^*q \cdot x + q \cdot Ax$. *In particular* (x, z) *is an extremal of* $(\mathscr{P})$ *and* (q, z') *an extremal of* $(\mathscr{P}^*)$.

Proof. We have already shown that (c) and (d) are equivalent. We shall be content with proving that (a) and (c) are equivalent; the proof that (b) and (d) are equivalent goes along the same lines.

Problem $(\mathscr{Q})$ can be written as

$$(\mathscr{Q}) \qquad \operatorname*{ext}_{\substack{(x,y,q)\in\\ V\times E\times V^*}} \int_\Omega \mathscr{F}_\omega(x(\omega), y(\omega), -A^*q(\omega), q(\omega))\, d\omega$$

where $\mathcal{F}_\omega$ is the characteristic map associated wth the Lagrangian submanifold $\mathcal{W}_\omega$ of $\mathbb{R}^{2k+2m} \times \mathbb{R}^{2k+2m} \times \mathbb{R}$ defined by

$$\mathcal{W}_\omega = \{(\xi, \eta, \pi, \rho; \pi - \sigma, \rho - \tau, \xi, \eta; \pi\xi + \rho\eta - \zeta) | \pi \in \mathbb{R}^k, \quad \rho \in \mathbb{R}^m, (\xi, \eta; \sigma, \tau; \zeta) \in W_\omega\}. \tag{4.24}$$

We now apply Definition 4.2 to the Hilbert space $\mathcal{V} = V \times E \times V^*$ and the map $\mathcal{A}$: $\mathcal{V} \to E$ defined by $\mathcal{A}(x, y, q) = -A^*q$; its adjoint will be the map $\mathcal{A}^*$: $V \to H \times E \times H$ defined by $\mathcal{A}^*(x') = (0, 0, -Ax')$. Conditions (4.2)–(4.4) then become

$$x \in V, \quad y \in E, \quad q \in V^*, \quad x' \in V, \quad z' \in L^1, \tag{4.25}$$

$$(x(\omega), y(\omega), -A^*q(\omega), q(\omega); 0, 0, x'(\omega), Ax'(\omega); z'(\omega)) \in \mathcal{W}_\omega \quad \text{a.e.}, \tag{4.26}$$

$$\tau x' \in \partial\Phi^*(-\tau q). \tag{4.27}$$

So $(x, y, q, z') \in V \times E \times V^* \times L^1$ is an extremal of $(\mathcal{Q})$ if and only if there exists $x' \in V$ such that (4.26) and (4.27) are satisfied. Now, comparing (4.26) with (4.24), we get

$$-A^*q(\omega) = \sigma, \tag{4.28}$$

$$q(\omega) = \tau, \tag{4.29}$$

$$x'(\omega) = x(\omega), \tag{4.30}$$

$$Ax'(\omega) = y(\omega), \tag{4.31}$$

$$z'(\omega) = -A^*q(\omega) \cdot x(\omega) + q(\omega)y(\omega) - \zeta, \tag{4.32}$$

$$x(\omega), y(\omega); \sigma, \tau; \zeta) \in W_\omega. \tag{4.33}$$

All this boils down to

$$(x(\omega), Ax(\omega); -A^*q(\omega), q(\omega); z(\omega)) \in W_\omega \quad \text{a.e.} \tag{4.34}$$

with $z(\omega) + z'(\omega) = -A^*q(\omega) \cdot x(\omega) + q(\omega) \cdot Ax(\omega)$. With (4.30) taken into account, (4.27) becomes

$$\tau x \in \partial\Phi^*(-\tau^* q) \tag{4.35}$$

which can be inverted to

$$-\tau^* q \in \partial\Phi(\tau x). \tag{4.36}$$

But (4.34) and (4.36) are just (c), and we have proved our claim. □

Proposition 4.4 can be considered a smooth version of the saddle-point property for Lagrange multipliers in convex optimization. Note that in the case where $F_\omega(\xi, \eta) = f(\omega; \xi, \eta)$, measurable in ω, C^∞ in (ξ, η), problem $(\mathcal{P}^*)$ involves $\mathcal{L}f(\omega; \xi, \eta)$ which typically is multivalued and cusped; working with problem $(\mathcal{Q})$ is a way of circumventing this inconvenience at the cost of increasing the dimension.

We now apply this idea of "smoothing out" Legendre transforms to another example.

PROPOSITION 4.5. *We are given a C^∞ function $\varphi: [0, T]\times\mathbb{R}^n \to \mathbb{R}$, a measurable function $f: [0, T]\to\mathbb{R}^n$, and a point $\xi_0\in\mathbb{R}^n$. We consider the differential equation*

$$(\mathscr{E}) \qquad \frac{dx}{dt}+\varphi'_\xi(t, x)=f \quad \text{a.e. on } [0, T],\ x(0)=\xi_0,$$

and the C.V. problems

$$(\mathscr{P}) \qquad \begin{aligned} &\operatorname{ext}\int_0^T\left[\varphi(t, x)+\mathscr{L}\varphi\left(t; f-\frac{dx}{dt}\right)+x\left(\frac{dx}{dt}-f\right)\right]dt,\\ &x\in H^1(0, T;\mathbb{R}^n), \qquad x(0)=x_0; \end{aligned}$$

$$(\mathscr{Q}) \qquad \begin{aligned} &\operatorname{ext}\int_0^T\left[\varphi(t, x)-\varphi(t, y)+\left(\frac{dx}{dt}-f\right)(x-y)\right]dt,\\ &x\in H^1(0, T;\mathbb{R}^n), \quad y\in H^1(0, T;\mathbb{R}^n), \quad x(0)=y(0)=\xi_0. \end{aligned}$$

If $(\mathscr{E})$ has no solution, then problems $(\mathscr{P})$ and $(\mathscr{Q})$ have no extremals. If $(\mathscr{E})$ has a solution $\bar{x}$, then problem $(\mathscr{P})$ has a unique extremal $(\bar{x}, 0)$, and problem $(\mathscr{Q})$ has a unique extremal $(\bar{x}, \bar{x}, 0)$.

Proof. Problem $(\mathscr{Q})$ arises from problem $(\mathscr{P})$ by replacing $\mathscr{L}\varphi(f-(dx/dt))$ by $y(f-(dx/dt))-\varphi(y)$, i.e. by smoothing out that part of the integrand which is a Legendre transform. Proposition 4.4 does not readily apply to this case, so we give a direct proof.

An extremal (x, y, z) of $(\mathscr{Q})$ is defined by the Euler equations

$$\varphi'_\xi(t, x)+\frac{dx}{dt}-f=\frac{d}{dt}[x-y], \tag{4.37}$$

$$-\varphi'_\xi(t, y)-\frac{dx}{dt}+f=0 \tag{4.38}$$

and the boundary conditions $x(0)=y(0)=x_0$. Together, they yield the system of differential equations on $[0, T]$

$$\frac{dy}{dt}+\varphi'_\xi(t, x)=f, \qquad y(0)=x_0, \tag{4.39}$$

$$\frac{dx}{dt}+\varphi'_\xi(t, y)=f, \qquad x(0)=x_0. \tag{4.40}$$

Now this is to be compared with equation

$$(\mathscr{E}) \qquad \frac{dx}{dt}+\varphi'_\xi(t, x)=f, \qquad x(0)=x_0.$$

The assumptions on φ imply that both system (4.39)–(4.40) and equation $(\mathscr{E})$ have at most one solution. If $\bar{x}$ is the solution of $(\mathscr{E})$, obviously $(\bar{x}, \bar{x})$ is the solution of (4.39)–(4.40). Conversely, if $(\bar{x}, \bar{y})$ is a solution of (4.39)–(4.40), then so is $(\bar{y}, \bar{x})$; from the uniqueness, it follows that $\bar{x}=\bar{y}$, obviously the solution of $(\mathscr{E})$. Writing $x=\bar{x}=\bar{y}=y$ in the integrand, we see that it is identically zero. We have proved the equivalence of equation $(\mathscr{E})$ and problem $(\mathscr{Q})$.

The equivalence of problems $(\mathscr{P})$ and $(\mathscr{Q})$ goes along the lines set up in Proposition 4.4. Indeed, (4.38) means simply that

$$-\varphi(t, y)-\left(\frac{dx}{dt}-f\right)y=\mathscr{L}\varphi\left(t; f-\frac{dx}{dt}\right) \tag{4.41}$$

and the integrands in $(\mathscr{P})$ and $(\mathscr{Q})$ become equal. With $x = y$, formula (4.41) yields, with a slight misuse of notations,

$$[\mathscr{L}\varphi]'_\xi\left(t, f-\frac{dx}{dt}\right)=x \tag{4.42}$$

and the Euler equation for $(\mathscr{P})$ turns out to be exactly equation $(\mathscr{E})$. □

Note that we have defined directly the extremals of a problem in the calculus of variations, without reference to any extremization problem. This is because the natural extremization problem involved is infinite-dimensional, and the results of the preceding sections do not extend readily to this case; indeed, smoothness assumptions which are natural in finite dimensions become preposterous in this new setting. In some particular cases, however, it can be made to work. Let us give an example, which will be recognized as an infinite-dimensional version of the example concluding § 3.

We consider the space $V = H_0^1(\Omega)$ and the function

$$f\colon V\setminus\{0\}\times L^2(\Omega)^n\to\mathbb{R}, \tag{4.43}$$

$$f(x, y)=|y|^2/|x|^2 \tag{4.44}$$

with $|\cdot|$ denoting the L^2-norm. Obviously f is a C^∞ function, with

$$p=f'_x(x, y)=-2x|y|^2/|x|^4\in L^2(\Omega), \tag{4.45}$$

$$q=f'_y(x, y)=2y/|x|^2\in L^2(\Omega)^n. \tag{4.46}$$

We now set

$$y=\operatorname{grad} x \tag{4.47}$$

to get the extremization problem

$$(\mathscr{P}) \qquad \begin{aligned} &\operatorname{ext}|\operatorname{grad} x|^2/|x|^2,\\ &x\in H_0^1(\Omega), \qquad x\neq 0. \end{aligned}$$

Let us write out the equation for a critical point, taking into account the fact that the transpose of grad: $H_0^1(\Omega)\to L^2(\Omega)^n$ is $-\operatorname{div}$: $L^2(\Omega)^n\to H^{-1}(\Omega)$:

$$0=p-\operatorname{div} q=-2(x|\operatorname{grad} x|^2/|x|^2+\operatorname{div}\operatorname{grad} x)/|x|^2. \tag{4.48}$$

Note that $|\operatorname{grad} x|$ cannot be zero unless x is, so (4.48) becomes

$$x=-\frac{|x|^2}{|\operatorname{grad} x|^2}\Delta x, \qquad x\neq 0. \tag{4.49}$$

In other words, the solutions of $(\mathscr{P})$ are the pairs $(x, 1/\lambda)$ where $-\lambda$ is a nonzero eigenvalue of the Laplacian under homogeneous boundary conditions, and x any nonzero eigenvector.

To get the dual problem, we note that (4.45) and (4.46) are invertible whenever $y \neq 0$; this yields

$$x = -2p/|q|^2, \qquad y = 2q|p|^2/|q|^4, \tag{4.50}$$

so we are in the particularly simple case where the Legendre transformation is one-to-one. Equations (4.48) and (4.47) become

$$p = \operatorname{div} q \in L^2, \tag{4.51}$$

$$2(q|p|^2/|q|^2 + \operatorname{grad} \operatorname{div} q)/|q|^2 = 0. \tag{4.52}$$

But this means exactly that $q \neq 0$ is a critical point of the function $q \to -|\operatorname{div} q|^2/|q|^2$ over the space

$$H(\Omega; \operatorname{div}) = \{q \in L^2(\Omega)^n | \operatorname{div} q \in L^2(\Omega)\}. \tag{4.53}$$

Finally, we get the dual problem

$$(\mathscr{P}^*) \qquad \begin{aligned} &\operatorname{ext} -|\operatorname{div} q|^2/|q|^2, \\ &q \in H(\Omega; \operatorname{div}) \end{aligned}$$

with the usual relationship (4.45)–(4.46) or (4.50). Note in particular that

$$\{\operatorname{ext} \mathscr{P}\} = -\{\operatorname{ext} \mathscr{P}^*\}. \tag{4.54}$$

5. Comments. The notion of a Lagrangian submanifold is central to the theory of Fourier integral operators. It is attributed to V. Arnold [1] or V. Maslov [13], and has been painstakingly investigated [11, especially § 3], [16], [9]. However, these authors define a Lagrangian submanifold of a symplectic manifold (dimension $2n$, fundamental 2-form Ω) as an n-dimensional submanifold on which Ω pulls back to zero. In our framework, this would mean an n-dimensional submanifold of $\mathbb{R}^n \times \mathbb{R}^n$ on which $\hat{\Omega} = \sum_{i=1}^n dp_i \wedge dx_i$ pulls back to zero. Noting $\omega = dz - \sum_{i=1}^n p_i\, dx_i$ as in (1.3), we see that $\Omega = d\omega$. It follows that if $V \subset \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ is a Lagrangian submanifold in the sense of Definition 1.1, if the projection $\pi_{xp}: V \to \mathbb{R}^n \times \mathbb{R}^n$ is proper and if its tangent map $T\pi_{xp}: TV \to T(\mathbb{R}^n \times \mathbb{R}^n)$ has rank n everywhere, then $\pi_{xp}V$ is a Lagrangian submanifold of $\mathbb{R}^n \times \mathbb{R}^n$ in the preceding sense. Our definition has the advantage of incorporating z, which is very useful for practical purposes.

For basic information about proper maps, we refer to any book on general topology, e.g. [4, Chaps. 1 and 2]. Sard's theorem in the C^∞ case, as well as basic information on submanifolds and the implicit function theorem, can be found in [12].

The definition (2.1) of the Legendre transformation is given in [6] as a particular case of a contact transformation. The contact transformation associated with a given C^∞ function $H(x, z; x', z')$ on $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$ is the mapping which associates with any point $(x, p, z) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ the point (x', p', z') defined by the formulae

$$H(x', z'; x, z) = 0,$$

$$\partial H/\partial x' + p'\, \partial H/\partial z' = 0,$$

$$\partial H/\partial x + p\, \partial H/\partial z = 0.$$

From the two last equations it follows (formally) that $p = \partial z/\partial x$ and $p' = \partial z'/\partial x'$. It follows (still formally) from the first one that $dz' + p'\,dx' = 0$ if and only if $dz + p\,dx = 0$. In other words, if we have no trouble with cusps or closedness, a contact transformation will send a Lagrangian manifold onto a Lagrangian manifold. It need not be involutive. In the special case where $H(x', z'; x, z) = z + z' - xx'$, we get the Legendre transformation.

Also related to the Legendre transform is the notion of dual varieties in algebraic geometry. Let a projective variety C be given by its equation $P(X_1, \cdots, X_n) = 0$, where P is a homogeneous polynomial of degree d. The dual variety, $\hat{C}$ is the set of tangents to C; its equation $\hat{P}(u_1, \cdots, u_n) = 0$ has as zeros all $(u_1, \cdots, u_n)$ such that the hyperplane $u_1X_1 + \cdots + u_nX_n$ is tangent to C. In particular, $\hat{\hat{C}} = C$. For instance, if $f\colon \mathbb{R}^n \to \mathbb{R}$ is a polynomial, setting

$$z = \frac{X_{n+1}}{X_{n+2}}, \qquad x_i = \frac{X_i}{X_{n+2}}$$

as is usual in projective geometry yields

$$\operatorname{graph} f = \left\{(X_1, \cdots, X_{n+2}) \middle| \frac{X_{n+1}}{X_{n+2}} = f\left(\frac{X_1}{X_{n+2}}, \cdots, \frac{X_n}{X_{n+2}}\right)\right\}.$$

The dual variety is simply the graph of the Legendre transform

$$\widehat{\operatorname{graph} f} = \operatorname{graph} \mathscr{L}f.$$

A particularly interesting case arises when $n = 1$ and complex numbers are used. It can be shown that if C (resp. $\hat{C}$) is a complex algebraic curve of degree d (resp. $\hat{d}$), having r (resp. $\hat{r}$) double points and s (resp. $\hat{s}$) cusps, with no other singularities, then we have the following symmetric relationship (Plücker's formulae)

$$\hat{d} = d(d-1) - 2r - 3s,$$
$$d = \hat{d}(\hat{d}-1) - 2\hat{r} - 3\hat{s},$$
$$\hat{s} - s = 3(\hat{d} - d).$$

(I am indebted to P. Deligne for this elementary algebraic geometry.)

Now let us proceed to providing §§ 2, 3, 4 with bibliographical references.

Fundamentals of convex analysis are given in [14] or [8]. Modern tools of different topology, included the Malgrange division theorem, Thom's transversality theorem and notions on stratifications, will be found in [15]; see [10] for a textbook on the subject. Note that the proof of Proposition 2.7 for $n = 1$ does not require the C^∞ division theorem.

Condition (3.7) can be interpreted as a necessary condition for optimality in a much broader context than indicated, i.e. the space need not be finite-dimensional and the g_i' need not be linearly independent; see [7]. Duality theory for finite-dimensional convex optimization problems will be found in [14].

Theorem 4.1 is due to J.-P. Aubin. Its proof will be found in [2] or [3]. Duality theory for convex problems in the calculus of variations is treated in [8], but here we follow rather the approach of [3]. Proposition 4.5 is a nonconvex analogue of [5].

Acknowledgments. I am indebted to R. Temam for suggesting to me the eigenvalue examples concluding §§ 3 and 4. Also I wish to acknowledge long and Numerous conversations with J.-P. Aubin and F. Clarke, and the expert typing of Mrs. Sally Ross.

REFERENCES

[1] V. I. ARNOLD, *Characteristic class entering in quantization conditions*, J. Functional Anal. Appl., 1 (1967), pp. 1–13.
[2] J.-P. AUBIN, *Approximation of Elliptic Boundary-value Problems*, John Wiley, New York, 1972.
[3] ———, *Mathematical methods of game and economic theory*, North-Holland Elsevier, Amsterdam, to appear.
[4] N. BOURBAKI, *Topologie Générale*, 2éme edition, Hermann, Paris, 1960.
[5] H. BREZIS AND I. EKELAND, *Un principe variational associé à certaines équations paraboliques*, C.R. Acad. Sci. Paris, Sér. A-B, 282 (1976), pp. 971–974 and 1197–1198.
[6] C. CARATHEODORY, *Variations rechnung und partielle differentialgleichungen erster Ordnung*, Teubner, Leipzig, 1935.
[7] F. CLARKE, *A new approach to Lagrange multipliers*, Mathematics of Operations Res., 1 (1976).
[8] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland Elsevier, Amsterdam, 1975.
[9] J. GUCKENHEIMER, *Catastrophes and partial differential equations*, Ann. Inst. Fourier (Grenoble), 23 (1973), 2, pp. 31–59.
[10] M. GOLUBISTSKY AND V. GUILLEMIN, *Stable Mappings and Their Singularities*, Springer-Verlag, Berlin, 1973.
[11] L. HÖRMANDER, *Fourier integral operators I*, Acta Math., 127 (1971), pp. 79–183.
[12] S. LANG, *Differentiable Manifolds*, Addison-Wesley, Reading, MA, 1972.
[13] V. MASLOV, *Theory of Perturbations and Asymptotic Methods*, Moskov. Gos. Univ., Moscow, 1965. (In Russian.)
[14] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1969.
[15] WALL, ed., *Proceeding of the Liverpool Singularities Symposium I*, Springer Lecture Notes in Mathematics 192, Springer-Verlag, Berlin, 1971.
[16] A. WEINSTEIN, *Symplectic manifolds and their Lagrangian submanifolds*, Advances in Math., 6 (1971), pp. 329–346.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

A CLASS OF NONLINEAR INTEGER PROGRAMS SOLVABLE BY A SINGLE LINEAR PROGRAM*

R. R. MEYER†

Abstract. Although the addition of integrality constraints to the existing constraints of an optimization problem will, in general, make the determination of an optimal solution more difficult, we consider here a class of nonlinear programs in which the imposition of integrality constraints on the variables makes it possible to solve the problem by a single, easily-constructed linear program. The class of problems addressed has a separable convex objective function and a totally unimodular constraint matrix. Such problems arise in logistic and personnel assignment applications.

1. Introduction.

Nonlinear integer programs of the form

$$\min_x \sum_{i=1}^{n} f_i(x_i)$$

$$\text{subject to } \sum_{i=1}^{n} x_i = r, \tag{1.1}$$

$$x = (x_1, \cdots, x_n)^T \geqq 0, \quad x_i \text{ integer } (i = 1, \cdots, n)$$

arise in logistic and personnel assignment applications and have been the subject of a number of studies [3], [9], [13], [14], [16]. Here, we consider the broader class of problems of the form

$$\min_x \sum_{i=1}^{n} f_i(x_i)$$

$$\text{subject to } Ax = b, \tag{1.2}$$

$$x \geqq 0, \quad x \text{ integer},$$

where A is a totally unimodular (T.U.) $m \times n$ matrix and b is integer,[1] (in the following, a vector is said to be *integer* if all its components are integer), and we will show that a solution to the problem (1.2) may be obtained by solving a *single* easily-constructed linear program provided that known bounds exist for the feasible set of (1.2) and each f_i is a *convex* function. This result thus also generalizes the well-known property [8] that, in the case that all the f_i are linear (so that (1.2) is a linear integer program), the solution of (1.2) may be obtained by solving a single linear program.

* Received by the editors April 2, 1976, and in revised form November 23, 1976.

† Computer Sciences Department, University of Wisconsin—Madison, Madison, Wisconsin 53706. This research was sponsored by the National Science Foundation under contract DCR 74-20584.

[1] Recall that a matrix is said to be *totally unimodular* if the determinant of each of its square submatrices has value 0 or ±1. Totally unimodular matrices typically arise in optimization problems defined on networks, but may also arise in other contexts such as bounds on sums or differences of subsets of variables. Although we assume here the equality constraints $Ax = b$, analogous results hold if $Ax = b$ is replaced by $Ax \leqq b$ or by *any combination of equations and inequalities* whose aggregate coefficient matrix is totally unimodular, since the conversion of such constraints to a set of equations (by the addition of slack variables) yields a new coefficient matrix that will also be totally unimodular.

Finally, this result may also be thought of as a generalization of a method suggested by Dantzig [5] for a class of transportation problems. (Although single-commodity network problems with convex costs on their arcs can generally be put in the form of (1.2), the study of the converse raises some interesting issues that are discussed in § 4.)

For the case in which bounds for the feasible set are *not* known, a column-generation procedure is developed, and it is shown that, if an optimal solution to (1.2) exists, this procedure will yield an optimal solution (and a proof of its optimality) by the solution of a finite number of linear programs. This column-generation procedure also has computational advantages in the bounded case if the bounds are very large and/or one or more of the f_i are "costly" to evaluate. It differs from parametric procedures proposed by Beale [1] and Hu [7] for certain convex network optimization problems in that it employs "global" rather than "local" cost function approximations.

2. An equivalent linear program. In this section we will establish the equivalence of the nonlinear integer program (1.2) to a linear program ((2.5) below) under the following *hypotheses*:

(A) there exist nonnegative integers l_i, u_i $(i = 1, \cdots, n)$ such that $F \equiv \{x | Ax = b, x \geqq 0\} \subseteq \{x | l_i \leqq x_i \leqq u_i, i = 1, \cdots, n\}$;

(B) $F \neq \varnothing$;

(C) the matrix A is an $m \times n$ totally unimodular matrix and b is integer;

(D) (for $i = 1, \cdots, n$) f_i is a real-valued convex function on $[l_i, u_i]$.

Note that under hypotheses (A) and (B), the nonlinear integer program (1.2) has an optimal solution since the number of feasible points is finite and nonzero. (The hypotheses (A) and (B) can, in fact, be deleted, as is shown in § 5, but the proofs for the more general case are straightforward extensions of the results of this section.) Let $\tilde{f}_i$ denote the continuous piecewise-linear function defined on $[l_i, u_i]$ that *coincides* with f_i at the integer points in $[l_i, u_i]$ and is *linear* between each pair of adjacent integers in $[l_i, u_i]$. It is easily seen that each $\tilde{f}_i$ is also convex. (In fact, it is convexity of the $\tilde{f}_i$ that is crucial rather than convexity of the f_i.) The problem (1.2) is therefore equivalent[2] to

$$\min_x \sum_{i=1}^{n} \tilde{f}_i(x_i)$$

$$\text{subject to } Ax = b, \tag{2.1}$$

$$x \geqq 0, \quad x \text{ integer},$$

since the objective functions of (1.2) and (2.1) *coincide* over their common feasible set. (Put another way, the values of the objective function terms at noninteger points are completely irrelevant to the optimization problem (1.2), so we can take advantage of this fact by "simplifying" the form of the objective function terms between consecutive integers.) We will now exploit properties of a

[2] If integrality constraints were not present in (1.2), then the $\tilde{f}_i$ would merely be approximations to the f_i; but, given the integrality constraints, no "error" is incurred over the discrete domain by replacing f_i by $\tilde{f}_i$. Thus in this context the $\tilde{f}_i$ should *not* be thought of as "approximations," as is the case when similar substitutions are done in the continuous variable case (see [2], [4]).

particular representation of the $\tilde{f}_i$ in order to get rid of the integrality constraints. (The overall strategy is thus to exploit the integrality constraints to modify the objective function, and then to exploit the modified objective function and total unimodularity to get rid of the integrality constraints.)

It is a well-known result of separable programming (see [5]) that, for $x_i \in [l_i, u_i]$, we have the following representation for the $\tilde{f}_i$ (R_i' denotes the integers in $[l_i, u_i]$):

$$\begin{aligned} \tilde{f}_i(x_i) &= \min_{\lambda_{i,j}} \sum_{j \in R_i'} f_i(j)\lambda_{i,j} \\ &\text{subject to } \sum_{j \in R_i'} j\lambda_{i,j} = x_i, \\ &\qquad \sum_{j \in R_i'} \lambda_{i,j} = 1; \quad \lambda_{i,j} \geqq 0. \end{aligned} \tag{2.2}$$

(In [10], which is an expanded version of this paper, it is shown that the so-called "δ-representation" of $\tilde{f}_i$, namely

$$\begin{aligned} \tilde{f}_i(x_i) &= \min_{\delta_{i,j}} f_i(l_i) + \sum_{j \in R_i} \delta_{i,j}[f_i(j+1) - f_i(j)] \\ &\text{subject to } l_i + \sum_{j \in R_i} \delta_{i,j} = x_i, \\ &\qquad 0 \leqq \delta_{i,j} \leqq 1, \quad j \in R_i, \end{aligned}$$

where $R_i = R_i'/\{u_i\}$, may also be used to obtain analogous results. In this paper we will concentrate on the "λ-representation" (2.2), which turns out to be more appropriate for a column-generation method to be discussed below.) Thus, the problem (2.1) is equivalent to the problem

$$\begin{aligned} &\min_{\lambda, x} \sum_{i=1}^{n} c_i\lambda_i \\ &\text{subject to } Ax = b, \quad x \geqq 0, \quad x \text{ integer}, \\ &\qquad D\lambda = x, \quad E\lambda = e, \quad \lambda \geqq 0 \end{aligned} \tag{2.3}$$

where $\lambda_i = (\lambda_{i,l_i}, \cdots, \lambda_{i,u_i})^T$, $\lambda = (\lambda_1, \cdots, \lambda_n)$, $c_i = (f_i(l_i), \cdots, f_i(u_i))$, $e = (1, \cdots, 1)^T$, and the constraints $D\lambda = x$, $E\lambda = e$, $\lambda \geqq 0$ represent the constraints of (2.2) as i ranges from 1 to n. The problem (1.2) has thus been transformed into an equivalent *linear* mixed-integer program (2.3). Now if the constraint matrix of (2.3) were totally unimodular, then the integrality constraints of (2.3) could be deleted without affecting the optimal value. However, because D contains integer entries other than 0 or ± 1, the constraint matrix of (2.3) is *not* totally unimodular, and if we consider the linear programming relaxation of (2.3)

$$\begin{aligned} &\min_{\lambda, x} \sum_{i=1}^{n} c_i\lambda_i \\ &\text{subject to } Ax = b, \quad x \geqq 0, \\ &\qquad D\lambda = x, \quad E\lambda = e, \quad \lambda \geqq 0, \end{aligned} \tag{2.4}$$

examples are easily constructed to show that the feasible set of (2.4) may have *noninteger* extreme points. Moreover, if $(\hat{\lambda}, \hat{x})$ is an extreme point of (2.4) then $\hat{x}$ need *not* be an extreme point of F. (This reflects the fact that (1.2) may have a unique optimal solution lying in the *interior* of F.) However, we will show that if $(\hat{\lambda}, \hat{x})$ is an extreme point of (2.4), then the vector $\hat{x}$ must be *integer*, and thus this condition is sufficient to guarantee that the optimal value of (2.4) is equal to the optimal value of (1.2). For notational convenience, we denote the equality constraints of (2.4) as

$$Ax = b, \tag{2.5}$$

$$D\lambda = x, \tag{2.6}$$

$$E\lambda = e. \tag{2.7}$$

THEOREM 2.1. *If* $(\hat{\lambda}, \hat{x})$ *is an extreme point of* (2.4), *then* $\hat{x}$ *is integer.*

Proof. Let λ_B and x_B be the basic variables corresponding to the extreme point $(\hat{\lambda}, \hat{x})$. It is easily seen from (2.6) and (2.7) that at least one and at most two variables from each λ_i must be in λ_B. Let $x'_B \equiv \{x_i | x_i$ is basic and λ_B contains exactly one variable of $\lambda_i\}$ and $x''_B \equiv \{x_i | x_i$ is basic and λ_B contains exactly two variables of $\lambda_i\}$, with corresponding definitions for λ'_B and λ''_B. If x_i is in x'_B, let μ_i denote the basic variable in λ_i, so that (2.6) and (2.7) imply $\hat{\mu}_i = 1$ and $\hat{x}_i = d_i\hat{\mu}_i = d_i$ for some integer d_i in R'_i.

Thus, the variables x'_B are all integer-valued, and we will now show that this is the case for x''_B also. For each variable x_i in x''_B, we let μ_i be one of the corresponding basic variables in λ_i, so that the other basic variables in λ_i can be replaced by $1-\mu_i$ because of (2.7). Denote the coefficient of the variable in (2.6) corresponding to $(1-\mu_i)$ as d_i and the coefficient of the other basic variable μ_i as $d_i + h_i$ (note that h_i is a *nonzero* integer). Using the change of variable $x_i = d_i + x''_i$, we have from (2.6), $d_i + x''_i = d_i(1-\mu_i) + (d_i + h_i)\mu_i$ or $x''_i = h_i\mu_i$, so that each such μ_i is uniquely determined by x''_i. We will now show that the columns of A corresponding to x''_B are linearly independent. For, suppose that they were not, and set all variables other than x''_B and λ''_B to their values in the solution $(\hat{\lambda}, \hat{x})$. If the columns of A corresponding to x''_B were linearly dependent, there would be infinitely many sets of values of x''_B for which (2.5) (with the other variables set to their values in $\hat{x}$) would be satisfied, and for each such set of values, values of μ_i could be determined so that (2.6) and (2.7) were also satisfied. This contradicts the fact that the system (2.5)–(2.7) must have a *unique* solution when the nonbasics are set to 0. Thus having shown that the columns of A corresponding to x''_B are linearly independent, we conclude from the T.U. of A that $\hat{x}''_B$ is integer. □

THEOREM 2.2. *The optimal value of* (2.4) *is equal to the optimal value of* (1.2), *and if* (λ^*, x^*) *is an optimal extreme point of* (2.4), *then* x^* *solves* (1.2).

Proof. Since the feasible set of the linear program [LP] (2.4) is nonempty and bounded, then (2.4) has an optimal solution. Thus, there exists an extreme point of (2.4) that is optimal, and the x-coordinates of this extreme point must be optimal for (1.2). □

Thus, the original nonlinear integer program (1.2) can be solved by computing the values of each f_i at the integer points in $[l_i, u_i]$ and solving the linear program (2.4) by the simplex method, which will generate an optimal extreme point.

It should also be noted that Theorem 2.2 also implies that (1.2) and (2.4) have the *same* optimal value as the problem obtained from (2.1) by deleting its integrality constraints, namely

$$\begin{aligned} &\min \sum_{i=1}^{n} \tilde{f}_i(x_i) \\ &\text{subject to } Ax = b, \quad x \geqq 0. \end{aligned} \tag{2.8}$$

The next two results show that optimal solutions can be obtained when each f_i is replaced by a piecewise-linear convex function that coincides with f_i at *some* rather than *all* of the integer points in the interval $[l_i, u_i]$. These more general results suggest the use of "column-generation" strategies in the event that evaluation of the f_i at *all* integer points in the intervals $[l_i, u_i]$ would be "costly". (Details of these "column-generation" procedures are given in § 3).

COROLLARY 2.3. *Let the functions $\hat{f}_i$ $(i = 1, \cdots, n)$ be convex piecewise-linear functions of the form*

$$\begin{aligned} \hat{f}_i(x_i) = \min_{\lambda_{i,j}} & \sum_{j \in R_i''} f_i(j)\lambda_{i,j} \\ \textit{subject to } & \sum_{j \in R_i''} j\lambda_{i,j} = x_i, \\ & \sum_{j \in R_i''} \lambda_{i,j} = 1, \quad \lambda_{i,j} \geqq 0, \end{aligned} \tag{2.9}$$

where each R_i'' is a finite, nonempty subset of the integers. If the optimal value of the problem

$$\begin{aligned} &\min \sum_{i=1}^{n} \hat{f}_i(x_i) \\ &\textit{subject to } Ax = b, \quad x \geqq 0, \quad x \textit{ integer} \end{aligned} \tag{2.10}$$

exists, then it is equal to the optimal value of

$$\begin{aligned} &\min \sum_{i=1}^{n} \hat{f}_i(x_i) \\ &\textit{subject to } Ax = b, \quad x \geqq 0. \end{aligned} \tag{2.11}$$

Proof. Since (2.10) is assumed to have an optimal solution, it is easily seen that (2.11) must also have an optimal solution, and by an argument analogous to the proof of Theorem 2.1, the LP equivalent to (2.11) must have an optimal solution with x integer-valued. □

In the case that the f_i are affine, the equivalence of (2.10) and (2.11) follows from the integrality of the extreme points of F, but note that the Corollary 2.3 cannot be based on this fact since the optimal solutions of (2.11) need not be extreme points of F. It should be recognized, however, that the conclusions of Corollary 2.3 need *not* hold if the $\hat{f}_i$ are general convex functions or if the $\hat{f}_i$ are even piecewise-linear convex functions with "breakpoints" at noninteger points. This is easily seen by letting the constraints $Ax = b$ be given by $x_1 + x_2 = 1$ $(n = 2)$ and letting $\hat{f}_i(x_i) = (x_i - \frac{1}{2})^2$ or $|x_i - \frac{1}{2}|$. Such convex functions must be replaced by

"equivalent" piecewise-linear functions with integral breakpoints before the integrality constraints may be deleted.

We will now show that the optimal value of the problem (2.11) coincides with the optimal value of the nonlinear problem (1.2) if the index sets R_i'' are sufficiently "fine" near an integer optimal solution of (2.11). This result is essentially equivalent to the fact that a local solution of (2.8) must also be a global solution of (2.8).

THEOREM 2.4. *If x^{**} is an integer optimal solution of* (2.11) *and if $R_i'' \supseteq \{x_i^{**}-1, x_i^{**}, x_i^{**}+1\} \cap [l_i, u_i]$ for $i = 1, \cdots, n$, then x^{**} is an optimal solution of the nonlinear integer program* (1.2).

Proof. Since the feasible sets of (2.8) and (2.11) coincide, x^{**} is feasible for (2.8). Moreover, $\hat{f}_i(y) = \tilde{f}_i(y)$ for $y \in [x_i^{**}-1, x_i^{**}+1] \cap [l_i, u_i]$, so x^{**} must be a local minimum of (2.8). Because of convexity, x^{**} is also a global minimum of (2.8), and the conclusion follows from Theorem 2.2 and the equivalence of (2.4) and (2.8). □

It should be noted that it is *not* sufficient for optimality to simply have $R_i'' \supseteq \{x_i^{**}\}$ for all i, as may be seen from following example: consider the following problem of the form (1.2)

$$\begin{aligned} &\min (x_1-1)^2+(x_2-1)^2 \\ &\text{subject to } x_i+x_2=2, \\ &\qquad x_i \geqq 0 \text{ and integer}, \end{aligned}$$

and let $l_i = 0$, $u_i = 2$, $R_i'' = \{0, 2\}$ for $i = 1, 2$; then $\hat{f}_i \equiv 1$ on $[0, 2]$, so that optimal solutions for the corresponding approximating problem occur at $x_1 = 0$, $x_2 = 2$ and $x_1 = 2$, $x_1 = 0$, but the unique optimal solution of the original problem is $x_1 = 1$, $x_2 = 1$.

Finally, note that these results do *not* generalize to the case in which the x_i are *vector* variables rather than the scalar components of x. This may be seen from the following example.

Example. Let $f(x_1, x_2, x_3, x_4) = f_1(x_1, x_2) + f_2(x_3, x_4)$, where f_1 is any convex function such that $f_1(0, 0) = f_1(1, 1) = 1$ and $f_1(0, 1) = f_1(1, 0) = -1$ (for example, f_1 could be taken as $2(x_1+x_2-1)^2-1$ or as a convex piecewise-linear function with those values), and f_2 is any convex function such that $f_2(0, 0) = f_2(1, 1) = -1$ and $f_2(0, 1) = f_2(1, 0) = 1$ (for example, f_2 could be taken as $2(x_3-x_4)^2-1$ or as a convex piecewise-linear function with those values). Consider the problem:

$$\begin{aligned} &\min f_1(x_1, x_2)+f_2(x_3, x_4) \\ &\text{subject to } x_1 \qquad -x_3 \quad = 0, \\ &\qquad\qquad x_2 \qquad -x_4 = 0, \\ &\qquad\qquad 0 \leqq x_i \leqq 1, \\ &\qquad\qquad\quad x_i \text{ integer } (i = 1, \cdots, 4). \end{aligned}$$

It is easily seen that the constraint matrix is totally unimodular. The four feasible points (0, 0, 0, 0), (1, 0, 1, 0), (0, 1, 0, 1), (1, 1, 1, 1) all have objective function

values of 0, and hence are all *optimal solutions*. Suppose, however, we replace the f_i by piecewise-linear *convex* functions $\bar{f}_i$ that agree with the f_i at (0, 0), (0, 1), (1, 0), (1, 1), and delete the integrality constraints to obtain the problem

$$\begin{aligned} &\min \bar{f}_1(x_1, x_2)+\bar{f}_2(x_3, x_4) \\ &\text{subject to } x_1 \qquad -x_3 \quad = 0, \\ &\qquad\qquad x_2 \qquad -x_4 = 0, \\ &\qquad\qquad 0 \leqq x_i \leqq 1 \quad (i = 1, \cdots, 4). \end{aligned}$$

It is easily seen that, as a result of convexity, the objective function of this new problem has a value no greater than -2 at the point $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, so that the deletion of the integrality constraints results in a change in the optimal solution and the optimal value. □

3. Computational considerations. Theorem 2.4 establishes the validity of the following *column-generation* procedure for solving the nonlinear integer program (1.2) under the hypotheses (A)–(D) of § 2: [3]

(3.1) Set the iteration index $k = 0$, and select an initial set of breakpoints

$$R_i^0 \supseteq \{l_i, u_i\} \, (i = 1, \cdots, n).$$

(3.2) Solve the LP (2.11) with $R_i'' = R_i^k$.

(3.3) If the optimal solution obtained for (2.11) satisfies the optimality conditions of Theorem 2.4, then it also solves (1.2), and the algorithm terminates; otherwise, increase k by 1 and add the breakpoints that would have been required to satisfy the breakpoint hypotheses of Theorem 2.4 for the solution obtained in (3.2) (thereby obtaining "finer" index sets R_i^k) and return to (3.2).

Since the maximum possible number of breakpoints is finite, and at least one new breakpoint is added at each iteration, this procedure must terminate in a *finite* number of iterations with an optimal solution of (1.2). As with other column-generation procedures, each succeeding iteration can be started with the optimal basis from the previous iteration. If function evaluations are much more "expensive" than pivot operations, the procedure could be modified by selecting the initial breakpoints close to an estimate of the optimal solution and adding only some of the "missing" breakpoints in step (3.3).

Linear programming can also be used to establish a *lower* bound on the optimal value of (1.2). (A lower bound may be useful if convergence is slow and one is content to have a feasible solution whose objective value is "close" to optimal.) To compute a lower bound, the f_i are replaced by convex, piecewise-linear functions f_i^* satisfying $f_i^*(x_i) \leqq f_i(x_i)$ for $x_i \geqq 0$ and integer. Such f_i^* may be

[3] This procedure may also be used if hypothesis (B) is violated (i.e., $F = \varnothing$), since $F = \varnothing$ if and only if the feasible set of (1.2) is empty, and the first attempt to analyze an LP in (3.2) will establish $F = \varnothing$ if this is the case. The case in which hypothesis (A) is violated (i.e., F is unbounded) is dealt with in § 5, where it is shown that an analogous procedure will converge in a finite number of iterations if (1.2) has an optimal solution.

obtained from a finite, nonempty set $R_i^{(k)}$ of integers ($R_i^{(k)} \subseteq [l_i, u_i - 1]$) by defining

$$f_i^*(x_i) \equiv \min_{z_i, \lambda_{i,j}} z_i$$

$$\text{subject to } z_i \geqq f_i(j) + \lambda_{i,j}(f_i(j+1) - f_i(j)),$$

$$x_i = j + \lambda_{i,j} \quad (j \in R_i^{(k)}).$$

Replacing the f_i by f_i^* and deleting the integrality constraints yields a linear program whose optimal value is a lower bound on the optimal value of (1.2). Conditions guaranteeing finiteness of this lower bound and further details and refinements may be found in [11].

In the case of the problem (1.1), it should be noted that, when the δ-form representation is used, the equivalent LP has such a simple structure that its solution is obvious. In fact, rather than dealing with (1.1), we can consider the more general class of problems of the form

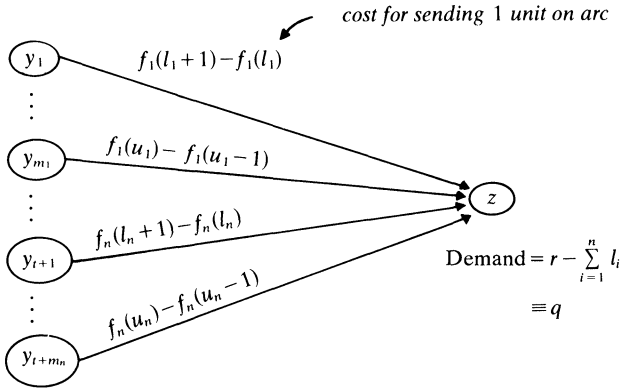
$$\min \sum_{i=1}^{n} f_i(x_i)$$

$$\text{subject to } \sum_{i=1}^{n} x_i = r, \tag{3.4}$$

$$l_i \leqq x_i \leqq u_i \quad (i = 1, \cdots, n),$$

$$x_i \text{ integer} \quad (i = 1, \cdots, n).$$

(Note that (1.1) is equivalent to a problem of the form (3.4) with $l_i = 0$ and $u_i = r$ $(i = 1, \cdots, n)$.) Figure 1 shows the network that results after the arcs corresponding to flows fixed at lower bounds have been accounted for by reducing the



$$m_i = u_i - l_i$$

$$t = \sum_{i=1}^{n-1} m_i$$

FIG. 1. *A network equivalent of* (3.4)

"demand" by $\sum_{i=1}^{n} l_i$. (We assume that the condition $\sum_{i=1}^{n} l_i \leqq r \leqq \sum_{i=1}^{n} u_i$, which is necessary and sufficient for feasibility, has been verified.) This network has $t+m_n$ supply points, each of which can ship at most 1 unit to the demand point z, at which the demand is $q \equiv r - \sum_{i=1}^{n} l_i$. If $q=0$, the optimal solution is obtained by setting $x_i = l_i$ for all i; otherwise, the optimal solution is obtained by sending one unit along each of the q "least expensive" arcs. Because of convexity, for each i we have $f_i(j) - f_i(j-1) \leqq f_i(j+1) - f_i(j)$ for j satisfying $l_i \leqq j-1 \leqq j+1 \leqq u_i$, and using this property it may be shown that the determination of the q least expensive arcs requires at most $2n + (r - \sum_{i=1}^{n} l_i) - 1$ function evaluations. (For further details on (1.1) see [9], [10], [13].)

4. Total unimodularity and networks. On the one hand, it is well-known that the standard minimum cost single-commodity network problem gives rise to a constraint matrix that is totally unimodular. In fact, in dealing with separable convex functions defined on the arcs of a network, it is possible to mimic the δ-formulation within the network context by replacing each arc by a set of arcs in a manner similar to that of Fig. 1 (for details, see [10] or [11]). Given the efficiency of special techniques for network optimization a network formulation will generally be more efficiently solved by such techniques than by using the ordinary simplex algorithm.

On the other hand, it is unclear whether, in all cases, network formulations can be constructed for problems of the form (1.2) with A totally unimodular. (Specifically, we would like to be able to construct from A and b a directed graph with node-arc incidence matrix B and a (possibly unbounded) hyper-rectangle R such that a vector x is feasible for (2.8) if and only if there exists a y such that the pair $\begin{pmatrix} x \\ y \end{pmatrix}$ satisfies $B\begin{pmatrix} x \\ y \end{pmatrix} = 0$, $\begin{pmatrix} x \\ y \end{pmatrix} \in R$. Certain results in matroid theory (see, for example [12], [15]), while not addressing questions of quite this generality, suggest that such a construction is not always possible.) In any event, the linear programming approach of the previous sections makes conversion to a network unnecessary, since it specifies the *algebraic* transformation of the original problem that will yield an equivalent LP.

5. The unbounded case. The results of this section show that the hypotheses regarding boundedness of the feasible set and finiteness of the number of breakpoints can be deleted, provided that the conclusions are appropriately generalized.

THEOREM 5.1. *Let the function $\bar{f}_i$ $(i=1, \cdots, n)$ be continuous piecewise-linear convex functions on $[0, +\infty)$ whose derivatives are also continuous except possibly on subsets of the positive* integers. *Let A be an $m \times n$ totally unimodular matrix, and b be an integer vector. Then the values*

$$\inf \sum_{i=1}^{n} \bar{f}_i(x_i)$$

(5.1)

$$\textit{subject to } Ax = b, \quad x \geqq 0, \quad x \textit{ integer}$$

and

$$\text{(5.2)} \qquad \begin{aligned} &\inf \sum_{i=1}^{n} \bar{f}_i(x_i) \\ &\textit{subject to } Ax = b, \quad x \geqq 0 \end{aligned}$$

coincide, and (5.1) *has an optimal solution if and only if* (5.2) *has an optimal solution. (The value is taken to be* $+\infty$ *when the constraints are infeasible.)*

Proof. If (5.2) has a feasible solution, then the feasible set of (5.2) has an extreme point, which is thus integer and therefore a feasible solution of (5.1).

The conclusions of the theorem are then easily proved by considering feasible or optimal solutions of (5.1) and (5.2), adding appropriate bounds to both problems, and applying the results of § 2. □

It should be noted that it is possible that (5.1) and (5.2) may have finite infima that are not attained, and the theorem shows that if this is the case for one of these problems, it must be true for the other also.

We now state the analogue of Theorem 2.4 in the absence of upper bounds.

THEOREM 5.2. *If the hypotheses of Theorem* 5.1 *are satisfied, and* x^{**} *is an integer optimal solution of* (5.2), *then* x^{**} *is an optimal solution of*

$$\text{(5.3)} \qquad \begin{aligned} &\inf \sum_{i=1}^{n} f_i(x_i) \\ &\textit{subject to } Ax = b, \quad x \geqq 0, \quad x \textit{ integer} \end{aligned}$$

where, for $i = 1, \cdots, n$, f_i *is any convex function that agrees with* $\bar{f}_i$ *on the set* $\{x_i^{**} - 1, x_i^{**}, x_i^{**} + 1\} \cap \{y | y \geqq 0\}$.

Proof. Suppose that there exists an $\bar{x}$ feasible for (5.3) such that $\sum_{i=1}^{n} f_i(\bar{x}_i) < \sum_{i=1}^{n} f_i(x_i^{**})$. Generate bounded variants of (5.2) and (5.3) by adding the constraints $\min\{\bar{x}_i, x_i^{**}\} \leqq x_i \leqq \max\{\bar{x}_i, x_i^{**}\}$ $(i = 1, \cdots, n)$. The bounded variant of (5.3) must have an optimal solution x^* such that $\sum_{i=1}^{n} f_i(x_i^*) < \sum_{i=1}^{n} f_i(x^{**})$. However, by applying Theorem 3.4 to the bounded variants of (5.2) and (5.3), we obtain a contradiction. □

Finite convergence of an extension to the unbounded case of the column-generation procedure of § 3 follows in a straightforward fashion from the next theorem, which applies to a broad class of integer programs, since total unimodularity of the constraint matrix A is *not* required for the result. (For notational convenience we define

$$f(x) \equiv \sum_{i=1}^{n} f_i(x_i) \quad \text{and} \quad F_I \equiv \{x | Ax = b, x \geqq 0, x \text{ integer}\}.$$

Details of an algorithm for the unbounded case are discussed in [11].)

THEOREM 5.3. *If* f_i $(i = 1, \cdots, n)$ *are convex functions on* $[0, +\infty)$ *and if the problem*

$$\text{(5.4)} \qquad \begin{aligned} &\inf \sum_{i=1}^{n} f_i(x_i) \\ &\textit{subject to } Ax = b, \quad x \geqq 0, \quad x \textit{ integer} \end{aligned}$$

has an optimal solution, then, for each real number M, the set $\{f(x)|x \in F_I\}$ contains a finite number (possibly 0) of distinct values in the range $(-\infty, M]$.

Proof. Suppose the result is false, for some M and let $\{x^{(k)}\}$ be a sequence contained in F_I with the property that $\{f(x^{(k)})\}$ is a sequence of distinct values in $(-\infty, M]$. Using the nonnegativity of the $x^{(k)}$, we shall construct an increasing subsequence I_n of integers and a partition J', J'' of the index set $\{1, \cdots, n\}$ such that if $r, s, t \in I_n$ with $r < s < t$, then $0 \leqq x_i^{(s)} - x_i^{(r)} \leqq x_i^{(t)} - x_i^{(s)}$, with strict inequalities holding for $i \in J''$. If $\{x_1^{(k)}\}$ is bounded, then there exists an integer $\bar{x}_1$ such that $x_1^{(k)} = \bar{x}_1$ for infinitely many k; in this case $1 \in J'$ and we denote by I_1 an increasing infinite subsequence of $\{1, 2, \cdots\}$ such that $x_1^{(k)} = \bar{x}_1$. If $\{x_1^{(k)}\}$ is not bounded, $1 \in J''$ and I_1 is taken to be an increasing infinite subsequence of $\{1, 2, \cdots\}$ such that $r, s, t \in I_1$ implies $0 < x_1^{(s)} - x_1^{(r)} < x_1^{(t)} - x_1^{(s)}$. Proceeding in an analogous fashion with the sequence $\{x_2^{(k)} | k \in I_1\}$, we place the index 2 in J' or J'' and extract from I_1 a subsequence I_2, and continue this process until all indices have been placed in J' or J'' and I_n has been extracted from I_{n-1}. Clearly J', J'' and I_n have the required properties, and note that $\{x_i^{(k)}|k \in I_n\}$ is constant for $i \in J'$. J' may be empty, but $J'' \neq \varnothing$ since otherwise $x^{(k)}$ would be constant for $k \in I_n$, contradicting the assumed distinctness of the elements of $\{f(x^{(k)})\}$.

Now $\{f(x^{(k)})|k \in I_n\}$ contains either a decreasing subsequence or an increasing subsequence. If it contains a decreasing subsequence, choose p such that $p \in I_n$ and $x_i^* \leqq x_i^{(p)}$ for $i \in J''$. Let $\Delta \equiv x^{(p+1)} - x^{(p)}$, and note that $\Delta \geqq 0$ and $A\Delta = 0$, so that $(x^* + \Delta) \in F_I$. However, using the convexity of f we have

$$\begin{aligned} f(x^*) - f(x^* + \Delta) &= \sum_{i=1}^{n} [f_i(x_i^*) - f_i(x_i^* + \Delta_i)] \\ &= \sum_{i \in J''} [f_i(x_i^*) - f_i(x_i^* + \Delta_i)] \\ &\geqq \sum_{i \in J''} [f_i(x_i^{(p)}) - f_i(x_i^{(p)} + \Delta_i)] \\ &= f(x^{(p)}) - f(x^{(p+1)}) > 0, \end{aligned}$$

which contradicts the assumed optimality of x^*. In the remaining case, $\{f(x^{(k)})|k \in I_n\}$ contains an increasing sequence, and since this sequence is bounded from above, there exist $r, s, t \in I_n$ such that $r < s < t$ and $f(x^{(t)}) - f(x^{(s)}) < f(x^{(s)}) - f(x^{(r)})$. However, $x^{(t)} - x^{(s)} \geqq x^{(s)} - x^{(r)} \geqq 0$ and the convexity and separability of f imply $f(x^{(t)}) - f(x^{(s)}) \geqq f(x^{(s)}) - f(x^{(r)})$, a contradiction. □

It might be noted that a straightforward extension of this result to the convex, *nonseparable* case is not possible, since it is easily seen that taking $f(x_1, x_2) = (x_1 - \sqrt{2}\, x_2)^2$, $A = 0$, $b = 0$ satisfies all of the hypotheses of the theorem except separability, but violates the conclusion of the theorem.

The following corollary, an immediate consequence of Theorem 5.3, establishes *finite convergence* for any "primal, nondegenerate" method for the class of problems considered in that theorem.

COROLLARY 5.4. *If the hypotheses of Theorem* 5.3 *hold, then any algorithm for the problem* (5.4) *that yields feasible iterates $x^{(0)}, x^{(1)}, \cdots$ satisfying $f(x^0) > f(x^1) > \cdots$ will generate an optimal solution for* (5.4) *in a finite number of iterations.*

6. Conclusions. We have shown how optimal solutions to bounded nonlinear integer programs of the form (1.2) (with f_i convex, A totally unimodular, and b integer) may be obtained by solving an easily-generated linear programming problem. These results generalize certain results in (linear) integer programming dealing with totally unimodular constraint matrices as well as results for nonlinear integer programs of the form (1.1), and provide a rigorous and finite approach for obtaining optimal solutions. Furthermore, in the case that known bounds are not available for the variables in (1.2), it is shown that an appropriate linear programming "column-generation" algorithm will yield an optimal solution in a finite number of iterations if (1.2) actually has an optimal solution.

REFERENCES

[1] E. M. L. BEALE, *An algorithm for solving the transportation problem when the shipping cost over each route is convex*, Naval Res. Logist. Quart., 6 (1959), pp. 43–56.

[2] CLAUDE BERGE AND A. GHOUILA-HOURI, *Programming, Games and Transportation Networks*, John Wiley, New York, 1965.

[3] L. B. BOZA, *The interactive flow simulator: A system for studying personnel flows*, presented at ORSA/TIMS Joint National Meeting, San Juan, Puerto Rico, Oct. 1974.

[4] A. CHARNES AND W. W. COOPER, *Nonlinear network flows and convex programming over incidence matrices*, Naval Res. Logist. Quart., 5 (1958), pp. 231–240.

[5] GEORGE B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.

[6] L. R. FORD, JR. AND D. R. FULKERSON, *Flows in Networks*, Princeton University Press, Princeton, NJ, 1962.

[7] T. C. HU, *Minimum cost flows in convex-cost networks*, Naval Res. Logist. Quart., 13 (1966), pp. 1–9.

[8] ROBERT S. GARFINKEL AND GEORGE L. NEMHAUSER, *Integer Programming*, John Wiley, New York, 1972.

[9] O. GROSS, *Class of discrete type minimization problems*, RM-1644, RAND Corp., Santa Monica, CA, 1956.

[10] R. R. MEYER, *A class of nonlinear integer programs solvable by a single linear program*, Computer Sci. Dept. Tech. Rep. 267, Univ. of Wisc., Madison, 1976.

[11] R. R. MEYER AND M. L. SMITH, *Algorithms for a class of "convex" nonlinear integer programs*, Computer Sci. Dept. Tech. Rep. 274, Univ. of Wisc., Madison, 1976.

[12] GEORGE J. MINTY, *On the axiomatic foundations of the theories of directed linear graphs, electrical networks and network-programming*, J. Math. and Mech., 15 (1966), pp. 485–520.

[13] THOMAS L. SAATY, *Optimization in Integers and Related Extremal Problems*, McGraw-Hill, New York, 1970.

[14] R. E. SCHWARTZ AND C. L. DYM, *An integer maximization problem*, Operations Res., 19 (1971), pp. 548–550.

[15] W. T. TUTTE, *An algorithm for determining whether a given binary matroid is graphic*, Proc. Amer. Math. Soc., 11 (1960), pp. 905–917.

[16] IRAM J. WEINSTEIN AND OLIVER S. YU, *Comment on an integer maximization problem*, Operations Res., 21 (1973), pp. 648–650.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

A WELL-POSED APPROXIMATE METHOD FOR INITIAL STATE DETERMINATION OF DISCRETE-TIME DISTRIBUTED PARAMETER SYSTEMS*

TOSHIHIRO KOBAYASHI†

Abstract. The purpose of this paper is to investigate the problem of initial state determination for a discrete-time distributed parameter system described by a differential-difference equation. This problem is not well-posed in general. After the problem formulation, a well-posed approximate method is presented. This method uniquely gives an approximate initial state which depends continuously on the measurement data. The method is analyzed on the assumption that the system is N output controllable with respect to the initial state. An a posteriori error estimate is also given.

1. Introduction. From the physical viewpoint, the system state functions may not be directly measurable. Only certain restricted measurements are actually obtained. In order to construct feedback controls, however, complete knowledge of the state functions is required. The system state has to be determined 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. In a distributed parameter system, observability ensures that an initial state can be uniquely determined from the measurement data. As the space of initial states is infinite-dimensional, observability does not generally ensure 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 that for a lumped parameter system [6], [7]. Thus an approximate method which reduces the non-well-posed problem to a well-posed one is of great importance.

Difference equations arise and are of utmost importance in the fields of, for example, numerical analysis and sample-data control systems [2], [8]. In the case of discrete-time observations, the distributed parameter systems are not observable in general. Therefore we analyze the problem of initial state determination without the assumption that the system is observable.

2. Problem statement. We shall use notations similar to those in Lions [3]. So let H and V be two Hilbert spaces with

$$V \subset H, \qquad V \text{ dense in } H;$$

the sign $\subset$ denotes both algebraic and topological inclusion. This means that the identity mapping of V into 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

$$V \subset H \subset V'.$$

* Received by the editors December 11, 1975, and in final revised form October 15, 1976.

† Department of Control Engineering, Kyushu Institute of Technology, Tobata, Kitakyushu, Japan.

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.

We are given a continuous bilinear form $a(u, v)$ on V, having the following properties: for any $u, v \in V$,

$$|a(u, v)| \leqq L\|u\|_V \cdot \|v\|_V, \tag{2.1}$$

where L is a constant independent of u, v.

For each fixed u in V, the linear form $v \to a(u, v)$ is continuous on V. Therefore it may be written as

$$a(u, v) = (Au, v), \qquad Au \in V'.$$

We deduce also from (2.1) that for any $u \in V$,

$$\|Au\|_{V'} \leqq L\|u\|_V, \tag{2.2}$$

where $\|\cdot\|_{V'}$ is the dual norm of $\|\cdot\|_V$. Equation (2.2) defines $A \in \mathscr{L}(V; V')$ (the space of continuous linear mappings from V into V').

Now let $a(u, v)$ be coercive, that is,

$$\begin{aligned} &\text{there exists } \alpha > 0 \text{ such that for any } v \in V, \\ &a(v, v) \geqq \alpha\|v\|_V^2. \end{aligned} \tag{2.3}$$

LEMMA 1 (see [3]). *Under the hypotheses* (2.1) *and* (2.3), *for every* $f \in V'$, *the equation*

$$Au = f \tag{2.4}$$

has a unique solution $u \in V$ *which depends continuously on* f. *Furthermore, if* $f \in H$, *the solution* u *belongs to a dense subspace* W *of* H. *Here* W *is a Hilbert space with a norm defined by*

$$\|u\|_W = (\|u\|_V^2 + \|Au\|_H^2)^{1/2}, \qquad u \in W.$$

In this paper, we consider the discrete-time distributed parameter system described by

$$\begin{aligned} Au_k &= u_{k-1}, \quad k \in \sigma = \{1, 2, \cdots, N\}, \\ u_0 &= \xi, \qquad \xi \text{ given in } H, \end{aligned} \tag{2.5}$$

where u_k is the state of the system at time k.

Remark 1. For example, we can obtain the discrete-time system (2.5) by replacing du/dt with a backward difference $(u_k - u_{k-1})/h$ in the evolution equation

$$\frac{du}{dt} + \frac{1}{h}(A - I)u(t) = 0,$$

$$u(0) = \xi,$$

where h is a sampling period and I is an identity operator.

From Lemma 1, there exists an operator U such that $U \in \mathscr{L}(H; W)$ and the solution sequence of the system (2.5) is given by

$$u_k = Uu_{k-1}, \qquad k \in \sigma. \tag{2.6}$$

It follows from (2.6) that for the initial state $\xi \in H$,

$$u_k = U^k \xi, \qquad k \in \sigma. \tag{2.7}$$

In physical situations, the space of observations K is finite-dimensional. We are given an observation equation

$$z_k = Mu_k + m_k, \qquad k \in \sigma, \tag{2.8}$$

where M is a continuous linear operator from W to K and $m_k \in K$ is a measurement error at time k. The observed outputs z_k's are written

$$z_k = MU^k \xi + m_k, \qquad k \in \sigma. \tag{2.9}$$

By virtue of Lemma 1, it follows that the output sequence $\{z_k\}_{k\in\sigma} \in l^2(\sigma; K)$.

Remark 2. $l^2(\sigma; K)$ denotes the Hilbert space consisting of all sequences $\{p_k\}_{k\in\sigma}$ with $p_k \in K$, $k \in \sigma$ and with an inner product defined by: for $p = \{p_k\}$, $q = \{q_k\}$ in $l^2(\sigma; K)$,

$$(p, q)_{l^2(\sigma;K)} = \sum_{k=1}^{N} (p_k, q_k)_K.$$

The space $l^2(\sigma; K)$ is finite dimensional when N is a finite natural number.

We denote by $J(\eta)$ a functional which measures the distance between the observations $z = \{z_k\}$ and the output $Mu = \{Mu_k\}$ computed for each initial state η from the system (2.5). Then, the initial state determination problem can be formulated as that of minimizing $J(\eta)$ with respect to η under the constraint (2.5). In the following, the functional $J(\eta)$ is taken as the mean-square error:

$$J(\eta) = \sum_{k=1}^{N} \|z_k - Mu_k(\eta)\|_K^2. \tag{2.10}$$

3. *N* observability and *N* output controllability with respect to the initial state. In this section, we investigate N observability and N output controllability with respect to the initial state of the discrete-time system (2.5) with the observation equation (2.8). We start with the following definition.

DEFINITION 1. The system described by (2.5) with the observation equation (2.8) is said to be *N output controllable with respect to the initial state* if $\{MU^k\eta\}_{k\in\sigma}$ generates a dense subspace $Q(N)$ of the space $l^2(\sigma; K)$, as η is varied without any constraints.

DEFINITION 2. The system (2.5) with (2.8) is said to be *N observable* if an initial state ξ can be uniquely determined from the observation $\{MU^k\xi\}_{k\in\sigma}$.

Let us define an operator $T: H \to l^2(\sigma; K)$ by

$$T\eta = \{MU^k\eta\}_{k\in\sigma} \tag{3.1}$$

for any $\eta \in H$. We get $T \in \mathcal{L}(H; l^2(\sigma; K))$. Let T^* be the adjoint of T. Now we can prove the following theorems.

THEOREM 1. *The following three conditions are equivalent*:

(i) *the system* (2.5) *with* (2.8) *is N output controllable with respect to the initial state*;

(ii) *the nullspace of* T^* *is* $\{0\}$;

(iii) TT^* *is positive.*

Proof. For any $p=\{p_k\}\in l^2(\sigma; K)$,

$$(Mu(\eta), p)_{l^2(\sigma;K)}=(T\eta, p)_{l^2(\sigma;K)}=\sum_{k=1}^{N}(MU^k\eta, p_k)_K$$

$$=\left(\eta, \sum_{k=1}^{N}(U^k)^*M^*p_k\right)_H \tag{3.2}$$

$$=(\eta, T^*p)_H.$$

Here the adjoint operator T^* of T is defined by

$$T^*p=\sum_{k=1}^{N}(U^k)^*M^*p_k \quad \text{if } p\in l^2(\sigma; K) \tag{3.3}$$

and T^* belongs to $\mathscr{L}(l^2(\sigma; K); H)$. From (3.2), $Q(N)$ is dense in $l^2(\sigma; K)$ if and only if the nullspace of T^* is $\{0\}$.

On the other hand, since

$$\|T^*p\|_H^2=(TT^*p, p)_{l^2(\sigma;K)}, \qquad p\in l^2(\sigma; K),$$

Null $(T^*)=\{0\}$ if and only if $TT^*\in\mathscr{L}(l^2(\sigma; K); l^2(\sigma; K))$ is a positive operator; i.e., for any $p\in l^2(\sigma; K)$,

$$(TT^*p, p)_{l^2(\sigma;K)}\geqq 0 \quad \text{and} \quad (TT^*p, p)=0 \quad \text{implies} \quad p=0.$$

THEOREM 2. *The following three conditions are equivalent*:

(i) *the system* (2.5) *with* (2.8) *is N observable*;

(ii) *the nullspace of T is* $\{0\}$;

(iii) T^*T *is positive.*

Proof. We can get the theorem from the calculation

$$\sum_{k=1}^{N}(MU^k\xi, MU^k\xi)_K=\|T\xi\|^2_{l^2(\sigma; K)}=(T^*T\xi, \xi)_H.$$

Theorem 1 and Theorem 2 show that the concept of N observability is dual to that of N output controllability with respect to the initial state.

4. Minimization of $J(\eta)$. We shall show that the minimizing solutions of $J(\eta)$ exist if the system (2.5) with (2.8) is N output controllable with respect to the initial state.

Since the operators U^k and M are continuous, the functional $J(\eta)$ is differentiable and convex. Hence the necessary condition for optimality is

$$J'(\xi)\eta=0 \quad \text{for all } \eta\in H. \tag{4.1}$$

From this equation, we obtain

$$\sum_{k=1}^{N}(MU^k\xi-z_k, MU^k\eta)_K=0, \tag{4.2}$$

that is,

$$(T^*(T\xi-z), \eta)_H=0. \tag{4.3}$$

Since (4.3) must hold for all $\eta \in H$, the minimizing solution ξ must satisfy

$$T^*(T\xi - z) = 0. \tag{4.4}$$

If the system (2.5) with (2.8) is N output controllable with respect to the initial state, the nullspace of T^* is $\{0\}$. Thus there exists at least one solution $\xi \in H$ such that

$$T\xi = z \quad \text{in } l^2(\sigma; K). \tag{4.5}$$

Remark 3. The element $\xi = T^*(TT^*)^{-1}z$ satisfies (4.5) evidently.

Let X be the set of elements $\xi \in H$ satisfying (4.5). X is a closed subset of H. For $\xi_1, \xi_2 \in X$ and for all $\eta \in H$,

$$J(\xi_k) \leqq J(\eta), \qquad k = 1, 2.$$

Since $J(\eta)$ is a convex functional, for $\theta \in (0, 1)$,

$$J((1-\theta)\xi_1 + \theta\xi_2) \leqq (1-\theta)J(\xi_1) + \theta J(\xi_2) \leqq J(\eta),$$

from which we get $(1-\theta)\xi_1 + \theta\xi_2 \in X$. Therefore X is a closed convex subset of H.

Next suppose that the system (2.5) with (2.8) is N observable. From (4.4), the initial state ξ is uniquely determined by

$$\xi = (T^*T)^{-1}T^*z = G^{-1}T^*z \tag{4.6}$$

from the observation z. Here $G = T^*T$ is the observability operator. However the discrete-time distributed parameter system (2.5) with (2.8) is not necessarily N observable. Therefore, we cannot seek the unique initial state ξ by (4.6) in general. Moreover, even if the system (2.5) with (2.8) is N observable, ξ determined by (4.6) is meaningless in general. It is because G^{-1} is not always continuous and the observation z has always errors which may be very small. The fact that G^{-1} is not always continuous is due to the following theorem.

THEOREM 3. *Suppose that an operator G defined on a Hilbert space H is self-adjoint and positive. Then, its inverse G^{-1} is continuous if and only if G is positive definite* [1], [5]; *that is, there exists a positive constant γ such that*

$$(G\eta, \eta)_H \geqq \gamma \|\eta\|_H^2 \quad \textit{for all } \eta \in H.$$

Remark 4. If H is finite dimensional, it is easily shown that a positive operator is always positive definite. In this case, γ is the minimum eigenvalue of G.

Now we should consider a new approximate method in order to determine uniquely an approximate initial state which depends continuously on the measurement data z.

5. A well-posed approximate method. In this section, by the method of regularization [3], we shall present the approximate method which

(i) chooses a unique element ξ_0 from X

and

(ii) gives a constructive procedure to obtain ξ_0.

This approximate method, from a different point of view, corresponds to approximating the nonnegative observability operator G by a family of positive definite ones.

Remark 5. It should be noticed that the problem (i) is dual to that of minimizing $J(\eta)$ (see [4]). An element $\xi = T^*(TT^*)^{-1}z$ is well-defined and solves the initial state determination problem when $l^2(\sigma; K)$ is a finite-dimensional space. However, it is not easy and not practical in general that we solve the dual problem to seek ξ_0.

To begin with, we consider the problem (i) on the assumption that the system (2.5) with (2.8) is N output controllable with respect to the initial state.

Let g be an element of H and let Λ be an operator of $\mathscr{L}(H; H)$ such that for any $\eta \in H$ and a positive constant κ,

$$(\Lambda\eta, \eta)_H \geqq \kappa\|\eta\|_H^2. \tag{5.1}$$

Then, there exists a unique element $\xi_0 \in X$ such that

$$(\Lambda\xi_0, \xi - \xi_0)_H \geqq (g, \xi - \xi_0)_H \quad \text{for all } \xi \in X, \tag{5.2}$$

since X is a closed convex subset of H (see [3]). In the special case of $\Lambda = I$ (identity operator in H) and $g = 0$, ξ_0 is the element having minimum norm in X and ξ_0 is given by

$$\xi_0 = T^*(TT^*)^{-1}z.$$

Next let us consider the problem (ii). We introduce a regularized functional $J_\varepsilon(\eta)$ corresponding to $J(\eta)$:

$$J_\varepsilon(\eta) = J(\eta) + \varepsilon[(\Lambda\eta, \eta)_H - 2(g, \eta)_H], \qquad \varepsilon > 0. \tag{5.3}$$

With similar arguments for $J(\eta)$, we can see that there exists a unique minimizing solution ξ_ε of $J_\varepsilon(\eta)$ determined by

$$\begin{aligned} \xi_\varepsilon &= (G + \varepsilon\Lambda)^{-1}(T^*z + \varepsilon g) \\ &= G_\varepsilon^{-1}(T^*z + \varepsilon g). \end{aligned} \tag{5.4}$$

Since G_ε is positive definite, G_ε^{-1} is continuous from Theorem 3. Therefore ξ_ε depends continuously on the measurement data z.

Now it should be noticed that ξ_0 and ξ_ε are the minimizing solutions of $J(\eta)$ and $J_\varepsilon(\eta)$ respectively with the measurement error $m = \{m_k\}$. Therefore ξ_0 is not a desired initial state. Let X^0 be the set of elements $\xi \in H$ satisfying

$$T\xi = z^0, \qquad z^0 = \{Mu_k\}. \tag{5.5}$$

Then X^0 is a closed convex subset of H. Define a unique element ξ^* of X^0 by

$$(\Lambda\xi^*, \xi - \xi^*) \geqq (g, \xi - \xi^*), \qquad \xi \in X^0. \tag{5.6}$$

Remark 6. If the system (2.5) with (2.8) is N observable, the element ξ^* is the actual initial state.

We now proceed to prove the following theorem.

THEOREM 4. (i) *ξ_ε depends continuously on the measurement data z.*
Suppose that the system (2.5) *with* (2.8) *is N output controllable with respect to the*

initial state. Then

(5.7) (ii) $\lim_{\varepsilon \to 0} \|\xi_\varepsilon - \xi_0\|_H = 0$.

(iii) *If the measurement error can be estimated by*

$$\|z - z^0\|_{l^2(\sigma;K)} \leqq \delta \quad \textit{and} \quad \frac{\delta}{\sqrt{\varepsilon}} \textit{ goes to } 0 \textit{ as } \varepsilon \to 0, \tag{5.8}$$

then

$$\lim_{\varepsilon,\delta \to 0} \|\xi_\varepsilon - \xi^*\|_H = 0. \tag{5.9}$$

Proof. Theorem 3 and (5.4) give (i) immediately.

Next we show (ii). Combining (4.4) and (5.4), we obtain

$$(G(\xi_\varepsilon - \xi_0), \eta)_H + \varepsilon(\Lambda\xi_\varepsilon, \eta)_H = \varepsilon(g, \eta)_H \quad \text{for any } \eta \in H. \tag{5.10}$$

Putting $\eta = \xi_\varepsilon - \xi_0$ and using the nonnegativeness of G, we have

$$(\Lambda\xi_\varepsilon, \xi_\varepsilon - \xi_0)_H \leqq (g, \xi_\varepsilon - \xi_0)_H, \tag{5.11}$$

or equivalently,

$$(\Lambda\xi_\varepsilon, \xi_\varepsilon)_H \leqq -(g, \xi_0)_H + (g + \Lambda^*\xi_0, \xi_\varepsilon)_H. \tag{5.12}$$

We obtain from (5.1) that for some constants c_1 and c_2,

$$\kappa\|\xi_\varepsilon\|_H^2 \leqq c_1 + c_2\|\xi_\varepsilon\|_H,$$

which implies there exists a positive constant c_3 such that

$$\|\xi_\varepsilon\|_H \leqq c_3.$$

Thus we can extract a subsequence μ from every sequence of $\varepsilon \to 0$ such that $\xi_\mu \to w$ weakly in H. Equation (5.10) becomes

$$(G(w - \xi_0), \eta)_H = 0 \quad \text{for any } \eta \in H.$$

Taking $\eta = w - \xi_0$, we get

$$(G(w - \xi_0), w - \xi_0)_H = 0$$

from which we have

$$T(w - \xi_0) = 0 \quad \text{in } l^2(\sigma; K).$$

This equation means $w \in X$. As $\mu \to 0$, (5.11) becomes

$$(\Lambda w, w - \xi_0)_H \leqq (g, w - \xi_0)_H. \tag{5.13}$$

On the other hand, putting $\xi = w$ in (5.2), we obtain

$$-(\Lambda\xi_0, w - \xi_0)_H \leqq -(g, w - \xi_0)_H. \tag{5.14}$$

From (5.13) and (5.14), adding, we have

$$(\Lambda(w - \xi_0), w - \xi_0)_H \leqq 0.$$

Again using (5.1), we get

$$\kappa\|w-\xi_0\|_H^2 \leqq 0$$

which implies $w=\xi_0$. Here $\{\xi_\mu\}$ is an arbitrary, weakly convergent subsequence and its weak limit ξ_0 does not depend on the subsequence. Therefore the extraction of a subsequence is unnecessary and $\xi_\varepsilon \to \xi_0$ weakly in H.

Moreover from (5.11), we get

$$(\Lambda(\xi_\varepsilon-\xi_0), \xi_\varepsilon-\xi_0) \leqq (g, \xi_\varepsilon-\xi_0)-(\Lambda\xi_0, \xi_\varepsilon-\xi_0).$$

This implies that $\xi_\varepsilon \to \xi_0$ strongly in H.

Finally we evaluate $\|\xi_\varepsilon-\xi^*\|_H$ in order to show (iii). Let us define ξ_ε^* by

$$G_\varepsilon\xi_\varepsilon^* = T^*z^0+\varepsilon g. \tag{5.15}$$

Then

$$\|\xi_\varepsilon-\xi^*\|_H \leqq \|\xi_\varepsilon-\xi_\varepsilon^*\|_H+\|\xi_\varepsilon^*-\xi^*\|_H. \tag{5.16}$$

For the second term on the right-hand side, we can apply the result (ii) in the case of $z=z^0$. Consequently we obtain

$$\lim_{\varepsilon\to 0}\|\xi_\varepsilon^*-\xi^*\|_H = 0. \tag{5.17}$$

As for the first term, we obtain

$$G_\varepsilon(\xi_\varepsilon-\xi_\varepsilon^*) = T^*(z-z^0), \tag{5.18}$$

which means that the element $(\xi_\varepsilon-\xi_\varepsilon^*)$ realizes the lower bound of the functional

$$I(\eta) = \|z-z^0-T\eta\|_{l^2(\sigma;K)}^2+\varepsilon(\Lambda\eta, \eta)_H, \qquad \eta\in H. \tag{5.19}$$

Therefore

$$I(\xi_\varepsilon-\xi_\varepsilon^*) \leqq I(0) = \|z-z^0\|_{l^2(\sigma;K)}^2. \tag{5.20}$$

From this, it follows that

$$\varepsilon(\Lambda(\xi_\varepsilon-\xi_\varepsilon^*), \xi_\varepsilon-\xi_\varepsilon^*)_H \leqq \|z-z^0\|_{l^2(\sigma;K)}^2. \tag{5.21}$$

If we can evaluate the measurement error by (5.8), we have

$$\varepsilon\kappa\|\xi_\varepsilon-\xi_\varepsilon^*\|_H^2 \leqq \delta^2$$

from (5.1) and (5.21). This becomes

$$\|\xi_\varepsilon-\xi_\varepsilon^*\|_H \leqq \frac{\delta}{\sqrt{\kappa\varepsilon}}. \tag{5.22}$$

The right-hand side of (5.22) tends to 0 as $\varepsilon, \delta \to 0$ with a relation $\delta = o(\sqrt{\varepsilon})$ (δ has a higher order than $\sqrt{\varepsilon}$). Thus we have shown (iii).

6. A posteriori estimate for $\|\xi_\varepsilon-\xi^*\|_H$. In this section, we shall give an a posteriori estimate for $\|\xi_\varepsilon-\xi^*\|_H$ by evaluating $\|(T^*)^{-1}\|$ in the case of $g=0$.

THEOREM 5. *Suppose the system* (2.5) *with* (2.8) *is N output controllable with respect to the initial state. If ε is chosen such that $\sqrt{\kappa}-\sqrt{\varepsilon}\|T^{*-1}\|\cdot\|\Lambda\|>0$, the value*

$\|\xi_\varepsilon - \xi^*\|_H$ can be estimated by

$$\|\xi_\varepsilon - \xi^*\|_H \leqq \frac{\frac{\delta}{\sqrt{\varepsilon}} + \frac{\sqrt{\varepsilon}}{\sqrt{\lambda}}\|\Lambda\| \cdot \|\xi_\varepsilon\|_H}{\sqrt{\kappa} - \frac{\sqrt{\varepsilon}}{\sqrt{\lambda}}\|\Lambda\|}, \tag{6.1}$$

where λ is the minimum eigenvalue of the operator TT^*.

Proof. Let us put $\xi^* = \xi_\varepsilon + y_\varepsilon$ in the identity $G\xi^* = T^* z^0$. Then we have

$$G(\xi_\varepsilon + y_\varepsilon) = T^* z^0.$$

From this, it follows that

$$G_\varepsilon \xi_\varepsilon - T^* z + G_\varepsilon y_\varepsilon - \varepsilon\Lambda(\xi_\varepsilon - y_\varepsilon) + T^*(z - z^0) = 0.$$

Since $G_\varepsilon \xi_\varepsilon - T^* z = 0$ from (5.4) in the case of $g = 0$, we get

$$G_\varepsilon y_\varepsilon - \varepsilon\Lambda(\xi_\varepsilon + y_\varepsilon) - T^*(z^0 - z) = 0.$$

This equation means that the element y_ε realizes the lower bound of the functional

$$L(\eta) = \|z^0 - z + \varepsilon(T^*)^{-1}\Lambda(\xi_\varepsilon + y_\varepsilon) - T\eta\|^2_{l^2(\sigma;K)} + \varepsilon(\Lambda\eta, \eta)_H. \tag{6.2}$$

Here $(T^*)^{-1}$ exists if the system (2.5) with (2.8) is N output controllable with respect to the initial state. Then

$$L(y_\varepsilon) \leqq L(0) = \|z^0 - z + \varepsilon(T^*)^{-1}\Lambda(\xi_\varepsilon + y_\varepsilon)\|^2_{l^2(\sigma;K)}. \tag{6.3}$$

We obtain from this

$$\varepsilon(\Lambda y_\varepsilon, y_\varepsilon)_H \leqq \|z^0 - z + \varepsilon(T^*)^{-1}\Lambda(\xi_\varepsilon + y_\varepsilon)\|^2_{l^2(\sigma;K)}. \tag{6.4}$$

By virtue of (5.1), (6.4) becomes

$$\kappa\varepsilon\|y_\varepsilon\|^2_H \leqq \|z_0 - z + \varepsilon(T^*)^{-1}\Lambda(\xi_\varepsilon + y_\varepsilon)\|^2_{l^2(\sigma;K)}. \tag{6.5}$$

Moreover

$$\sqrt{\kappa\varepsilon}\,\|y_\varepsilon\|_H \leqq \|z - z^0\|_{l^2(\sigma;K)} + \varepsilon\|(T^*)^{-1}\| \cdot \|\Lambda\|(\|\xi_\varepsilon\|_H + \|y_\varepsilon\|_H). \tag{6.6}$$

Since the space $l^2(\sigma; K)$ is finite dimensional in our case, we can evaluate $\|(T^*)^{-1}\|$. From Theorem 1, TT^* is a positive operator on $l^2(\sigma; K)$. There exists the minimum eigenvalue $\lambda > 0$ of TT^* such that for any $p \in l^2(\sigma; K)$

$$\|T^* p\|^2_H = (TT^* p, p)_{l^2(\sigma;K)} \geqq \lambda\|p\|^2_{l^2(\sigma;K)}$$

from Remark 4. Thus we obtain

$$\|T^* p\|_H \geqq \sqrt{\lambda}\|p\|_{l^2(\sigma;K)} \quad \text{if } p \in l^2(\sigma; K). \tag{6.7}$$

Furthermore for any $\eta \in H$,

$$\|\eta\|_H = \|T^*(T^*)^{-1}\eta\|_H \geqq \sqrt{\lambda}\|(T^*)^{-1}\eta\|_{l^2(\sigma;K)}$$

from (6.7). Therefore we get

$$\|(T^*)^{-1}\| \leqq \frac{1}{\sqrt{\lambda}}. \tag{6.8}$$

If ε is chosen as $\sqrt{\kappa}-\sqrt{\varepsilon}\|(T^*)^{-1}\| \cdot \|\Lambda\| > 0$, the a posteriori error estimate (6.1) is obtained from (6.6) and (6.8).

7. The system with pointwise observation. We shall apply the theory developed in the preceding sections to the simple system with pointwise observations.

Consider the system described by the following differential difference equation:

$$\left(-a\frac{d^2}{dx^2}+qI\right)u_k(x) = u_{k-1}(x), \qquad k \in \sigma, \quad x \in \Omega = (0, 1), \tag{7.1}$$

where a is a positive constant and $q \geqq 1$. The boundary condition is given by

$$\frac{du_k(0)}{dx} = \frac{du_k(1)}{dx} = 0, \qquad k = 0, 1, \cdots, N. \tag{7.2}$$

The initial condition is

$$u_0(x) = \xi(x), \qquad x \in \bar{\Omega} = [0, 1]. \tag{7.3}$$

As two Hilbert spaces H and V, we choose [3]

$$H = L^2(\Omega);$$

$$V = H^1(\Omega) = \left\{ v \in L^2(\Omega) \text{ such that } \frac{dv}{dx} \in L^2(\Omega) \right\}.$$

In this case, since

$$(Au_k, u_k) = -\int_0^1 a\frac{d^2u_k(x)}{dx^2}u_k(x)\,dx + q\int_0^1 u_k^2(x)\,dx$$

$$= a\int_0^1 \left(\frac{du_k(x)}{dx}\right)^2 dx + q\int_0^1 u_k^2(x)\,dx,$$

$$(Au_k, u_k) \leqq \max(a, q)\|u_k\|^2_{H^1(\Omega)}.$$

Therefore the condition $A \in \mathscr{L}(V; V')$ is satisfied. Moreover α in (2.3) is $\alpha = \min(a, q)$.

Using the eigenvalues $\{\lambda_n\} = \{n^2\pi^2 a + q\}$ and the eigenfunctions $\{\phi_n(x)\} = \{1, \sqrt{2}\cos \pi x, \sqrt{2}\cos 2\pi x, \cdots\}$, we can express the solution of the system (7.1) and (7.2) by

$$u_k(x) = \sum_{n=0}^{\infty} \frac{h_n\phi_n(x)}{(\lambda_n)^k}, \qquad k \in \sigma, \tag{7.4}$$

$$h_n = \int_0^1 h(x)\phi_n(x)\,dx, \qquad n = 0, 1, \cdots,$$

for any initial state $h \in L^2(0, 1)$. Thus the right-hand side of (7.4) is $U^k h$.

Now let the observation equation be

$$z_k = u_k(y_k) + m_k, \qquad y_k \in [0, 1], \quad k \in \sigma. \tag{7.5}$$

Here y_k is an observation point at time k. In this case K is a one dimensional Euclidean space. The operator M in (2.8) is defined by

$$Mu_k = \int_0^1 \delta(x - y_k) u_k(x)\, dx, \qquad u_k \in H^1(0, 1). \tag{7.6}$$

Since $\int_0^1 \delta(x - y_k)(\cdot)\, dx$ is a continuous linear functional on $H^1(0, 1)$ and $H^1(0, 1) \subset C(0, 1)$, the pointwise observation (7.6) is meaningful.

From (7.4) and (7.6), it follows that

$$MU^k h = \sum_{n=0}^{\infty} \frac{h_n \phi_n(y_k)}{(\lambda_n)^k}. \tag{7.7}$$

Thus we obtain

$$U^{k*} M^* z_k = \sum_{n=0}^{\infty} \frac{\phi_n(y_k)\phi_n(x)}{(\lambda_n)^k} z_k. \tag{7.8}$$

In order to show that the system (7.1), (7.2) and (7.5) is N output controllable with respect to the initial state, it is sufficient to show that TT^* is positive. TT^* is an $N \times N$ symmetric matrix such as

$$TT^* = ((t_{jk})), \qquad t_{jk} = \sum_{n=0}^{\infty} \frac{\phi_n(y_k)\phi_n(y_j)}{(\lambda_n)^{k+j}}, \qquad j, k = 1, \cdots, N. \tag{7.9}$$

Define N vectors $v_1, v_2, \cdots, v_N$ by

$$v_k = \left(\frac{\phi_0(y_k)}{(\lambda_0)^k}, \frac{\phi_1(y_k)}{(\lambda_1)^k}, \cdots \right)^T, \qquad k = 1, \cdots, N. \tag{7.10}$$

Then TT^* is the Gram matrix of $v_1, v_2, \cdots, v_N$ and $t_{jk} = (v_j, v_k)_{l^2}$ $(j, k = 1, \cdots, N)$. If the vectors $v_1, v_2, \cdots, v_N$ are linearly independent, TT^* is nonsingular, that is, positive. Consider the case $y_1 = y_2 = \cdots = y_N = y^*$. If there are at least N nonnegative integers n such as $\phi_n(y^*) \neq 0$, the vectors $v_1, v_2, \cdots, v_N$ are linearly independent. In this case, the system (7.1), (7.2) and (7.5) is N output controllable with respect to the initial state.

As for the a posteriori estimate (6.1), λ in (6.1) is the minimum eigenvalue of the $N \times N$ symmetric, positive matrix TT^*.

Finally we show that G is not positive definite in this case. Let

$$s = \inf_{h \in H} \frac{(Gh, h)_H}{(h, h)_H} \tag{7.11}$$

and then show $s = 0$. Putting an eigenfunction ϕ_n into h, we have

$$\frac{(Gh, h)}{(h, h)} = \frac{(G\phi_n, \phi_n)}{\|\phi_n\|^2} = \sum_{k=1}^{N} (MU^k \phi_n, MU^k \phi_n) = \sum_{k=1}^{N} \frac{\phi_n^2(y_k)}{(\lambda_n)^{2k}}. \tag{7.12}$$

The right-hand side tends to zero as $n \to \infty$. Thus G is not positive definite.

From the above facts, we should apply the well-posed approximate method presented in § 5 for the system (7.1), (7.2) and (7.3), in order to determine uniquely an initial state dependent continuously on the measurement data.

8. Conclusions. In this paper, we have investigated the initial state determination problem for a discrete-time distributed parameter system. An approximate method has been presented in order to determine the approximate initial state continuously dependent on the measurement data. We have analyzed this method on the assumption that the system is N output controllable with respect to the initial state.

Lastly we state N observability for the system discussed in the last section. This system is N observable if and only if for any $h \in L^2(\Omega)$, $\sum_{n=0}^{\infty} h_n \phi_n(y_k)/(\lambda_n)^k = 0$ $(k = 1, 2, \cdots, N)$ implies $h_n = 0$ $(n = 0, 1, \cdots)$. This fact does not hold in general. However, it is expected that for any $\varepsilon > 0$, there exists a number N such that $|h_n| \leqq \varepsilon$ $(n = 0, 1, \cdots, N-1)$ under a suitable assumption on the observation points. This is the concept of ε, N-mode observability.

REFERENCES

[1] S. G. KLEIN, *Linear Differential Equations in Banach Spaces*, Nauka, Moscow, 1967.

[2] K. Y. LEE, S. CHOW AND R. BARR, *On the control of discrete-time distributed parameter systems*, this Journal, 10 (1972), pp. 361–367.

[3] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[4] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[5] S. G. MIKHLIN, *The Problem of the Minimum of a Quadratic Functional*, Holden-Day, San Francisco, 1965.

[6] S. ROLEWICZ, *On optimal observability of linear systems with infinite-dimensional states*, Studia Math., 48 (1972), pp. 411–416.

[7] Y. SAKAWA, *Observability and related problems for partial differential equations of parabolic type*, this Journal, 13 (1975), pp. 15–27.

[8] J. ZABCZYK, *Remarks on the control of discrete-time distributed parameter systems*, this Journal, 12 (1974), pp. 721–735.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

SEMISMOOTH AND SEMICONVEX FUNCTIONS IN CONSTRAINED OPTIMIZATION*

ROBERT MIFFLIN†

Abstract. We introduce semismooth and semiconvex functions and discuss their properties with respect to nonsmooth nonconvex constrained optimization problems. These functions are locally Lipschitz, and hence have generalized gradients. The author has given an optimization algorithm that uses generalized gradients of the problem functions and converges to stationary points if the functions are semismooth. If the functions are semiconvex and a constraint qualification is satisfied, then we show that a stationary point is an optimal point.

We show that the pointwise maximum or minimum over a compact family of continuously differentiable functions is a semismooth function and that the pointwise maximum over a compact family of semiconvex functions is a semiconvex function. Furthermore, we show that a semismooth composition of semismooth functions is semismooth and give a type of chain rule for generalized gradients.

1. Introduction. In this paper we are interested in an inequality constrained optimization problem where the functions need not be differentiable or convex. More precisely, consider the problem of finding an $x \in R^n$ to

$$\text{minimize} \quad f(x)$$

$$\text{subject to} \quad h_i(x) \leqq 0 \quad \text{for } i = 1, 2, \cdots, m$$

where $h_1, h_2, \cdots, h_m$ and f are real-valued functions defined on R^n.

We utilize the "generalized gradient" introduced by Clarke [1], [2] for "locally Lipschitz" functions. A necessary condition [2] (of the Karush [5]–John [4] type) for optimality of a point $\bar{x}$ is that the zero vector is a certain convex combination of generalized gradients of $h_1, h_2, \cdots, h_m$ and f at $\bar{x}$. In § 5 of this paper, this "stationarity" condition is concisely stated in terms of a map as given by Merrill [10] depending on the problem function generalized gradients. Our implementable algorithm for nonsmooth nonconvex optimization given in [11] uses this map and converges to such stationary points if the problem functions are "semismooth" as defined here in § 2. This algorithm can be viewed as a modification and extension of the "conjugate subgradient" type algorithms for nondifferentiable unconstrained optimization given by Lemarechal [8] and Wolfe [16] for convex functions and by Feuer [3] for min-max objectives.

Semismooth functions possess a semicontinuous relationship between their generalized gradients and directional derivatives. They are related to, but different from, the "almost differentiable" functions of Shor [14]. Notable examples of such functions are convex, concave and continuously differentiable functions.

In § 2 we also define "semiconvex" functions. These functions are "quasidifferentiable" (Pshenichnyi [12]) and essentially "semiconvexe" in the sense of Tuy [15] and, if also differentiable, are "pseudoconvex" (Mangasarian

* Received by the editors January 1, 1977.

† International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria. This research was done in part at the University of Oslo while the author was on leave from Yale University. This work was supported in part by the Air Force Office of Scientific Research, Air Force Systems Command, under Grant AFOSR-74-2695.

[9]). In § 5 we show that the above stationarity condition is sufficient for optimality if the problem functions are semiconvex and a constraint qualification is satisfied. This is a nondifferentiable analogue of a sufficient optimality result in [9, Thm. 10.1.1].

In §§ 3 and 4, we give some important properties of semismooth and semiconvex functions. Starting from the work in [1] and [3] on min-max objectives, we show that the pointwise maximum or minimum over a compact family of continuously differentiable functions is a semismooth function. We also give an example of a semismooth function that is an extremal combination not of continuously differentiable functions, but of semismooth functions. This leads us to show that a semismooth composition of semismooth functions is semismooth and to give a type of "chain rule" for generalized gradients. Special cases of this chain rule may be found in [2].

In § 3 we also show that the pointwise maximum over a compact family of semiconvex functions is a semiconvex function. Thus, semiconvex functions behave as do convex functions with respect to the maximization operation, while pseudoconvex functions do not because of the loss of differentiability due to this nonsmooth operation.

2. Definitions and examples of semismooth and semiconvex functions. Let B be an open subset of R^n and $F: R^n \to R$ be *Lipschitz* on B: i.e. there exists a positive number K such that

$$|F(y)-F(z)| \leqq K|y-z| \quad \text{for all } y, z \in B.$$

If F is Lipschitz on each bounded subset of R^n then F is called *locally Lipschitz*.

Let $x \in B$ and $d \in R^n$. As in Clarke [2], let

$$F^0(x; d) = \limsup_{\substack{h \to 0 \\ t \downarrow 0}} [F(x+h+td)-F(x+h)]/t$$

and let $\partial F(x)$ denote the *generalized gradient* of F at x defined by

$$\partial F(x) = \{g \in R^n : \langle g, d\rangle \leqq F^0(x; d) \text{ for all } d \in R^n\}.$$

The following two propositions collect together useful properties of F^0 and ∂F from Clarke [1], [2] and Lebourg [7], respectively.

PROPOSITION 1.

(a) *$\partial F(x)$ is a nonempty convex compact subset of R^n.*

(b) $F^0(x; d) = \max [\langle g, d\rangle : g \in \partial F(x)]$.

(c) *F is differentiable almost everywhere in B and $\partial F(x)$ is the convex hull of all the points g of the form*

$$g = \lim_{k \to \infty} \nabla F(x_k)$$

where $\{x_k\} \to x$ and F has a gradient ∇F at each $x_k \in B$.

(d) *If $\{x_k\} \subset B$ converges to x and $g_k \in \partial F(x_k)$ for each k then $|g_k| \leqq K$ and each accumulation point g of $\{g_k\}$ satisfies $g \in \partial F(x)$; i.e. ∂F is bounded on bounded subsets of B and ∂F is* uppersemicontinuous *on B.*

PROPOSITION 2. *Let y and z be in a convex subset of B. Then there exists $\lambda \in (0, 1)$ and $g \in \partial F(y + \lambda(z - y))$ such that*

$$F(z) - F(y) = \langle g, z - y \rangle;$$

i.e. a mean value result holds.

By combining part (d) of Proposition 1 with Proposition 2 one may easily establish the following useful result:

LEMMA 1. *Let $\{t_k\} \downarrow 0$, $\{h_k\} \to 0 \in R^n$ and F^* be any accumulation point of*

$$\{[F(x + h_k + t_k d) - F(x + h_k)]/t_k\}.$$

Then there exists $g \in \partial F(x)$ such that

$$F^* = \langle g, d \rangle.$$

If $\lim_{t \downarrow 0} [F(x + td) - F(x)]/t$ exists it is denoted by $F'(x; d)$ and called the *directional derivative* of F at x in the direction d. If $F'(x; d)$ exists and equals $F^0(x; d)$ for each $d \in R^n$ then F is said to be *quasidifferentiable* at x (Pshenichnyi [12]). Note that if $F'(x; d)$ exists then, by Lemma 1, there exists $g \in \partial F(x)$ such that

$$F'(x; d) = \langle g, d \rangle$$

and, if, in addition, F is quasidifferentiable at x, then, by parts (a) and (b) of Proposition 1, g is a maximizer of $\langle \cdot, d \rangle$ over $\partial F(x)$.

DEFINITION 1. $F: R^n \to R$ is *semismooth* at $x \in R^n$ if

(a) F is Lipschitz on a ball about x and

(b) for each $d \in R^n$ and for any sequences $\{t_k\} \subset R_+$, $\{\theta_k\} \subset R^n$ and $\{g_k\} \subset R^n$ such that

$$\{t_k\} \downarrow 0, \quad \{\theta_k / t_k\} \to 0 \in R^n \quad \text{and} \quad g_k \in \partial F(x + t_k d + \theta_k),$$

the sequence $\{\langle g_k, d \rangle\}$ has exactly one accumulation point.

LEMMA 2. *If F is semismooth at x then for each $d \in R^n$, $F'(x; d)$ exists and equals $\lim_{k \to \infty} \langle g_k, d \rangle$ where $\{g_k\}$ is any sequence as in Definition 1.*

Proof. Suppose $\{\tau_k\} \downarrow 0$. By Proposition 2, there exist $t_k \in (0, \tau_k)$ and $g_k \in \partial F(x + t_k d)$ such that

$$F(x + \tau_k d) - F(x) = \tau_k \langle g_k, d \rangle.$$

Then, by Definition 1 with $\theta_k = 0 \in R^n$, since $\{t_k\} \downarrow 0$,

$$\lim_{k \to \infty} [F(x + \tau_k d) - F(x)]/\tau_k = \lim_{k \to \infty} \langle g_k, d \rangle.$$

Since $\{\tau_k\}$ is an arbitrary positive sequence converging to zero, $F'(x; d)$ exists and equals the desired limit. □

DEFINITION 2. Let X be a subset of R^n. $F: R^n \to R$ is *semiconvex* at $x \in X$ (with respect to X) if

(a) F is Lipschitz on a ball about x,

(b) F is quasidifferentiable at x and

(c) $x + d \in X$ and $F'(x; d) \geqq 0$ imply $F(x + d) \geqq F(x)$.

Tuy's [15] earlier concept of semiconvexity does not include quasidifferentiability, but we include it in order to obtain Theorems 8 and 9 given below. A

semiconvex function that is also differentiable is called "pseudoconvex" (Mangasarian [9, Chap. 9]).

We say that F is semismooth (quasidifferentiable, semiconvex) on $X \subset R^n$ if F is semismooth (quasidifferentiable, semiconvex) at each $x \in X$. We denote the convex hull of a set S by conv S.

From convex analysis [13, §§ 23 and 24] and [2, Proposition 3] we have the following:

PROPOSITION 3. *If $F: R^n \to R$ is convex (concave) then $F(F)$ is locally Lipschitz,*

$$\partial F(x) = \{g \in R^n : F(y) \geqq (\leqq) F(x) + \langle g, y - x \rangle \text{ for all } y \in R^n\} \quad \text{for each } x \in R^n;$$

i.e. ∂F is the subdifferential *of F, $F(-F)$ is semiconvex on R^n and $F(F)$ is semismooth on R^n.*

From [2, Proposition 4] and the properties of continuously differentiable functions we have the following:

PROPOSITION 4. *If $F: R^n \to R$ is continuously differentiable then F is locally Lipschitz, $\partial F(x) = \{\nabla F(x)\}$ for each $x \in R^n$, and F is quasidifferentiable and semismooth on R^n.*

An example of a locally Lipschitz function on R that is not semismooth (nor quasidifferentiable) is the following differentiable function that is not continuously differentiable:

$$F(x) = \begin{cases} x^2 \sin(1/x) & \text{for } x \neq 0, \\ 0 & \text{for } x = 0. \end{cases}$$

Note that $F'(0; 1) = 0$ and $\partial F(0) = \text{conv}\{-1, 1\}$ is the set of possible accumulation points of $F'(x; 1)$ as $x \downarrow 0$.

An example of a function that is semiconvex and semismooth on R, but not convex nor differentiable, is

$$F(x) = \log(1 + |x|)$$

where

$$\partial F(x) = \begin{cases} 1/(1+x) & \text{for } x > 0, \\ \text{conv}\{-1, 1\} & \text{for } x = 0, \\ -1/(1-x) & \text{for } x < 0. \end{cases}$$

Note that in a neighborhood of $x = 0$

$$F(x) = \max[\log(1+x), \log(1-x)];$$

i.e. F is a pointwise maximum of smooth functions. General functions of this type are the subject of the next section.

3. Semismooth and semiconvex extremal-valued functions. In this section we supplement developments in Feuer [3] and Clarke [1] to show that certain extremal-valued functions E are semismooth and/or semiconvex.

Suppose $E: R^n \to R$ is defined on B, an open subset of R^n, as follows in terms of $f: R^n \times T \to R$ where T is a topological space:

Suppose there exists a sequentially compact subspace U of T such that

(a) $f(x, u)$ is continuous for $(x, u) \in B \times U$,

(b) $f(x, u)$ is Lipschitz for $x \in B$ uniformly for $u \in U$,

(c) $\partial_x f(x, u)$ is uppersemicontinuous for $(x, u) \in B \times U$

and for each $x \in B$ either

(d) $E(x) = \max [f(x, u): u \in U]$ and

(e) $f'_x(x, u; d) = f^0_x(x, u; d)$ for all $(u, d) \in U \times R^n$

or

(d′) $E(x) = \min [f(x, u): u \in U]$ and

(e′) $f'_x(x, u; d) = -f^0_x(x, u; -d)$ for all $(u, d) \in U \times R^n$.

For each $x \in B$ let

$$A(x) = \{u \in U: E(x) = f(x, u)\}.$$

Note that E and A are well defined by the continuity and compactness assumptions. Furthermore, for each $x \in B$, $A(x)$ is compact and $\partial_x f(x, \cdot)$ is uppersemicontinuous and bounded on U, and a direct consequence of [1, Thm. 2.1] is the following:

THEOREM 1. *Let the above assumptions on E and f hold. Then E is Lipschitz on B and for each* $x \in B$

$$\partial E(x) = \operatorname{conv} \{\partial_x f(x, u): u \in A(x)\},$$

and for each $d \in R^n$

$$E'(x; d) = E^0(x; d) = \max [\langle g, d \rangle: g \in \partial_x f(x, u), u \in A(x)]$$

if (d) *and* (e) *hold, or*

$$E'(x; d) = -E^0(x; -d) = \min [\langle g, d \rangle: g \in \partial_x f(x, u), u \in A(x)]$$

if (d′) *and* (e′) *hold.*

Remark. Feuer [3] shows the results of Theorem 1 under the stronger assumptions of our next theorem and proves a result [3, p. 57] close to semismoothness from which our next proof is adapted.

THEOREM 2. *Suppose that* (a) *and* (d) *or* (d′) *hold and that* $f(\cdot, u)$ *is differentiable on B for each* $u \in U$ *and* $\nabla_x f(\cdot, \cdot)$ *is continuous and bounded on* $B \times U$. *Then E is semismooth on B.*

Proof. Note that the additional assumption implies (b), (c), (e), and (e′) and that $\partial_x f = \nabla_x f$ on $B \times U$. Suppose E has the max form (d). (The proof of semismoothness for the min form (d′) is similar.) Let $x \in B$, $d \in R^n$, $x_k = x + t_k d + \theta_k$ and $g_k \in \partial E(x_k)$ where $\{t_k\} \downarrow 0$ and $\{\theta_k / t_k\} \to 0 \in R^n$. From Theorem 1 and Proposition 1 we have that

$$E'(x; d) = E^0(x; d) = \max [\langle g, d \rangle: g \in \partial E(x)]$$

and ∂E is bounded and uppersemicontinuous on a ball about x, so

$$\limsup_{k \to \infty} \langle g_k, d \rangle \leqq E'(x; d).$$

Suppose

$$\liminf_{k\to\infty} \langle g_k, d\rangle < E'(x; d),$$

i.e. there is an $\varepsilon > 0$ and a subsequence of $\{g_k\}$ such that on this subsequence

$$\{\langle g_k, d\rangle\} \to E'(x; d) - \varepsilon. \tag{3.1}$$

For each k corresponding to this subsequence choose $\bar{g}_k \in \partial E(x_k)$ and $u_k \in A(x_k)$ such that

$$\bar{g}_k = \nabla_x f(x_k, u_k) \in \operatorname{conv}\{\nabla_x f(x_k, u): u \in A(x_k)\} = \partial E(x_k)$$

and

$$\langle \bar{g}_k, d\rangle = \min[\langle g, d\rangle: g \in \partial E(x_k)] \leqq \langle g_k, d\rangle. \tag{3.2}$$

Since $\nabla_x f$ is continuous on $B \times U$, $\{x_k\} \to x$ and $\{u_k\}$ is in the compact set U, $\{\bar{g}_k\}$ and $\{u_k\}$ have accumulation points $\bar{g}$ and $\bar{u}$, respectively, such that

$$\bar{g} = \nabla_x f(x, \bar{u}).$$

Thus, by (3.1) and (3.2),

$$\langle \nabla_x f(x, \bar{u}), d\rangle = \langle \bar{g}, d\rangle \leqq E'(x; d) - \varepsilon.$$

Let $u^* \in A(x)$ be such that

$$E'(x; d) = E^0(x; d) = \max[\langle \nabla_x f(x, u), d\rangle: u \in A(x)] = \langle \nabla_x f(x, u^*), d\rangle.$$

Then

$$\langle \nabla_x f(x, \bar{u}), d\rangle \leqq \langle \nabla_x f(x, u^*), d\rangle - \varepsilon$$

and, since $\langle \nabla_x f(\cdot, \cdot), \cdot\rangle$ is continuous, there exist neighborhoods $B(x)$, $V(\bar{u})$ and $D(d)$ such that

$$\langle \nabla_x f(z, u), \delta\rangle \leqq \langle \nabla_x f(z, u^*), \delta\rangle - \varepsilon/2 \quad \text{for all } (z, u, \delta) \in B(x) \times V(\bar{u}) \times D(d).$$

Choose k so large that $u_k \in V(\bar{u})$, $t_k|d| + |\theta_k|$ is less than the radius of a ball about x contained in $B(x)$ and $2|\theta_k/t_k|$ is less than the radius of a ball about d contained in $D(d)$. Then for all $t \in [0, t_k]$,

$$x(t) \equiv x + td + (t/t_k)^2 \theta_k \in B(x),$$

and

$$x'(t) = d + 2(t/t_k)(\theta_k/t_k) \in D(d).$$

Then

$$\langle \nabla_x f(x(t), u_k), x'(t)\rangle \leqq \langle \nabla_x f(x(t), u^*), x'(t)\rangle - \varepsilon/2 \quad \text{for all } t \in [0, t_k].$$

Integrating from $t = 0$ to $t = t_k$ gives

$$f(x(t_k), u_k) - f(x(0), u_k) \leqq f(x(t_k), u^*) - f(x(0), u^*) - t_k\varepsilon/2.$$

But $x(t_k) = x_k$, $x(0) = x$, $u_k \in A(x_k)$ and $u^* \in A(x)$, so

$$E(x_k) - f(x, u_k) \leqq f(x_k, u^*) - E(x) - t_k\varepsilon/2,$$

or

$$E(x_k) + E(x) \leqq f(x_k, u^*) + f(x, u_k) - t_k\varepsilon/2.$$

But this leads to a contradiction, because $f(x_k, u^*) \leqq E(x_k)$, $f(x, u_k) \leqq E(x)$, $t_k > 0$ and $\varepsilon > 0$. Thus, $\lim_{k\to\infty} \langle g_k, d\rangle = E'(x; d)$, so E is semismooth at x. □

THEOREM 3. *Let X be a subset of B. Suppose that* (a), (b), (c), (d), *and* (e) *hold, i.e. E is a max function, and suppose that $f(\cdot, u)$ is semiconvex at $x \in X$ (with respect to X) for each $u \in U$. Then E is semiconvex at $x \in X$ (with respect to X).*

Proof. By Theorem 1, E is Lipschitz on a ball about x, quasidifferentiable at x, and for $d \in R^n$ there exist $\bar{u} \in A(x)$ and $\bar{g} \in \partial_x f(x, \bar{u})$ such that

$$E'(x; d) = \langle \bar{g}, d\rangle = \max[\langle g, d\rangle : g \in \partial_x f(x, u), u \in A(x)].$$

Suppose $x + d \in X$ and $E'(x; d) \geqq 0$. Then, by the quasidifferentiability of $f(\cdot, \bar{u})$ at x, we have

$$f'_x(x, \bar{u}; d) = f^0_x(x, \bar{u}; d) = \max[\langle g, d\rangle : g \in \partial_x f(x, \bar{u})] \geqq \langle \bar{g}, d\rangle \geqq 0.$$

Thus, by the semiconvexity of $f(\cdot, \bar{u})$ at x,

$$f(x + d, \bar{u}) \geqq f(x, \bar{u}).$$

But $x + d \in X \subset B$ and assumption (d) imply

$$E(x + d) \geqq f(x + d, \bar{u})$$

and $\bar{u} \in A(x)$ implies

$$E(x) = f(x, \bar{u}),$$

so

$$E(x + d) \geqq f(x + d, \bar{u}) \geqq f(x, \bar{u}) = E(x)$$

and the semiconvexity of E at x is established. □

The following function F is an example of a semismooth function on R^2 which is not an extremal-valued function in the sense of Theorem 2, because in any ball about (0, 0) there is a point at which the value of F is neither the maximum nor the minimum of the three underlying linear functions that define F:

$$F(x_1, x_2) = \begin{cases} x_1 & \text{for } x_2 \geqq 0 \text{ and } x_2 \geqq x_1 \geqq 0, \\ x_2 & \text{for } x_1 \geqq 0 \text{ and } x_1 \geqq x_2 \geqq 0, \\ 0 & \text{for } x_1 \leqq 0 \text{ or } x_2 \leqq 0. \end{cases}$$

Note that $F(x_1, x_2) = \max[0, \min(x_1, x_2)]$. This raises the question of whether or not a finite extremal composition of extremal-valued functions is a semismooth function. This is indeed the case, as is shown in more generality in the next section.

4. Semismooth composition. In this section we show that a semismooth composition of semismooth functions results in a semismooth function. In order to prove this useful result we first establish a type of "chain rule" for generalized

gradient sets. For $v^1, v^2, \cdots, v^m \in R^n$ let $[v^1 v^2 \cdots v^m]$ denote that $n \times m$ matrix whose ith column is v^i for $i = 1, 2, \cdots, m$.

THEOREM 4. *Let $f_i: R^n \to R$ for $i = 1, 2, \cdots, m$ and $E: R^m \to R$ be locally Lipschitz. For $x \in R^n$ define*

$$Y(x) = (f_1(x), f_2(x), \cdots, f_m(x)),$$

$$F(x) = E(Y(x))$$

and

$$G(x) = \operatorname{conv}\{g \in R^n : g = [g^1 g^2 \cdots g^m]w,\ g^i \in \partial f_i(x),\ i = 1, 2, \cdots, m,\ w \in \partial E(Y(x))\}.$$

Then F is locally Lipschitz and

$$\partial F(x) \subset G(x) \quad \text{for each } x \in R^n. \tag{4.1}$$

Remarks. Clarke [2] establishes (4.1) for the three cases where 1) E is continuously differentiable and $m = 1$, 2) $E(y_1, y_2) = y_1 + y_2$ and 3) $E(y) = \max[y_i : i \in \{1, 2, \cdots, m\}]$ for $y = (y_1, y_2, \cdots, y_m)$.

Note that the containment in (4.1) may be strict, because, as suggested to us by M. J. D. Powell, for $E(y_1, y_2) = y_1 - y_2$, $x \in R$ and $f_1(x) = f_2(x) = |x|$, we have $\partial F(0) = \{0\}$ and $G(0) = \operatorname{conv}\{-2, 2\}$.

Proof. It is not difficult to show that F is locally Lipschitz and to show that G is uppersemicontinuous. Hence, by part (c) of Proposition 1, F is differentiable almost everywhere, and if we show

$$\nabla F(\bar{x}) \in G(\bar{x}) \tag{4.2}$$

where $\bar{x}$ is any point of differentiability of F, then (4.1) follows from the convexity and uppersemicontinuity of G.

In order to show (4.2), let $\nabla F(\bar{x})$ exist, $d \in R^n$ and $\{t_k\} \downarrow 0$. Then

$$\begin{aligned} \langle \nabla F(\bar{x}), d \rangle &= F'(\bar{x}; d) \\ &= \lim_{k \to \infty} [F(\bar{x} + t_k d) - F(\bar{x})]/t_k \\ &= \lim_{k \to \infty} [E(Y(\bar{x} + t_k d)) - E(Y(\bar{x}))]/t_k. \end{aligned} \tag{4.3}$$

Choose a subsequence of $\{t_k\}$ such that for each $i = 1, 2, \cdots, m$

$$\{[f_i(\bar{x} + t_k d) - f_i(\bar{x})]/t_k\} \to f_i^*$$

on the subsequence. By Lemma 1,

$$f_i^* = \langle g^i, d \rangle \quad \text{for some } g^i \in \partial f_i(\bar{x}), \tag{4.4}$$

so

$$\{[f_i(\bar{x} + t_k d) - f_i(\bar{x}) - t_k \langle g^i, d \rangle]/t_k\} \to 0$$

on the subsequence. Let

$$v = (f_1^*, f_2^*, \cdots, f_m^*) = (\langle d, g^1 \rangle, \langle d, g^2 \rangle, \cdots, \langle d, g^m \rangle). \tag{4.5}$$

Then

$$\{[Y(\bar{x}+t_k d)-Y(\bar{x})-t_k v]/t_k\}\to 0\in R^m$$

and, by Lipschitz continuity of E,

$$\{[E(Y(\bar{x}+t_k d))-E(Y(\bar{x})+t_k v)]/t_k\}\to 0 \tag{4.6}$$

on the subsequence. Now choose a sub-subsequence of $\{t_k\}$ such that

$$\{[E(Y(\bar{x})+t_k v)-E(Y(\bar{x}))]/t_k\}\to E^* \tag{4.7}$$

on this sub-subsequence. Then, by combining (4.6) and (4.7),

$$\{[E(Y(\bar{x}+t_k d))-E(Y(\bar{x}))]/t_k\}\to E^*$$

on the sub-subsequence and, by (4.3),

$$\langle \nabla F(\bar{x}), d\rangle = E^*. \tag{4.8}$$

From (4.7) and Lemma 1,

$$E^* = \langle v, w\rangle \quad \text{for some } w\in \partial E(Y(\bar{x})). \tag{4.9}$$

Let

$$g=[g^1 g^2\cdots g^m]w,$$

so that combining (4.8), (4.9), (4.5) and (4.4) and recalling the definition of G yields

$$\langle \nabla F(\bar{x}), d\rangle = E^* = \langle v, w\rangle = \langle(\langle d, g^1\rangle, \langle d, g^2\rangle, \cdots, \langle d, g^m\rangle), w\rangle = \langle d, g\rangle$$

where $g\in G(\bar{x})$. Since this result holds for each $d\in R^n$, and $G(\bar{x})$ is convex, we have that the desired result (4.2) holds, for if not, then a strict separation theorem [9, Thm. 3.2.6] gives a contradiction. □

THEOREM 5. *Suppose, in addition to the assumptions of Theorem* 4, *that f_i for each $i=1,2,\cdots,m$ is semismooth at $x\in R^n$ and E is semismooth at $Y(x)\in R^m$. Then F is semismooth at x.*

Proof. Suppose $x_k = x+t_k d+\theta_k$ and $g_k\in\partial F(x_k)$ where $d\in R^n$, $\{t_k\}\downarrow 0$ and $\{\theta_k/t_k\}\to 0\in R^n$. Since $\partial F(x_k)$ is contained in the compact convex set $G(x_k)$, by minimizing and maximizing the linear function $\langle\cdot, d\rangle$ over $G(x_k)$, we may find $\bar{g}_k$, $\hat{g}_k\in G(x_k)$ such that

$$\langle \bar{g}_k, d\rangle \leqq \langle g_k, d\rangle \leqq \langle \hat{g}_k, d\rangle$$

and

$$\bar{g}_k = [\bar{g}_k^1 \bar{g}_k^2\cdots \bar{g}_k^m]\bar{w}_k, \qquad \hat{g}_k = [\hat{g}_k^1, \hat{g}_k^2, \cdots, \hat{g}_k^m]\hat{w}_k$$

where

$$\bar{g}_k^i, \hat{g}_k^i\in\partial f_i(x_k) \quad \text{for each } i=1,2,\cdots,m$$

and

$$\bar{w}_k, \hat{w}_k\in\partial E(Y(x_k)).$$

By the uppersemicontinuity and local boundedness of the various maps, $\{\bar{g}_k\}$ and $\{\hat{g}_k\}$ are bounded and there are accumulation points $\bar{g}$ of $\{\bar{g}_k\}$ and $\hat{g}$ of $\{\hat{g}_k\}$ and corresponding accumulation points $\bar{g}^i$ of $\{\bar{g}_k^i\}$ and $\hat{g}^i$ of $\{\hat{g}_k^i\}$ for each $i = 1, 2, \cdots, m$ and $\bar{w}$ of $\{\bar{w}_k\}$ and $\hat{w}$ of $\{\hat{w}_k\}$ such that

$$\bar{g} = [\bar{g}^1 \bar{g}^2 \cdots \bar{g}^m]\bar{w}, \qquad \hat{g} = [\hat{g}^1 \hat{g}^2 \cdots \hat{g}^m]\hat{w}$$

and

$$\langle \bar{g}, d\rangle \leqq \liminf_{k\to\infty} \langle g_k, d\rangle \leqq \limsup_{k\to\infty} \langle g_k, d\rangle \leqq \langle \hat{g}, d\rangle.$$

By the semismoothness of each f_i, we have

$$\langle d, \bar{g}^i\rangle = \langle d, \hat{g}^i\rangle = f_i'(x; d);$$

so, by defining

$$z = (f_1'(x; d), f_2'(x; d), \cdots, f_m'(x; d)),$$

we have

$$\langle d, \bar{g}\rangle = \langle d, [\bar{g}^1 \bar{g}^2 \cdots \bar{g}^m]\bar{w}\rangle = \langle z, \bar{w}\rangle,$$

$$\langle d, \hat{g}\rangle = \langle d, [\hat{g}^1 \hat{g}^2 \cdots \hat{g}^m]\hat{w}\rangle = \langle z, \hat{w}\rangle,$$

and

$$\langle z, \bar{w}\rangle \leqq \liminf_{k\to\infty} \langle g_k, d\rangle \leqq \limsup_{k\to\infty} \langle g_k, d\rangle \leqq \langle z, \hat{w}\rangle.$$

So, if we show that

$$\langle z, \bar{w}\rangle = \langle z, \hat{w}\rangle, \tag{4.10}$$

then $\{\langle g_k, d\rangle\}$ has only one accumulation point and we are done.

To show (4.10) we will show that

$$Y(x_k) = Y(x) + t_k z + \phi_k \tag{4.11}$$

where

$$\{\phi_k / t_k\} \to 0 \in R^m, \tag{4.12}$$

and then, since $\bar{w}_k, \hat{w}_k \in \partial E(Y(x_k))$, we have, by the semismoothness of E, that $\{\langle \bar{w}_k, z\rangle\}$ and $\{\langle \hat{w}_k, z\rangle\}$ have the same limit, which implies (4.10), because $\bar{w}$ and $\hat{w}$ are accumulation points of $\{\bar{w}_k\}$ and $\{\hat{w}_k\}$, respectively.

For each $i = 1, 2, \cdots, m$ let

$$\phi_k^i = f_i(x_k) - f_i(x) - t_k f_i'(x; d),$$

so that (4.11) is satisfied with $\phi_k = (\phi_k^1, \phi_k^2, \cdots, \phi_k^m)$ and

$$\phi_k^i / t_k = [f_i(x_k) - f_i(x)]/t_k - f_i'(x; d). \tag{4.13}$$

Note that, by using the definition of x_k and adding and subtracting $f_i(x + t_k d)$, we have

$$\begin{aligned}[f_i(x_k) - f_i(x)]/t_k \\ = [f_i(x + t_k d + \theta_k) - f_i(x + t_k d)]/t_k + [f_i(x + t_k d) - f_i(x)]/t_k.\end{aligned} \tag{4.14}$$

As $k \to \infty$, the first term of the right-hand side of (4.14) converges to zero, because each f_i is Lipschitz and $\{\theta_k/t_k\} \to 0 \in R^n$. The second term converges to $f_i'(x; d)$, so we have that

$$\{[f_i(x_k) - f_i(x)]/t_k\} \to f_i'(x; d),$$

which, by (4.13), implies (4.12) and completes the proof. □

5. Stationarity and optimality. Consider the following problem that is equivalent to the optimization problem of § 1:

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad h(x) \leqq 0 \end{aligned}$$

where

$$h(x) = \max_{1 \leqq i \leqq m} h_i(x) \quad \text{for } x \in R^n.$$

We say that $x \in R^n$ is *feasible* if $h(x) \leqq 0$ and *strictly feasible* if $h(x) < 0$. We say that $\bar{x} \in R^n$ is *optimal* if $\bar{x}$ is feasible and $f(\bar{x}) \leqq f(x)$ for all feasible x.

Let X be a subset of R^n and for each $x \in R^n$ let

$$A(x) = \{i \in \{1, 2, \cdots, m\}: h(x) = h_i(x)\}.$$

Then, from Theorems 4,5,1 and 3, we have the following:

THEOREM 6. *Suppose $h_1, h_2, \cdots, h_m$ are locally Lipschitz.*

(a) *Then h is locally Lipschitz and for each $x \in R^n$*

$$\partial h(x) \subset \operatorname{conv}\{\partial h_i(x): i \in A(x)\}.$$

(b) *If $h_1, h_2, \cdots, h_m$ are semismooth on X then h is semismooth on X.*

(c) *If $h_1, h_2, \cdots, h_m$ are semiconvex (quasidifferentiable) on X then h is semiconvex (quasidifferentiable) on X and for each $x \in X^n$*

$$\partial h(x) = \operatorname{conv}\{\partial h_i(x): i \in A(x)\}.$$

A key idea for dealing with the above optimization problem is to define the point-to-set map $M: R^n \to 2^{R^n}$ by

$$M(x) = \begin{cases} \partial f(x) & \text{if } h(x) < 0 \\ \operatorname{conv}\{\partial f(x) \cup \partial h(x)\} & \text{if } h(x) = 0 \\ \partial h(x) & \text{if } h(x) > 0 \end{cases} \quad \text{for } x \in R^n.$$

This map was introduced and used by Merrill [10, Chap. 12] for problems with differentiable and/or convex functions, i.e. problems with functions having gradients and/or subgradients. It is used by our algorithm in [11] for problems with functions having generalized gradients.

We say that $\bar{x} \in R^n$ is *stationary* for the optimization problem if $h(\bar{x}) \leqq 0$ and $0 \in M(\bar{x})$. Our algorithm in [11] is shown to converge to stationary points for problems with semismooth functions. The next result shows that stationarity is necessary for optimality. It follows from a very general theorem in Clarke [2]. Here we give an independent proof using a strict separation theorem for convex sets.

THEOREM 7. *Suppose f and h are locally Lipschitz. If $\bar{x}$ is optimal then $\bar{x}$ is stationary.*

Proof. Consider the case where $h(\bar{x})=0$. Suppose, for contradiction purposes, that $\bar{x}$ is not stationary. Then $0 \notin M(\bar{x})$. Since $\partial f(\bar{x})$ and $\partial h(\bar{x})$ are compact, $M(\bar{x})$ is closed and convex and, thus, from a strict separation theorem [9, Cor. 3.2.4], there exists a $d \in R^n$ such that

$$\langle g, d\rangle < 0 \quad \text{for all } g \in M(\bar{x}). \tag{5.1}$$

Since $\bar{x}$ is optimal, it must be the case that either $f^0(\bar{x}; d) \geqq 0$ or $h^0(\bar{x}; d) \geqq 0$, for if not, we can find a $t>0$ such that $f(\bar{x}+td)<f(\bar{x})$ and $h(\bar{x}+td)<h(\bar{x})=0$, which contradicts the optimality of $\bar{x}$. Thus, by Proposition 1, there is a $\bar{g} \in (\partial f(\bar{x}) \cup \partial h(\bar{x})) \subset M(\bar{x})$ such that $\langle \bar{g}, d\rangle \geqq 0$. But this contradicts (5.1). So $0 \in M(\bar{x})$. We omit the proof of the case where $h(\bar{x})<0$ which is similar, but simpler. □

Remark. This theorem, when specialized, gives two well-known necessary optimality theorems. If $h_1, h_2, \cdots, h_m$ and f are differentiable then the above result combined with part (a) of Theorem 6 shows that an optimal $\bar{x}$ solves the Karush [5]–John [4] stationary point problem [9, p. 93]. Alternatively, if $h_1, h_2, \cdots, h_m$ and f are convex then Theorems 6 and 7 and Proposition 3 show that an optimal $\bar{x}$ solves the corresponding saddle-point problem [9, p. 71].

As usual, in order to have stationarity be sufficient for optimality, we need stronger assumptions on the problem functions. We now proceed to show that if the problem functions are semiconvex and there is a strictly feasible point then stationarity implies optimality. In order to demonstrate this we require the following preliminary result for semiconvex functions on convex sets:

THEOREM 8. *If F is semiconvex on a convex set $X \subset R^n$, $x \in X$ and $x+d \in X$ then*

$$F(x+d) \leqq F(x) \quad \textit{implies} \quad F'(x; d) \leqq 0.$$

Proof. Suppose, for contradiction purposes, $F(x+d) \leqq F(x)$ and $F'(x; d)>0$. Then there exists $t>0$ such that $t<1$ and $F(x+td)>F(x)$. Let $\bar{t} \in (0, 1)$ maximize the continuous function $a(t)=F(x+td)$ over $t \in [0, 1]$. Clearly, by the maximality of $\bar{t}$,

$$a(1)=F(x+d) \leqq F(x)=a(0)<a(\bar{t})=F(x+\bar{t}d), \tag{5.2}$$

$$F'(x+\bar{t}d; d) \leqq 0 \quad \text{and} \quad F'(x+\bar{t}d; -d) \leqq 0.$$

Now by the quasidifferentiability of F there exist $g^+ \in \partial F(x+\bar{t}d)$ and $g^- \in \partial F(x+\bar{t}d)$ such that

$$0 \geqq F'(x+\bar{t}d; d)=F^0(x+\bar{t}d; d)=\langle g^+, d\rangle \geqq \langle g^-, d\rangle$$

and

$$0 \geqq F'(x+\bar{t}d; -d)=F^0(x+\bar{t}d; -d)=\langle g^-, -d\rangle \geqq \langle g^+, -d\rangle.$$

So,

$$F'(x+\bar{t}d; d)=0,$$

and, by the positive homogeneity of $F'(x+\bar{t}d;\cdot)$, since $1-\bar{t}>0$, we have

$$F'(x+\bar{t}d;(1-\bar{t})d)=(1-\bar{t})F'(x+\bar{t}d;d)=0.$$

Then the semiconvexity of F implies

$$F(x+d)\geqq F(x+\bar{t}d)$$

which contradicts (5.2). □

Remark. The above proof follows one in Mangasarian [9, pp. 143–144] and a slight modification shows that a semiconvex function on a convex set is "strictly quasiconvex" and, hence, "quasiconvex" [9, Chap. 9].

THEOREM 9. *Suppose f and h are semiconvex on R^n and $\bar{x}\in R^n$ is such that $0\in M(\bar{x})$.*

(a) *If $h(\bar{x})>0$ then $h(x)\geqq h(\bar{x})>0$ for all $x\in R^n$, i.e. the optimization problem has no feasible points.*

(b) *If $h(\bar{x})\leqq 0$ then at least one of the following holds:*

(i) *$\bar{x}$ is optimal,*

(ii) *$h(x)\geqq 0$ for all $x\in R^n$, i.e. the optimization problem has no strictly feasible points.*

Proof. If $h(\bar{x})>0$ then $0\in\partial h(\bar{x})$ and it is clear from the semiconvexity of h that $\bar{x}$ minimizes h over R^n and the desired result (a) follows. If $h(\bar{x})<0$ then $0\in\partial f(\bar{x})$ and similar reasoning shows that $\bar{x}$ minimizes f over R^n which implies (b)(i). Suppose $h(\bar{x})=0$. Then there exist $\lambda\in[0,1]$, $\bar{g}\in\partial f(\bar{x})$ and $\hat{g}\in\partial h(\bar{x})$ such that

$$\lambda\bar{g}+(1-\lambda)\hat{g}=0.$$

If $\lambda=0$, then $\hat{g}=0$, $\bar{x}$ minimizes h over R^n and (b)(ii) holds. Alternatively, if $\lambda>0$ then

$$\bar{g}+[(1-\lambda)/\lambda]\hat{g}=0,$$

and for all $x\in R^n$

$$\langle\bar{g},x-\bar{x}\rangle+[(1-\lambda)/\lambda]\langle\hat{g},x-\bar{x}\rangle=0.$$

For all $x\in R^n$ such that $h(x)\leqq 0=h(\bar{x})$, we have, by the semiconvexity of h, Theorem 8 and the fact that $\hat{g}\in\partial h(\bar{x})$, that

$$0\geqq h'(\bar{x};x-\bar{x})=h^0(\bar{x},x-\bar{x})\geqq\langle\hat{g},x-\bar{x}\rangle.$$

Thus, since $[(1-\lambda)/\lambda]\geqq 0$, we have that

$$\langle\bar{g},x-\bar{x}\rangle\geqq 0\quad\text{for all } x \text{ such that } h(x)\leqq 0.$$

So, by the semiconvexity of f, since $\bar{g}\in\partial f(\bar{x})$, we have that

$$f'(\bar{x};x-\bar{x})=f^0(\bar{x};x-\bar{x})\geqq\langle\bar{g},x-\bar{x}\rangle\geqq 0$$

and, hence,

$$f(x)\geqq f(\bar{x})\quad\text{for all } x \text{ such that } h(x)\leqq 0.$$

Thus, $\bar{x}$ is optimal and we have that $\lambda>0$ implies that (b)(i) holds. □

Remark. If $h(\bar{x})=0$ and $\lambda>0$ in the above proof then, in order to show optimality of $\bar{x}$, we need only assume that h is quasidifferentiable and satisfies the conclusion of Theorem 8 rather than assume h is semiconvex. This observation corresponds to a sufficient optimality theorem in Mangasarian [9, Thm. 10.1.1] and says that if $\bar{x}$ satisfies generalized Karush [5]–Kuhn–Tucker [6] conditions, f is semiconvex and h is quasidifferentiable and "quasiconvex" [9, Chap. 9] then $\bar{x}$ is optimal. A constraint qualification that implies $\lambda>0$ is that $0\notin\partial h(\bar{x})$.

Acknowledgment. I wish to thank Claude Lemarechal for his many helpful suggestions.

REFERENCES

[1] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), 247–262.

[2] ———, *A new approach to Lagrange multipliers*, Math. of Operations Res., 1 (1976), pp. 165–174.

[3] A. FEUER, *An implementable mathematical programming algorithm for admissible fundamental functions*, Ph.D. dissertation, Dept. of Math., Columbia Univ., New York, 1974.

[4] F. JOHN, *Extremum problems with inequalities as subsidiary conditions*, Studies and Essays: Courant Anniversary Volume, K. O. Frederichs, O. E. Neugebauer and J. J. Stoker, eds., Interscience, New York, 1948, pp. 187–204.

[5] W. KARUSH, *Minima of functions of several variables with inequalities as side conditions*, M. S. dissertation, Dept. of Math., Univ. of Chicago, 1939.

[6] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., University of California Press, Berkeley, 1951, pp. 481–492.

[7] G. LEBOURG, *Valeur moyenne pour gradient généralisé*, C. R. Acad. Sc. Paris Sér. A, 281 (1975), pp. 795–797.

[8] C. LEMARECHAL, *An extension of Davidon methods to nondifferentiable problems*, Nondifferentiable Optimization. M. L. Balinski and P. Wolfe, eds. Mathematical Programming Study 3, North-Holland, Amsterdam, 1975, pp. 95–109.

[9] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[10] O. H. MERRILL, *Applications and extensions of an algorithm that computes fixed points of certain upper semicontinuous point to set mappings*, Ph.D. dissertation, Univ. of Michigan, Ann Arbor, 1972.

[11] R. MIFFLIN, *An algorithm for constrained optimization with semismooth functions*, International Institute for Applied Systems Analysis, Laxenburg, Austria, RR-77-3, Math. of Operations Res., to appear.

[12] B. N. PSHENICHNYI, *Necessary Conditions for an Extremum*, Marcel Dekker, New York, 1971.

[13] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[14] N. Z. SHOR, *A class of almost-differentiable functions and a minimization method for functions of this class*, Kibernetika (Kiev), 4 (1972), pp. 65–70 = Cybernetics, July (1974), pp. 599–606.

[15] HOÁNG TUY, *Sur les inégalites linéaires*, Colloq. Math., 13 (1964), pp. 107–123.

[16] P. WOLFE, *A method of conjugate subgradients for minimizing nondifferentiable functions*, Nondifferentiable Optimization, M. L. Balinski and P. Wolfe, eds., Mathematical Programming Study 3, North-Holland, Amsterdam 1975, pp. 145–173.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

BOUNDARY VALUE CONTROL OF A CLASS OF HYPERBOLIC EQUATIONS IN A GENERAL REGION*

JOHN LAGNESE†

Abstract. Let $c(t)$ be a real valued function which is analytic for $t \geqq 0$ and which is such that, for some positive integer $N \geqq 3$, the operator

$$L_N = \sum_{i=1}^{N} \frac{\partial^2}{\partial x_i^2} - \frac{\partial^2}{\partial t^2} - c(t)$$

satisfies Huygens' principle in the sense of Hadamard's "minor premise". Let Ω be a smooth, bounded domain in R^n, $n \geqq 2$. We show that control processes which are modeled by an equation $L_n u = 0$ in the cylindrical region $\Omega \times [0, \infty)$ are exactly controllable in any finite time T which exceeds the diameter of Ω by control forces applied on the wall of the cylinder.

1. Introduction. Let $n \geqq 2$ be a positive integer and Ω be a bounded, open, connected region in R^n with a smooth boundary Γ. We denote by $\nu(x)$ the outward unit normal vector at each $x \in \Gamma$. Let $c(t)$ be a real valued function which is analytic for $t \geqq 0$ and define

$$L_n u = \Delta_n u - \left(\frac{\partial^2 u}{\partial t^2} + c(t)u\right), \qquad \Delta_n = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}. \tag{1.1}$$

Let u_0 and v_0 be functions defined in Ω and α, β be constants with $\alpha^2 + \beta^2 > 0$, $\alpha\beta \geqq 0$, and let T be a specified positive number. We consider the following control problem: Does there exist a control function f defined on $\Gamma \times [0, T]$ such that the solution of the problem

$$L_n u = 0, \qquad (x, t) \in \Omega \times [0, T], \tag{1.2}$$

$$u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = v_0(x), \quad x \in \Omega, \tag{1.3}$$

$$\alpha u(x, t) + \beta \frac{\partial u}{\partial \nu}(x, t) = f(x, t), \qquad (x, t) \in \Gamma \times [0, T], \tag{1.4}$$

achieves a specified state

$$u(x, T) = u_1(x), \quad \frac{\partial u}{\partial t}(x, T) = v_1(x), \quad x \in \Omega. \tag{1.5}$$

We shall show that this question has an affirmative answer for a certain class of operators of the form (1.1) for initial and final states in $H^r(\Omega) \times H^{r-1}(\Omega)$, $r \geqq 2$, and moreover, (1.5) can be achieved in any time T which exceeds the diameter of Ω.

* Received by the editors September 2, 1976.
† Department of Mathematics, Georgetown University, Washington, D.C. 20057.

The particular class of operators L_n we shall consider consists of those for which L_N satisfies Huygens' principle in the sense of Hadamard's "minor premise" [6] for some N. Included in this class are the ordinary wave operators

$$\Box_n = \Delta_n - \frac{\partial^2}{\partial t^2},$$

the EPD (Euler–Poisson–Darboux) operators in self-adjoint form

$$\Box_n + \frac{k(k+1)}{(at+b)^2}, \quad ab > 0, \quad k = 1, 2, \cdots,$$

and many others. For additional examples and detailed information about the structure of such operators we refer to [8]–[10].

Boundary control problems for hyperbolic equations have been considered by a number of authors, most notably D. L. Russell; see [11]–[14] and references contained therein; also [2], [3], [5]. The results of this paper are most closely related to those of [13]. In that paper it was proved, for processes governed by the wave equation, that the set of controllable states includes $H^2(\Omega) \times H^1(\Omega)$ and that, if n is an odd positive integer, T can be taken as any number greater than the diameter of Ω. The proof makes essential use of the fact that Huygens' principle is valid for such operators. That T also can be taken in this way for the wave equation in even space dimensions has, as far as we know, been proved only when the geometry of Ω is of certain specific types, e.g., when Ω is a unit ball (see [5]).

The result we shall prove is the following

THEOREM. *Let the initial state (u_0, v_0) and final state (u_1, v_1) be given in $H^r(\Omega) \times H^{r-1}(\Omega)$, $r \geqq 2$. Suppose $c(t)$ is analytic for $t \geqq 0$ and that L_N is a Huygens' operator for some N. Let T be any number greater than the diameter of Ω. Then there exists a control function $f \in H^s(\Gamma \times [0, T])$ ($s = r - \frac{3}{2}$ if $\beta \neq 0$, $s = r - \frac{1}{2}$ if $\beta = 0$) such that the unique solution of* (1.2)–(1.4) *lies in $H^r(\Omega \times [0, T])$ and satisfies* (1.5). *Moreover, there is a constant $K = K(r, T)$ such that*

$$\|f\|_{H^s(\Gamma \times [0,T])} \leqq K(\|u_0\|_{H^r(\Omega)} + \|u_1\|_{H^{r-1}(\Omega)} + \|v_0\|_{H^{r-1}(\Omega)} + \|v_1\|_{H^{r-1}(\Omega)}). \tag{1.6}$$

We shall establish this result for the wave equation in even space dimensions in the next section. The proof in the general case, which is considered in § 3, is then obtained by exploiting the fact that all operators of the specific type considered here are transforms, in a certain sense, of the wave operator. This fact also allows the determination of a boundary control for the operators L_n in terms of a certain boundary control for the wave operator.

2. The wave equation in even space dimensions. Because of the time reversibility of the wave equation, it suffices to consider the case $u_1 = v_1 \equiv 0$.

Our proof is an extension of a method, used by Russell in [13], which may be summarized as follows:

Set

$$V^r(\Omega) = H^r(\Omega) \times H^{r-1}(\Omega), \qquad V_0^r(\Omega) = H_0^r(\Omega) \times H_0^{r-1}(\Omega).$$

Let $\delta > 0$ and set

$$\Omega_\delta = \{x \in R^n : \|x - \hat{x}\| < \delta \text{ for some } \hat{x} \in \Omega\}.$$

Russell first extends (u_0, v_0) to a pair $(u_\delta, v_\delta) \in V_0^r(\Omega_\delta)$ by a bounded linear transformation. Setting these functions equal to zero outside of Ω_δ, one then solves

$$\Delta_n u - \frac{\partial^2 u}{\partial t^2} = 0 \tag{2.1}$$

in $R^n \times [0, \infty)$ using (u_δ, v_δ) for initial data. The solution $w_\delta(x, t)$ so obtained is in $C^\infty(\bar{\Omega}_\delta \times [T_0, \infty))$, where T_0 is any fixed number which satisfies

$$T_0 > 2\delta + \text{diameter } (\Omega).$$

Let $\phi \in C_0^\infty(\Omega_\delta)$ such that $\phi(x) = 1$ if $x \in \Omega$. Let $T \geqq T_0$ and define $z(x, t)$ to be the solution of (2.1) in $R^n \times (-\infty, T]$ which satisfies the terminal conditions

$$z(x, T) = \phi(x) w_\delta(x, T), \qquad \frac{\partial z}{\partial t}(x, T) = \phi(x) \frac{\partial w_\delta}{\partial t}(x, T).$$

Let $(\hat{u}_0, \hat{v}_0)$ be the restriction of $(z(x, 0), (\partial z/\partial t)(x, 0))$ to Ω. $(\hat{u}_0, \hat{v}_0)$ depends, of course, on (u_0, v_0). Setting

$$u(x, t) = w_\delta(x, t) - z(x, t)$$

and defining f to be the restriction of $\alpha u + \beta(\partial u/\partial \nu)$ to $\Gamma \times [0, T]$, one obtains a solution of (1.2), (1.4) whose initial data is $(u_0 - \hat{u}_0, v_0 - \hat{v}_0)$ and whose final state is zero at time $t = T$. The question is then whether $(u_0 - \hat{u}_0, v_0 - \hat{v}_0)$ spans all of $V^r(\Omega)$ as (u_0, v_0) does the same.

To answer this question, Russell considers the map $K_T: (u_0, v_0) \to (\hat{u}_0, \hat{v}_0)$ and proves that K_T is a linear contraction on $V^r(\Omega)$ for all sufficiently large T and hence $(I - K_T)^{-1}$ exists as an everywhere defined bounded linear operator. Thus the control problem is solved for such values of T.

As Russell notes in his paper, K_T is compact for each $T \geqq T_0$. Let Σ_0 be the region of the complex plane given by

$$\Sigma_0 = \left\{ \zeta : \zeta = T_0 + z, |\arg z| \leqq \frac{\pi}{4} \right\}.$$

Our extension of Russell's proof consists of showing that the family $\{K_T : T \geqq T_0\}$ can be extended to a family $\{K_\zeta : \zeta \in \Sigma_0\}$ of compact operators which depend holomorphically on ζ. One can then utilize a result of Atkinson [1] (see also [7, p. 370]) to the effect that either 1 is an eigenvalue of each of the operators K_ζ, $\zeta \in \Sigma_0$, or else $(I - K_\zeta)^{-1}$ exists for all but at most a finite number of values of ζ in each compact subset of Σ_0. This latter possibility must be the case since K_ζ is a contraction for sufficiently large positive values of ζ. Thus for all $T \geqq T_0$ with the possible exception of a finite number of values, $(I - K_T)^{-1}$ exists. The conclusions of the theorem now follow for all such values as in [13]. One may then conclude that the control problem is solvable for every $\tilde{T} > T_0$. In fact, for each such $\tilde{T}$ one may choose a number T between T_0 and $\tilde{T}$ for which the problem (2.1), (1.3), (1.4) has a solution $u \in H^r(\Omega \times [0, T])$ satisfying

$$u(x, T) = \frac{\partial u}{\partial t}(x, T) = 0, \qquad x \in \Omega, \tag{2.2}$$

this being true for every choice of $(u_0, v_0) \in V^r(\Omega)$. Extend u to a function $\tilde{u}$ defined on $\Omega \times [0, \tilde{T}]$ by setting $\tilde{u} = 0$ in $\Omega \times [T, \tilde{T}]$. If $\tilde{u} \in H^r(\Omega \times [0, \tilde{T}])$, the trace $\tilde{f}$ of $\alpha\tilde{u} + \beta(\partial\tilde{u}/\partial\nu)$ on $\Gamma \times [0, \tilde{T}]$ is a boundary control of the proper type which steers (u_0, v_0) to zero in time $t = \tilde{T}$.

To verify that $\tilde{u} \in H^r(\Omega \times [0, \tilde{T}])$ we note that $(\partial^k u/\partial t^k)(\cdot, t) \in H^{r-k}(\Omega)$ for each $t \in [0, T]$, $k = 0, 1, \cdots, r$. This fact follows from standard energy estimates (cf. [4, p. 652]). One may then use (2.1) and (2.2) to conclude that

$$\frac{\partial^k u}{\partial t^k}(x, T) = 0, \qquad x \in \Omega, \quad k = 0, 1, \cdots, r,$$

which implies that $\tilde{u} \in H^r(\Omega \times [(0, \tilde{T}])$.

To obtain the holomorphic extension of the family $\{K_T\colon T \geqq T_0\}$ we examine these operators in detail.

Let $(u_\delta, v_\delta) \in V_0^r(\Omega_\delta)$. For $t \geqq T_0$ and $x \in \bar{\Omega}_\delta$ define

$$w_\delta(x, t) = \frac{2}{1 \cdot 3 \cdots (n-1)\sigma_n}\left\{\frac{\partial}{\partial t}\left(\frac{1}{t\partial t}\right)^{(n-2)/2} \int_{\Omega_\delta} \frac{u_\delta(\eta)\, d\eta}{(t^2 - \|\eta - x\|^2)^{1/2}} + \left(\frac{1}{t\partial t}\right)^{(n-2)/2} \int_{\Omega_\delta} \frac{v_\delta(\eta)\, d\eta}{(t^2 - \|\eta - x\|^2)^{1/2}}\right\}$$

where σ_n is the surface area of the unit ball in R^{n+1}. w_δ is a solution of (2.1) on $R^n \times [T_0, \infty)$. Let $\lambda > 0$ satisfy

$$\frac{\operatorname{diam} \Omega + 2\delta}{T_0} < \lambda < 1.$$

For $t \geqq T_0$, $\eta \in \bar{\Omega}_\delta$ and $x \in \bar{\Omega}_\delta$ one has $\|\eta - x\| \leqq \lambda t$. It follows that $w_\delta \in C^\infty(\bar{\Omega}_\delta \times [T_0, \infty))$ and all differentiations may be carried out beneath the integral. There results, for example, that

$$w_\delta(x, t) = \frac{2(-1)^{n/2+1}}{(n-1)\sigma_n}\left[(1-n)t \int_{\Omega_\delta} \frac{u_\delta(\eta)\, d\eta}{(t^2 - \|\eta - x\|^2)^{(n+1)/2}} + \int_{\Omega_\delta} \frac{v_\delta(\eta)\, d\eta}{(t^2 - \|\eta - x\|^2)^{(n-1)/2}}\right], \tag{2.3}$$

$$\frac{\partial w_\delta}{\partial t}(x, t) = \frac{2(-1)^{n/2+1}}{\sigma_n}\left[(n+1)t^2 \int_{\Omega_\delta} \frac{u_\delta(\eta)\, d\eta}{(t^2 - \|\eta - x\|^2)^{(n+3)/2}} - \int_{\Omega_\delta} \frac{u_\delta(\eta) + t v_\delta(\eta)}{(t^2 - \|\eta - x\|^2)^{(n+1)/2}}\, d\eta\right] \tag{2.4}$$

as long as $t \geqq T_0$ and $x \in \bar{\Omega}_\delta$. If ρ is a positive integer, one has the following estimate (see [13]), valid for $t \geqq T_0$:

$$\|w_\delta(\cdot, t)\|^2_{H^\rho(\Omega_\delta)} + \left\|\frac{\partial w_\delta}{\partial t}(\cdot, t)\right\|^2_{H^{\rho-1}(\Omega_\delta)} \leqq K(r, \rho, t)(\|u_\delta\|^2_{H^r(\Omega_\delta)} + \|v_\delta\|^2_{H^{r-1}(\Omega_\delta)}). \tag{2.5}$$

Equations (2.3) and (2.4) define a linear mapping $S(t)$: $(u_\delta, v_\delta) \to (w_\delta(\cdot, t), (\partial w_\delta/\partial t)(\cdot, t))$ of $V_0^r(\Omega_\delta)$ with $V^r(\Omega_\delta)$ which, in view of (2.5), is compact for each $t \geqq T_0$.

Let $T \geqq T_0$ and $(\hat{u}_1, \hat{v}_1) \in V_0^r(\Omega_\delta)$. Extend these functions to R^n by setting each equal to zero outside Ω_δ, and let $z(x, t)$ be the solution of (2.1) in $R^n \times (-\infty, T]$ which assumes the terminal data

$$z(x, T) = \hat{u}_1(x), \quad \frac{\partial z}{\partial t}(x, T) = \hat{v}_1(x), \quad x \in R^n.$$

If $t \leqq T - T_0$ and $x \in \bar{\Omega}_\delta$ the solution z has a form similar to (2.3). In particular,

$$z(x, 0) = \frac{2(-1)^{n/2+1}}{(n-1)\sigma_n}\left[(1-n)T\int_{\Omega_\delta} \frac{\hat{u}_1(\eta)\, d\eta}{(T^2 - \|\eta - x\|^2)^{(n+1)/2}} - \int_{\Omega_\delta} \frac{\hat{v}_1(\eta)\, d\eta}{(T^2 - \|\eta - x\|^2)^{(n-1)/2}}\right], \tag{2.6}$$

$$\frac{\partial z}{\partial t}(x, 0) = \frac{2(-1)^{n/2+1}}{\sigma_n}\left[-(n+1)T^2 \int_{\Omega_\delta} \frac{\hat{u}_1(\eta)\, d\eta}{(T^2 - \|\eta - x\|^2)^{(n+3)/2}} + \int_{\Omega_\delta} \frac{\hat{u}_1(\eta) - T\hat{v}_1(\eta)}{(T^2 - \|\eta - x\|^2)^{(n+1)/2}}\, d\eta\right] \tag{2.7}$$

if $x \in \bar{\Omega}_\delta$. Equations (2.6) and (2.7) define a compact mapping $\hat{S}(T)$: $(\hat{u}_1, \hat{v}_1) \to (z(\cdot, 0), (\partial z/\partial t)(\cdot, 0))$ of $V_0^r(\Omega_\delta)$ into $V^r(\Omega_\delta)$ for each $T \geqq T_0$. $\hat{S}(T)$ is related to $S(T)$ as follows: Let P_i be the projection of $V^r(\Omega_\delta)$ onto $H^{r-i}(\Omega_\delta)$, $i = 0, 1$. Then

$$P_0\hat{S}(T)(\hat{u}_1, \hat{v}_1) = P_0 S(T)(\hat{u}_1, -\hat{v}_1), \tag{2.8}$$

$$P_1\hat{S}(T)(\hat{u}_1, \hat{v}_1) = P_1 S(T)(-\hat{u}_1, \hat{v}_1). \tag{2.9}$$

The mapping K_T of [13] may now be expressed as follows. Let E be a bounded linear operator from $V^r(\Omega)$ into $V_0^r(\Omega_\delta)$ such that restriction of $E(u_0, v_0)$ to Ω coincides with (u_0, v_0). Let $\phi \in C_0^\infty(\Omega_\delta)$ such that $\phi(x) \equiv 1$ on Ω and define a bounded linear mapping M_ϕ: $V^r(\Omega_\delta) \to V_0^r(\Omega_\delta)$ by $M_\phi(u, v) = (\phi u, \phi v)$. Let R: $V^r(\Omega_\delta) \to V^r(\Omega)$ be the bounded linear operator defined by $R(u, v) = (u|_\Omega, v|_\Omega)$. Then

$$K_T = R\hat{S}(T)M_\phi S(T)E, \qquad T \geqq T_0.$$

To obtain a holomorphic extension of $\{K_T : T \geqq T_0\}$ it clearly suffices to do the same for $\hat{S}(T)$ and $S(T)$. In view of (2.8) and (2.9), it is enough to show that $P_i S(T)$: $V_0^r(\Omega_\delta) \to H^{r-i}(\Omega_\delta)$ has such a holomorphic extension, $i = 0, 1$.

Let $\zeta \in \Sigma_0$. Then for all $\eta \in \bar{\Omega}_\delta$ and all $x \in \bar{\Omega}_\delta$

$$\begin{aligned}\operatorname{Re}(\zeta^2 - \|\eta - x\|^2) &= T_0^2 - \|\eta - x\|^2 + 2T_0 \operatorname{Re}(z) + (\operatorname{Re}(z))^2 - (\operatorname{Im}(z))^2 \\ &\geqq (1 - \lambda^2)T_0^2\end{aligned}$$

and $|\arg(\zeta^2 - \|\eta - x\|^2)| \leqq \pi/2$. By choosing that value of $(\zeta^2 - \|\eta - x\|^2)^{1/2}$ which has positive real part, one obtains, for fixed η and x, a holomorphic function of ζ

whose values lie in the sector

$$\Sigma_\lambda = \left\{\zeta\colon \zeta = \sqrt{1-\lambda^2}\,T_0 + z,\ |\arg z| \leqq \frac{\pi}{4}\right\}$$

for all $\eta \in \bar{\Omega}_\delta$ and $x \in \bar{\Omega}_\delta$. The functions $F_k(\zeta) = (\zeta^2 - \|\eta - x\|^2)^{k/2}$ are likewise holomorphic in Σ_0 and satisfy

$$\|F_k(\zeta)\| \geqq (1-\lambda^2)^{k/2} T_0^k, \qquad \zeta \in \Sigma_0. \tag{2.10}$$

The operators $P_iS(T)$ are extended to operators $P_iS(\zeta)$ using (2.3) and (2.4), in which t is replaced by ζ. Inequality (2.10) implies that the quantities $w_\delta(x, \zeta)$, $(\partial w/\partial t)(x, \zeta)$ are holomorphic in Σ_0 for each $x \in \bar{\Omega}_\delta$ and differentiations may be carried out under the integral. Just as in [13], one can obtain an estimate like (2.5) with t replaced by $\zeta \in \Sigma_0$, from which follows that each $P_iS(\zeta)$ is compact. To show, for example, that $P_0S(\zeta)$ is holomorphic in Σ_0 it suffices to prove its weak holomorphicity, that is, to prove

$$F(\zeta) = \sum_{|\alpha| \leqq r} \int_{\Omega_\delta} D_x^\alpha w_\delta(x, \zeta) \overline{D_x^\alpha v(x)}\, dx$$

is holomorphic in Σ_0 for each (u_δ, v_δ) in $V_0^r(\Omega_\delta)$ and each $v \in H^r(\Omega_\delta)$, where

$$D_x^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha} \cdots \partial x_n^{\alpha_n}}, \qquad |\alpha| = \alpha_1 + \cdots + \alpha_n.$$

For any nonzero real number m, one computes

$$D_x^\alpha[(\zeta^2 - \|\eta - x\|^2)^{-m}] = \sum_{\beta = [(\alpha-1)/2]}^{\alpha} C_\beta \frac{(x-\eta)^{q_\beta}}{(\zeta^2 - \|\eta - x\|^2)^{m+|\beta|}}$$

where the C_β are constants,

$$\left[\frac{\alpha - 1}{2}\right] = \left(\left[\frac{\alpha_1 - 1}{2}\right], \cdots, \left[\frac{\alpha_n - 1}{2}\right]\right),$$

and $q_\beta = (q_{\beta_1}, \cdots, q_{\beta_n})$ satisfies $|q_\alpha| = |\alpha|$ and $|q_\beta| < |\beta|$ if $\beta \neq \alpha$. Thus $F(\zeta)$ is a linear combination of integrals of the form

$$\int_{\Omega_\delta} A(x) \int_{\Omega_\delta} \frac{B(\eta)(\eta - x)^\gamma}{(\zeta^2 - \|\eta - x\|^2)^{k/2}}\, d\eta\, dx$$

and such integrals multiplied by ζ, where k is an odd positive integer and A and B are (at least) in $L^2(\Omega_\delta)$. Each such integral is a holomorphic function in Σ_0 because of (2.10) and, therefore, $P_0S(\zeta)$ is holomorphic there.

3. The general case. To prove the theorem in the general case, we exploit the fact that every analytic operator of the form (1.1) such that L_N is a Huygens' operator for some N is a certain transformation of the ordinary wave operator $\Box_n$. To describe this transformation, consider (1.1) with $c(t)$ analytic for $t \geqq 0$. Let $\mu(t)$ be a solution of

$$\ddot{\mu} + c(t)\mu = 0, \qquad \left(\dot{\ } = \frac{d}{dt}\right), \tag{3.1}$$

which is analytic and nonzero for $t \geqq 0$ and define the operators

$$l = \frac{\partial}{\partial t} - \frac{\dot{\mu}}{\mu}, \qquad l^* = -\frac{\partial}{\partial t} - \frac{\dot{\mu}}{\mu}. \tag{3.2}$$

Then one may write

$$L_n = \Delta_n + l^* l$$

and therefore

$$lL_n = \tilde{L}_n l \tag{3.3}$$

where

$$\tilde{L}_n = \Delta_n + ll^* = \Delta_n - \frac{\partial^2}{\partial t^2} + \left[c(t) + 2\left(\frac{\dot{\mu}}{\mu}\right)^2 \right].$$

One of the important properties of this mapping of L_n to $\tilde{L}_n$ is the fact that if L_N is a Huygens' operator, the same is true of $\tilde{L}_{N+2}$ [10]. Starting with the wave operator $\Box_n$ and repeatedly applying such transformations results in

$$l_q l_{q-1} \cdots l_0 \Box_n = L_n l_q l_{q-1} \cdots l_0 \tag{3.4}$$

where

$$l_k = \frac{\partial}{\partial t} - \frac{\dot{\mu}_k}{\mu_k}, \qquad \mu_0 = at + b \quad (ab > 0), \tag{3.5}$$

$$(l_k l_k^*)\mu_{k+1} = 0, \qquad k = 0, 1, \cdots, q-1. \tag{3.6}$$

It was proved in [8] that *every* operator L_n which satisfies the conditions of the theorem satisfies (3.4) for some (nonunique) sequence $\{l_k\}$. It is possible to choose $q = (N-5)/2$ in (3.4), but not smaller. In this case the sequence $\{l_k\colon k = 0, 1, \cdots, (N-5)/2\}$ is uniquely determined. In addition, one has

$$c(t) = 2\frac{d^2}{dt^2}\left[\log \prod_{k=0}^{(N-5)/2} |\mu_k(t)| \right]$$

and $\prod_{k=0}^{m} \mu_k(t)$ is a polynomial of degree $\frac{1}{2}(m+1)(m+2)$, $0 \leqq m \leqq (N-5)/2$. These and other facts concerning the operators L_n are proved in [9].

Since the theorem is proved for $\Box_n$, to prove it in the general case it is sufficient to show that if it is true for some operator L_n it is also true for the operator $\tilde{L}_n$ defined by (3.1)–(3.3).

Let $T \geqq T_0$, $u_0 \in H^{r+1}(\Omega)$, $v_0 \in H^r(\Omega)$ and $f \in H^{s+1}(\Gamma \times [0, T])$ be chosen so that the unique solution $u \in H^{r+1}(\Omega \times [0, T])$ of (1.2)–(1.4) satisfies (1.5). Set $\tilde{u} = lu$ and note that

$$\frac{\partial \tilde{u}}{\partial t} = u_{tt} - \left(\frac{\ddot{\mu}}{\mu}\right)u + \left(\frac{\dot{\mu}}{\mu}\right)^2 u - \left(\frac{\dot{\mu}}{\mu}\right)u_t = \left[\Delta_n + \left(\frac{\dot{\mu}}{\mu}\right)^2\right]u - \left(\frac{\dot{\mu}}{\mu}\right)u_t.$$

Therefore $\tilde{u} \in H^r(\Omega \times [0, T])$ satisfies

$$\tilde{L}_n \tilde{u} = 0 \quad \text{in } \Omega \times [0, T],$$

$$\tilde{u}(x, 0) = \tilde{u}_0(x) \equiv v_0(x) - \frac{\dot{\mu}(0)}{\mu(0)} u_0(x), \qquad x \in \Omega, \tag{3.7}$$

$$\frac{\partial \tilde{u}}{\partial t}(x, 0) = \tilde{v}_0(x) \equiv \left[\Delta_n + \left(\frac{\dot{\mu}(0)}{\mu(0)}\right)^2\right] u_0(x) - \frac{\dot{\mu}(0)}{\mu(0)} v_0(x), \qquad x \in \Omega, \tag{3.8}$$

$$\alpha \tilde{u}(x, t) + \beta \frac{\partial \tilde{u}}{\partial \nu}(x, t) = lf(x, t) \quad \text{on } \Gamma \times [0, T].$$

In addition, at $t = T$, one has

$$\tilde{u}(x, T) = \tilde{u}_1(x), \quad \frac{\partial \tilde{u}}{\partial t}(x, T) = \tilde{v}_1(x), \quad x \in \Omega,$$

where $\tilde{u}_1$, $\tilde{v}_1$ are defined in the same manner as $\tilde{u}_0$, $\tilde{v}_0$ but with u_0, v_0 replaced by u_1, v_1, respectively. Thus the control function $lf \in H^s(\Gamma \times [0, T])$ steers $(\tilde{u}_0, \tilde{v}_0)$ to $(\tilde{u}_1, \tilde{v}_1)$ in time $t = T$. In addition, the map $(u_0, v_0) \to (\tilde{u}_0, \tilde{v}_0)$ defined by (3.7), (3.8) maps $V^{r+1}(\Omega)$ *onto* $V^r(\Omega)$. Indeed, given $(\tilde{u}_0, \tilde{v}_0)$, one may choose u_0 as the unique solution in $H^{r+1}(\Omega)$ of

$$\Delta_n u_0 = \tilde{v}_0 + \frac{\dot{\mu}(0)}{\mu(0)} \tilde{u}_0, \qquad x \in \Omega, \tag{3.9}$$

$$u_0 = 0 \quad \text{on } \Gamma, \tag{3.10}$$

and then set

$$v_0 = \tilde{u}_0 + \frac{\dot{\mu}(0)}{\mu(0)} u_0, \qquad x \in \Omega. \tag{3.11}$$

It follows that an arbitrary initial state $(\tilde{u}_0, \tilde{v}_0) \in V^r(\Omega)$ can be steered in time $t = T$ to an arbitrary final state $(\tilde{u}_1, \tilde{v}_1) \in V^r(\Omega)$ with a boundary control $lf \in H^s(\Gamma \times [0, T])$.

To obtain the estimate (1.6), we define u_0, v_0 by (3.9)–(3.11) and similarly define u_1, v_1 in terms of $\tilde{u}_1$, $\tilde{v}_1$. Using the well known estimate

$$\|w\|_{k+2} \leqq c \|\Delta_n w\|_k, \quad w \in H^{r+1}(\Omega) \cap H_0^1(\Omega), \quad k \leqq r - 1,$$

one obtains

$$\begin{aligned} \|lf\|^2_{H^s(\Gamma \times [0,T])} &\leqq C \|f\|^2_{H^{s+1}(\Gamma \times [0,T])} \\ &\leqq K(r, T)(\|u_0\|^2_{H^{r+1}(\Omega)} + \|u_1\|^2_{H^{r+1}(\Omega)} + \|v_0\|^2_{H^r(\Omega)} + \|v_1\|^2_{H^r(\Omega)}) \\ &\leqq \tilde{K}(r, T)(\|\tilde{u}_0\|^2_{H^r(\Omega)} + \|\tilde{u}_1\|^2_{H^r(\Omega)} + \|\tilde{v}_0\|^2_{H^{r-1}(\Omega)} + \|\tilde{v}_1\|^2_{H^{r-1}(\Omega)}). \end{aligned}$$

Remark 1. We have shown that if the control problem for the wave equation is solvable in time T, the same is true for all Huygens' operators (1.1) or, equivalently, for operators L_n defined by (3.4)–(3.6). Conversely, if for some Huygens' operator (1.1) the control problem is solvable in time T, the same is true

for the wave operator. This follows from the relation

$$l_0 l_1 \cdots l_q L_n = \Box_n l_0 l_1 \cdots l_q$$

(see [8]) where the $\{\mu_k\}$ are certain solutions of

$$\ddot{\mu}_q + c(t)\mu_q = 0, \quad (l_k l_k^*)\mu_{k-1} = 0, \quad k = q, q-1, \cdots, 1,$$

and $q = (N-5)/2$ where N is the smallest number of space dimensions for which L_N satisfies Huygens' principle. Russell has proved [11] that if $\hat{T}$ is less than the "minimal distance across Ω", that is,

$$\hat{T} < 2 \min_{t>0} \{t: \Gamma_t \supseteq \Omega\}$$

where

$$\Gamma_t = \{x \in R^n: \exists\, \hat{x} \in \Gamma, \|x - \hat{x}\| \leqq t\},$$

the control problem for the wave equation is not solvable in time $\hat{T}$ and, in fact, the set of initial states which can be steered to zero at time $\hat{T}$ with controls (1.4) is not even dense in $V^r(\Omega)$. The same result must therefore hold for every analytic Huygens' operator (1.1).

Remark 2. Just as in the case of the wave equation, the control function f for which the solution of (1.2)–(1.4) satisfies (2.2) may be realized as the solution of a certain moment problem. For definiteness we assume $\alpha \neq 0$, although this is inessential. Following [13], let

$$0 < \lambda_1 < \lambda_2 < \cdots < \lambda_k < \cdots \to \infty$$

be the eigenvalues of the problem

$$\Delta_n w + \lambda w = 0 \quad \text{in } \Omega$$

subject to the boundary condition

$$\alpha w(x) + \beta \frac{\partial w}{\partial \nu}(x) = 0 \quad \text{on } \Gamma,$$

and let $\{\phi_{k,l}: k = 1, 2, \cdots; l = 1, 2, \cdots, m_k\}$ be an orthonormal basis for $L^2(\Omega)$ such that $\{\phi_{k,l}: l = 1, 2, \cdots, m_k\}$ spans the eigenspace corresponding to λ_k. Write

$$u_0(x) = \sum_{k=1}^{\infty} \sum_{l=1}^{m_k} \mu_{k,l} \phi_{k,l},$$

$$v_0(x) = \sum_{k=1}^{\infty} \sum_{l=1}^{m_k} \nu_{k,l} \phi_{k,l}.$$

These series converge in $H^r(\Omega)$ and $H^{r-1}(\Omega)$, respectively.

Let u be the solution of the boundary control problem (1.2)–(1.4), (2.2) and $z \in H^r(\Omega \times [0, T])$ be a solution of (1.2) in $\Omega \times [0, T]$ which satisfies homogeneous boundary conditions of the type (1.4). One can establish, exactly as in [13], the relation

$$(3.12) \quad \int_\Omega \left[z(x, 0) v_0(x) - u_0(x) \frac{\partial z}{\partial t}(x, 0) \right] dx = \frac{1}{\alpha} \int_{\Gamma \times [0,T]} \frac{\partial z}{\partial \nu}(x, t) f(x, t)\, dx\, dt.$$

To obtain the moment problem, one inserts into this relation special solutions $z(x, t)$. Since $\phi_{i,j}(x)\cos(\omega_i t)$ and $\phi_{i,j}(x)\sin(\omega_i t)$ $(\omega_i = \sqrt{\lambda_i})$ are solutions of $\Box_n u = 0$ which satisfy homogeneous boundary conditions, (3.4) suggests that one introduce solutions of the form

$$z_{i,j} = \phi_{i,j}(x)(l_q l_{q-1} \cdots l_0)\cos(\omega_i t),$$
$$\hat{z}_{i,j} = \phi_{i,j}(x)(l_q l_{q-1} \cdots l_0)\sin(\omega_i t)$$

for $i = 1, 2, \cdots; j = 1, 2, \cdots, m_k$, where $q = (N-5)/2$.

Substituting these solutions into (3.12) and using the orthonormality of the $\phi_{k,l}$ results in the moment problem

$$\tilde{\nu}_{i,j} = \frac{1}{\alpha}\int_{\Gamma\times[0,T]} \frac{\partial\phi_{i,j}}{\partial\nu}(x)[(l_q l_{q-1}\cdots l_0)\cos(\omega_i t)]f(x,t)\,dx\,dt, \tag{3.13}$$

$$\tilde{\mu}_{i,j} = \frac{1}{\alpha}\int_{\Gamma\times[0,T]} \frac{\partial\phi_{i,j}}{\partial\nu}(x)[(l_q l_{q-1}\cdots l_0)\sin(\omega_i t)]f(x,t)\,dx\,dt, \tag{3.14}$$

where $\tilde{\nu}_{i,j}$, $\tilde{\mu}_{i,j}$ are functions of $\mu_{i,j}$, $\nu_{i,j}$ and ω_i.

As a specific example we consider

$$\Delta_n u - \frac{\partial^2 u}{\partial t^2} + \frac{2}{(t+1)^2}u = 0.$$

In this case $q = 0$,

$$l_0 = \frac{\partial}{\partial t} - \frac{1}{t+1},$$
$$z_{i,j} = -\left[\omega_i \sin(\omega_i t) + \frac{1}{t+1}\cos(\omega_i t)\right]\phi_{i,j}(x),$$
$$\hat{z}_{i,j} = \left[\omega_i \cos(\omega_i t) - \frac{1}{t+1}\sin(\omega_i t)\right]\phi_{i,j}(x).$$

The left hand members in the moment problem (3.13), (3.14) become, respectively, $-[\nu_{i,j} + (1-\omega_i^2)\mu_{i,j}]$ and $\omega_i(\nu_{i,j} + \mu_{i,j})$.

For the equation

$$\Delta_n - \frac{\partial^2 u}{\partial t^2} + \frac{6}{(t+1)^2}u = 0$$

one has $q = 1$ and

$$l_1 = \frac{\partial}{\partial t} - \frac{2}{t+1}.$$

In this case the left members of (3.13) and (3.14) are, respectively, $(2-\omega_i^2)\nu_{i,j} + (4-3\omega_i^2)\mu_{i,j}$ and $-\omega_i[3\nu_{i,j} + (5-\omega_i^2)\mu_{i,j}]$.

Remark 3. D. L. Russell has considered the question of exact controllability of solutions of the wave equation when the control is applied on only a portion of the boundary, and has shown that data (u_0, v_0) in $H^1(\Omega)\times L^2(\Omega)$ can be steered to the zero state in some finite time T with a control $f \in L^2(\Gamma\times[0, T])$. Here Γ is a relatively open subset of Ω such that the pair (Ω, Γ) is "star-complemented"; see [14] for details. By using the transformation of § 3, analogous results can be obtained for operators of the type considered in the present paper.

REFERENCES

[1] F. V. ATKINSON, *A spectral problem for completely continuous operators*, Acta Math. Acad. Sci. Hungar., 3 (1952), pp. 53–60.

[2] W. C. CHEWNING, *Controllability of the nonlinear wave equation in several space variables*, this Journal, 14 (1976), pp. 19–25.

[3] B. M. N. CLARKE, *Boundary controllability of linear elastodynamic systems*, Quart. J. Mech. Appl. Math., 28 (1975), pp. 495–515.

[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. II, Interscience, New York, 1962.

[5] K. D. GRAHAM AND D. L. RUSSELL, *Boundary value control of the wave equation in a spherical region*, this Journal, 13 (1975), pp. 174–196.

[6] J. HADAMARD, *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale University Press, New Haven, CT, 1923.

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

[8] J. LAGNESE, *A solution of Hadamard's problem for a restricted class of operators*, Proc. Amer. Math. Soc., 19 (1968), pp. 981–988.

[9] ———, *The structure of a class of Huygens' operators*, J. Math. Mech., 18 (1969), pp. 1195–1201.

[10] J. LAGNESE AND K. L. STELLMACHER, *A method of generating classes of Huygens' operators*, Ibid., 17 (1967), pp. 461–472.

[11] D. L. RUSSELL, *Boundary value control of the higher dimensional wave equation, Parts I and II*, this Journal, 9 (1971), pp. 29–42 and 401–419.

[12] ———, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, J. Math. Anal. Appl., 40 (1972), pp. 336–368.

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

[14] ———, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, Differential Games and Control Theory, Roxin, Liu and Steinberg, eds., Marcel Dekker, New York, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

DISCRETE MAXIMUM PRINCIPLE WITH STATE CONSTRAINED CONTROL*

JOSÉ ANTONIO ORTEGA† AND RICHARD JEFFREY LEAKE‡

Abstract. The validity of a discrete maximum principle is proved for a class of problems in which the control constraint set depends on the system state using the formalism and basic optimization theorem of Cannon, Cullum, and Polak [Optimization, Control and Algorithm, 1970].

1. Introduction. Although the maximum principle for continuous systems was derived by Pontryagin et al. [2] over 15 years ago, the mathematical details of the proof of the discrete maximum principle have been cleared up only recently.

First mention of a maximum principle for discrete time systems can be attributed to Rozonoer [3]. Rozonoer's claim that "the extension of the maximum principle to discrete systems is possible, generally speaking, only in the linear case," together with faulty proofs by early researchers caused considerable confusion in the subject.

Beginning in 1964 and later in 1966 Halkin [4], [5] presented a careful mathematical proof of the discrete, maximum principle. Almost concurrently, Propoi [7] arrived at the same conclusion, namely that a rather strong convexity assumption is required. A study of state constrained controls using Halkin's methods was given by Bruckner and Wu [8]. Holtzman and Halkin [9]–[11] greatly extended the usefulness of Halkin's results, however, by introducing the concept of directional convexity.

Cannon, Cullum, and Polak [12] and Da Cunha and Polak [13] later presented a method for handling such problems, using a basic theorem of optimization and a systematic approach to conical approximations. Using this method they have greatly simplified the proof of the discrete maximum principle and have given extensions to include state space constraints and vector valued performance criteria. Their work is summarized in the book by Cannon, Cullum, and Polak [1], and to conserve space the notation and formalism of [1] will be assumed familiar to the reader in this work.

In the definition of $U_i(x_i)$, we allow that the set be empty. In this case it should be noted that further, implicit, constraints are imposed on the state in order that there exist an admissible control. The essential additional assumption which we have made in order to allow this type of constraint are the differentiability of the next state function in u, and the requirement that the set

$$X_i = \{(x, u): u \in U_i(x)\} \tag{1}$$

be convex. If X_i is defined by a constraint $R_i(x, u) \leqq 0$ by a differentiable function

* Received by the editors September 14, 1973, and in revised form November 11, 1976. This research was supported in part by BNDE and CNPq, Brasil, and by the U.S. Air Force Office of Scientific Research under Grant AFOSR-76-3036.

† University of California at Berkeley, on leave from COPPE, Universidade Federal do Rio de Janeiro, Brasil.

‡ Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana 46556.

R_i then explicit formulas can be given relating the various multipliers as in [1].

Remark. A set valued function U_i is biconvex if

1) $U_i(x)$ is convex for each x;

2) $\theta U_i(x)+\theta' U_i(x')\subset U_i(\theta x+\theta' x')$, where $\theta, \theta' \geqq 0$, $\theta+\theta'=1$.

It is easy to show that U_i is biconvex if and only if

$$X_i=\{(x, u): u\in U_i(x)\} \quad \text{is convex.}$$

Thus, our assumption of convexity on X_i is equivalent to the biconvexity of U_i. Halkin [6] uses this concept to obtain a very elegant proof of the "discrete maximum principle" and the slightly stronger "equilibrium price conditions" of mathematical economics. Halkins results are easily obtained from the results of this paper by taking $f_i(x_i, u_i)=u_i$.

2. Problem definition. We consider the system

$$x_{i+1}-x_i=f_i(x_i, u_i) \tag{2}$$

with $x_i\in E^n$, $u_i\in E^m$, $i=0, 1, \cdots, k-1$ and the problem of minimizing the scalar performance function

$$J=\sum_{i=0}^{k-1} f_i^0(x_i, u_i) \tag{3}$$

subject to (2), to $u_i\in U_i(x_i)$, $i=0, 1, \cdots, k-1$, where each X_i of (1) is a convex subset of $E^n\times E^m$, and to the terminal constraints

$$g_0(x_0)=0, \qquad g_k(x_k)=0 \tag{4}$$

where all functions are continuously differentiable and the Jacobians of g_0 and g_k have maximum rank where evaluated. Defining

$$F_i(x, u)=(f_i^0(x, u), f_i(x, u)) \tag{5}$$

we make the additional assumption that $F_i(x, U_i(x))$ is directionally convex for each $i=0, 1, \cdots, k-1$ and each x. That is, given x and u' and u'' in $U_i(x)$ and $0\leqq\lambda\leqq 1$ there exists $u(\lambda)$ in $U_i(x)$ such that

$$\begin{aligned} &f_i(x, u(\lambda))=\lambda f_i(x, u')+(1-\lambda) f_i(x, u''),\\ &f_i^0(x, u(\lambda))\leqq\lambda f_i^0(x, u')+(1-\lambda) f_i^0(x, u''). \end{aligned} \tag{6}$$

If $U_i(x)$ is empty then $F_i(x, U_i(x))$ is also empty and hence taken to be directionally convex. In what follows however, existence of optimal controls guarantees the nonemptiness of $U_i(\bar{x}_i)$.

3. Main results. In this section we present an extension of the discrete maximum principle. Let us define the Hamiltonian as

$$H_i(x_i, p_{i+1}, u_i)=p^0 f_i^0(x_i, u_i)+\langle p_{i+1}, f_i(x_i, u_i)\rangle \tag{7}$$

and the radial cone

$$\mathrm{RC}((\bar{x}_i, \bar{u}_i), X_i)=\{\lambda(x_i-\bar{x}_i, u_i-\bar{u}_i): \lambda\geqq 0, (x_i, u_i)\in X_i\}. \tag{8}$$

THEOREM 1. *If* $(\bar{u}_0, \bar{u}_1, \cdots, \bar{u}_{k-1})$, $(\bar{x}_0, \bar{x}_1, \cdots, \bar{x}_k)$ *is an optimal solution of the problem in* § 2, *then there exist vectors* $p_0, p_1, \cdots, p_k, \mu_0, \mu_k$, *and a scalar* $p^0 \leqq 0$, *not all zero, such that*

$$\left\langle \frac{\partial H_i(\bar{x}_i, p_{i+1}, \bar{u}_i)}{\partial(x, u)}^T, (\delta x_i, \delta u_i) \right\rangle + \langle p_{i+1} - p_i, \delta x_i \rangle \leqq 0 \tag{9}$$

for all

$$(\delta x_i, \delta u_i) \in \overline{\mathrm{RC}}((\bar{x}_i, \bar{u}_i), X_i) \tag{10}$$

where the bar denotes closure, for $i = 0, 1, \cdots, k-1$. *Furthermore,*

$$H_i(\bar{x}_i, p_{i+1}, u_i) \leqq H_i(\bar{x}_i, p_{i+1}, \bar{u}_i) \tag{11}$$

for all $u_i \in U_i(\bar{x}_i)$, $i = 0, 1, \cdots, k-1$ *and the transversality conditions*

$$p_0 = -\frac{\partial g_0(\bar{x}_0)}{\partial x}^T \mu_0, \qquad p_k = \frac{\partial g_k(\bar{x}_k)}{\partial x}^T \mu_k \tag{12}$$

are satisfied.

Proof. We follow the approach in [1], translating to the mathematical programming problem:

$$\text{minimize:} \quad f(z) = \sum_{i=0}^{k-1} v_i^0 \tag{13}$$

$$\text{subject to:} \quad r(z) = \begin{bmatrix} x_1 - x_0 - v_0 \\ x_2 - x_1 - v_1 \\ \vdots \\ x_k - x_{k-1} - v_{k-1} \\ g_0(x_0) \\ g_k(x_k) \end{bmatrix} = 0, \tag{14}$$

$$z \in \Omega' = \{z = ((x_0, u_0), \cdots, (x_k, u_k), V_0, \cdots, V_{k-1}):$$

$$(x_i, u_i) \in X_i,\ V_i \in \operatorname{co} F_i(x_i, U_i(x_i))\} \tag{15}$$

where co denotes convex hull,

$$v_i^0 = f_i^0(x_i, u_i), \quad v_i = f_i(x_i, u_i), \quad V_i = (v_i^0, v_i), \tag{16}$$

$$X_i = \{(x_i, u_i): u_i \in U_i(x_i)\} \tag{17}$$

for $i = 0, 1, \cdots, k-1$. $X_k = E^n \times E^m$ and u_k are artificial quantities added for symmetry. Assuming $\bar{z}$ is an optimal solution of the corresponding problem with Ω' replaced by Ω which is defined by replacing $F_i(x, U_i(x_i))$ by co $F_i(x_i, U_i))$ we introduce the convex cone $C(\bar{z}, \Omega')$ as the set of vectors

$$\delta z = ((\delta x_0, \delta u_0), \cdots, (\delta x_k, \delta u_k), \delta V_0, \cdots, \delta V_{k-1}) \tag{18}$$

such that

$$(\delta x_i, \delta u_i) \in \mathrm{RC}((\bar{x}_i, \bar{u}_i), X_i) \tag{19}$$

and

$$\delta V_i - \frac{\partial F_i(\bar{x}_i, \bar{u}_i)}{\partial(x, u)}(\delta x_i, \delta u_i) \in \mathrm{RC}(\bar{V}_i, \mathrm{co}\, F_i(\bar{x}_i, U_i(\bar{x}_i))) \tag{20}$$

where

$$\bar{V}_i = F_i(\bar{x}_i, \bar{u}_i) \tag{21}$$

and

$$\mathrm{RC}(\bar{V}_i, \mathrm{co}\, F_i(\bar{x}_i, U_i(\bar{x}_i))) = \{\lambda(V - \bar{V}_i) : \lambda \geqq 0,\ V \in \mathrm{co}\, F_i(\bar{x}_i, U_i(\bar{x}_i))\}. \tag{22}$$

Now let $\delta z_1, \delta z_2, \cdots, \delta z_p$ be a linearly independent set of vectors in $C(\bar{z}, \Omega')$ with

$$\delta z_j = ((\delta x_{0j}, \delta u_{0j}), \cdots, (\delta x_{kj}, \delta u_{kj}), \delta V_{0j}, \cdots, \delta V_{k-1j}). \tag{23}$$

Choose $\varepsilon > 0$ such that for $j = 1, 2, \cdots, p$

$$(\bar{x}_i, \bar{u}_i) + \varepsilon(\delta x_{ij}, \delta u_{ij}) \in X_i \tag{24}$$

for $i = 0, 1, \cdots, k$ and also

$$\bar{V}_i + \varepsilon\left(\delta V_{ij} - \frac{\partial F_i(\bar{x}_i, \bar{u}_i)}{\partial(x, u)}(\delta x_{ij}, \delta u_{ij})\right) \in \mathrm{co}\, F_i(\bar{x}_i, U_i(\bar{x}_i)) \tag{25}$$

for $i = 0, 1, \cdots, k-1$. Now there exist coefficients $\mu^j(z) \geqq 0$, $\sum_{j=1}^{p} \mu^j(z) \leqq 1$ such that for any $z \in \mathrm{co}\,(\bar{z}, \bar{z} + \varepsilon\delta z_1, \cdots, \bar{z} + \varepsilon\delta z_p)$ we have

$$\delta z = z - \bar{z} = \varepsilon \sum_{j=1}^{p} \mu^j(z)\delta z_j. \tag{26}$$

In fact, as in [1] the independence of the δz_j implies the existence of a matrix Y such that for $\mu(z) = (\mu^1(z), \cdots, \mu^p(z))$,

$$\mu(z) = Y\delta z. \tag{27}$$

Equation (25) implies the existence of controls

$$u_{ij}^{\alpha}(\bar{z}) \in U_i(\bar{x}_i) \tag{28}$$

$\alpha = 1, 2, \cdots, s_i$ and $\lambda_{ij}^{\alpha} \geqq 0$, $\sum_{\alpha=1}^{s_i} \lambda_{ij}^{\alpha} = 1$, such that

$$\bar{V}_i + \varepsilon\left(\delta V_{ij} - \frac{\partial F_i(\bar{x}_i, \bar{u}_i)}{\partial(x, u)}(\delta x_{ij}, \delta u_{ij})\right) = \sum_{\alpha=1}^{s_i} F_i(\bar{x}_i, u_{ij}^{\alpha}(\bar{z}))\lambda_{ij}^{\alpha}. \tag{29}$$

We now make the crucial definition

$$u_{ij}^{\alpha}(z) = \left(1 - \sum_{j=1}^{p} \mu^j(z)\right)u_{ij}^{\alpha}(\bar{z}) + \sum_{j=1}^{p} \mu^j(z)(\bar{u}_i + \varepsilon\delta u_{ij}) \tag{30}$$

such that (24), (28), and the convexity of X_i imply

$$u_{ij}^{\alpha}(z) \in U_i(z_i) \tag{31}$$

for $z \in \mathrm{co}\,(\bar{z}, \bar{z} + \varepsilon\delta z_i, \cdots, \bar{z} + \varepsilon\delta z_p)$ and x_i a component of z. Introducing

matrices B_i^α and $Z_i(z)$ such that the

$$jth\ column\ of\ Z_i(z) = \sum_{\alpha=1}^{s_i} \lambda_{ij}^\alpha F_i(x_i, u_{ij}^\alpha(z)) - F_i(x_i, u_i), \tag{32}$$

$$jth\ column\ of\ B_i^\alpha = \bar{u}_i - u_{ij}^\alpha(\bar{z}), \tag{33}$$

we have

$$u_{ij}^\alpha(z) = u_{ij}^\alpha(\bar{z}) + \delta u_i + B_i^\alpha Y \delta z \tag{34}$$

and

$$\delta V_i = \frac{\partial F_i(\bar{x}_i, \bar{u}_i)}{\partial(x, u)}(\delta x_i, \delta u_i) + Z_i(\bar{z}) Y \delta z \tag{35}$$

we next apply the extended basic theorem [1, p. 85] which shows that if there exists a continuous map

$$\zeta\colon \operatorname{co}(\bar{z}, \bar{z} + \varepsilon\delta z_1, \cdots, \bar{z} + \varepsilon\delta z_p) \to \Omega' \tag{36}$$

such that

$$\zeta(\bar{z} + \delta z) = \bar{z} + \delta z + o(\delta z) \tag{37}$$

with

$$\lim_{\|\delta z\| \to 0} \frac{\|o(\delta z)\|}{\|\delta z\|} = 0 \tag{38}$$

then there exist $p^0 \leqq 0$ and vectors $p_0, p_1, \cdots, p_k$, not all zero such that

$$\begin{aligned} p^0 \sum_{i=0}^{k-1} \delta v_i^0 + \sum_{i=0}^{k-1} \langle -p_{i+1}, (\delta x_{i+1} - \delta x_i - \delta v_i) \rangle \\ + \left\langle \mu_0, \frac{\partial g_0(\bar{x}_0)}{\partial x} \delta x_0 \right\rangle + \left\langle \mu_k, \frac{\partial g_k(\bar{x}_k)}{\partial x} \delta x_k \right\rangle \leqq 0 \end{aligned} \tag{39}$$

for all $\delta z \in C(z, \Omega')$. We define $\zeta(z) = (y_0(z), \cdots, y_k(z), W_0(z), \cdots, W_{k-1}(z))$ with

$$y_i(z) = (x_i, u_i), \tag{40}$$

$i = 0, 1, \cdots, k-1$, and

$$W_i(z) = F_i(x_i, u_i) + Z_i(z) Y \delta z. \tag{41}$$

It is now routine, using the crucial relation (31), to establish that $W_i(z)$ is a convex combination of elements of $F_i(x_i, U_i(x_i))$ and hence the range of ζ is in Ω'; and further, using (34) and continuous differentiability, that

$$Z_i(\bar{z} + \delta z) Y \delta z = Z_i(\bar{z}) Y \delta z + O''(\delta z), \tag{42}$$

so

$$W(\bar{z} + \delta z) = \bar{V}_i + \delta V_i + V_i + O'(\delta z). \tag{43}$$

Thus, all conditions of the extended basic theorem [1] are satisfied, and using

respectively the variations

$$\delta z = (0, \cdots, 0, (\delta x_i, \delta u_i), 0, \cdots, 0, \delta V_i, 0, \cdots, 0),$$

$$\delta z = (0, \cdots, 0, \delta V_i, 0, \cdots, 0), \tag{44}$$

$$\delta z = (0, \cdots, 0, (\delta x_k, \delta u_k), 0, \cdots, 0)$$

in (39) yields all the conditions of the theorem when the relation in p_0 is taken as a definition.

THEOREM 2. *Suppose that all of the assumptions of the previous theorem are satisfied with*

$$X_i = \{(x_i, u_i): R_i(x_i, u_i) \leqq 0\}, \tag{45}$$

$$U_i(x_i) = \{u_i: R_i(x_i, u_i) \leqq 0\}, \tag{46}$$

$i = 0, 1, \cdots, k-1$, *where each* $R_i: E^n \times E^m \to E^{1_i}$ *is continuously differentiable and the gradients of the active constraints are linearly independent; that is,* $\{\nabla R_i^j(\bar{x}_i, \bar{u}_i): j \in I(\bar{x}_i, \bar{u}_i)\}$ *is a linearly independent set, where* $I(\bar{x}_i, \bar{u}_i) = \{j: R_i^j(\bar{x}_i, \bar{u}_i) = 0,\ i \leqq j \leqq 1_i\}$. *Then there exist vectors* $p_0, p_1, \cdots, p_k, \mu_0, \mu_k$ *and a scalar* $p^0 \leqq 0$, *not all zero, and also vectors* $\lambda_0, \lambda_1, \cdots, \lambda_{k-1}, \lambda_i \leqq 0\ i = 0, 1, \cdots, k$ *such that*

$$-(p_{i+1} - p_i) = \frac{\partial H(\bar{x}_i, p_{i+1}, \bar{u}_i)^T}{\partial x} + \frac{\partial R_i(\bar{x}_i, \bar{u}_i)^T}{\partial x}\lambda_i, \tag{47}$$

$$0 = \frac{\partial H_i(\bar{x}_i, p_{i+1}, \bar{u}_i)^T}{\partial u} + \frac{\partial R_i(\bar{x}_i, \bar{u}_i)^T}{\partial u}\lambda_i, \tag{48}$$

$$\langle R_i(\bar{x}_i, \bar{u}_i), \lambda_i \rangle = 0, \ i = 0, 1, \cdots, k-1. \tag{49}$$

Furthermore,

$$H_i(\bar{x}_i, p_{i+1}, u_i) \leqq H_i(\bar{x}_i, p_{i+1}, \bar{u}_i) \tag{50}$$

for all $u_i \in U_i(\bar{x}_i)$, $i = 0, 1, \cdots, k-1$, *and the transversality conditions*

$$p_0 = -\frac{\partial g_0(\bar{x}_0)^T}{\partial x}\mu_0, \qquad p_k = \frac{\partial g_k(\bar{x}_k)^T}{\partial x}\mu_k \tag{51}$$

are satisfied.

Proof. The linear independence of the gradient vector implies that the set

$$\left\{(\delta x_i, \delta u_i): \left\langle \frac{\partial R_i^j(\bar{x}_i, \bar{u}_i)^T}{\partial(x, u)}, (\delta x_i, \delta u_i) \right\rangle \leqq 0 \quad \text{for } j \in I(\bar{x}_i, \bar{u}_i)\right\} \tag{52}$$

is contained in $\mathrm{RC}((\bar{x}_i, \bar{u}_i), X_i)$ and as such we may apply Farkas' lemma [1] to (9) and (52) to yield (47), (48), (49). The equations (50) and (51) follow from Theorem 1.

Comment. Note that convexity of X_i is required in (45), and this is implied in the instance that each $R_i^j(x_i, u_i)$ is a convex function.

4. Conclusions. The principal contribution of this work is a valid proof of the extension of the results of Cannon, Cullum, and Polak [1] to allow dependency of the control constraint set U_i on the state x_i. Problems of this type are frequently encountered in economic and industrial problems. These results compliment the dynamic programming approach, in which there is little difficulty in handling state constrained controls.

REFERENCES

[1] M. D. CANNON, C. D. CULLUM AND E. POLAK, *Optimization, Control and Algorithm*, McGraw-Hill, New York, 1970.

[2] L. S. PONTRYAGIN ET AL., *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[3] L. I. ROZONOER, *The maximum principle of L. S. Pontryagin in optimal-system theory, Part III*, Automat. Remote Control, 20 (1970), pp. 1515–1532.

[4] H. HALKIN, *Optimal control for systems described by difference equations*, Advances in Control Systems, C. T. Leondes, ed., Academic Press, New York, 1964, Chap. 4.

[5] ———, *A maximum principle of the Pontryagin type for systems described by nonlinear difference equations*, this Journal, 4 (1966), pp. 90–111.

[6] ———, *External properties of biconvex contingent equations*, Ordinary Differential Equations 1971 NRL–MRC Conference, Leonard Weiss, ed., Academic Press, New York, 1972, pp. 109–119.

[7] A. I. PROPOI, *The maximum principle for discrete systems*, Automat. Remote Control, 26 (1965), pp. 1169–1177.

[8] J. BRUCKNER AND S. WU, *A maximum principle for discrete systems with control variable inequality constraints*, Proc. Sixth Annual Allerton Conf. on Cir. and Sys. Theory, 1968, pp. 475–484.

[9] J. M. HOLTZMAN, *Convexity and the maximum principle for discrete systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 30–36.

[10] J. M. HOLTZMAN AND H. HALKIN, *Directional convexity and the maximum principle for discrete systems*, this Journal, 4 (1966), pp. 263–275.

[11] J. M. HOLTZMAN, *On the maximum principle for nonlinear discrete-time systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 273–274.

[12] M. CANNON, C. CULLUM AND E. POLAK, *Constrained minimization problems in finite-dimensional spaces*, this Journal, 4 (1966), pp. 528–547.

[13] N. O. DA CUNHA AND E. POLAK, *Constrained minimization under vector-valued criteria in finite dimensional spaces*, J. Math. Anal. Appl., 19 (1967), pp. 103–124.

[14] E. R. VON STOCKERT, *Condicoes de Otimalidade para Sistemas Discretos no Tempo com Controles Limitados pelo Estado*, M. Sc. Thesis, COPPE, Rio de Janeiro, Brasil, 1973.

[15] M. L. J. HAUTUS, *Necessary conditions for multiple constraint optimization problems*, this Journal, 11 (1973), pp. 653–669.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

ON LOWER SEMICONTINUITY OF INTEGRAL FUNCTIONALS. II*

A. D. IOFFE†

Abstract. A necessary and sufficient condition for the integral functional $\int_G f(t, x(t), y(t))\, d\mu$ to be lower semicontinuous with respect to norm convergence of $x(\cdot)$-components in $L_s(G, R^m)$ and weak convergence of $y(\cdot)$-components in $L_q(G, R^n)$ $(1<q<\infty)$ is established.

1. Introduction. Let G be a measure space with finite positive measure μ, and let $f(t, x, y)$ be a function on $G\times R^m\times R^n$ with values in $(-\infty, \infty]$. In [1] we proved a theorem containing a general necessary and sufficient condition for the integral functional

$$I(x(\cdot), y(\cdot)) = \int_G f(t, x(t), y(t))\, d\mu$$

to be sequentially lower semicontinuous (l.s.c.) relative to a spectrum of mixed strong-weak topologies. Here we consider in more detail the case when G is a *bounded domain* in a finite dimensional Euclidean space, μ is the ordinary *Lebesgue measure* on G and the type of convergence in question is that defined by the norm topology of $L_s(G, R^m)$ for $x(\cdot)$-components and by the weak topology of $L_q(G, R^n)$ $(1<q<\infty)$ for $y(\cdot)$-components. Here $L_p(G, R^k)$ is the space of measurable mappings from G into R^k with the usual p-norm. For simplicity, we shall write merely L_p, not $L_p(G, R^k)$, since this can lead to no confusion. We exclude the case $q=1$ fully investigated by Olech [4]. (As has been shown in [1], Theorem 5, Olech's result is a direct corollary of the main theorem of [1].)

In [1] we assumed f to be $\mathscr{L}\otimes\mathscr{B}$-measurable. Here the particular choice of the measure space allows us to deal with a more convenient, though somewhat broader, class of functions. Let $A\subset G\times R^k$. We shall say that A is an *almost Borel set* if there is a Borel set $A'\subset G\times R^k$ such that $\mu(\mathrm{pr}_G(A\triangle A'))=0$ ($\triangle$ stands for symmetrical difference, pr_G denotes the projection on G). Note that every $\mathscr{L}\otimes\mathscr{B}$-measurable subset in $G\times R^k$ ($\mathscr{L}$ and $\mathscr{B}$ being the algebras of Lebesgue measurable subsets of G and Borel subsets of R^k respectively) is almost Borel, since each Lebesgue measurable set in G contains an F_σ-subset of the same measure. It is clear that almost Borel sets form a σ-algebra. Mappings or functions measurable with respect to this algebra will be called *almost B-measurable*. It is likewise clear that $\mathscr{L}\otimes\mathscr{B}$-measurable mappings are almost B-measurable and that every almost B-measurable mapping coincides with a B-measurable mapping up to a set with measure-negligible projection on G. We shall suppose:

(A) *f is almost B-measurable, l.s.c. in (x, y) and convex in y.*

Note that, according to Theorem 2 in [1], these assumptions are quite natural and, in fact, almost necessary for $I(\cdot, \cdot)$ to be l.s.c.

The purpose of the paper is to prove the following result.

* Received by the editors February 17, 1976.
† Profsojuznaja 97-1-203, Moscow B-279, USSR.

THEOREM. *Let* (A) *be satisfied, and let* $|I(x_0(\cdot), y_0(\cdot))| < \infty$ *for some* $x_0(\cdot) \in L_s$, $y_0(\cdot) \in L_q$. *For* $I(\cdot, \cdot)$ *to be l.s.c. with respect to norm convergence of* $x(\cdot)$*-components in* L_s $(1 \leqq s < \infty)$ *and weak convergence of* $y(\cdot)$*-components in* L_q $(1 < q < \infty)$, *it is necessary and sufficient that there exist an almost B-measurable mapping* $p(t, x)$: $G \times R^m \to R^n$ *such that*

$$\text{(i)} \qquad f(t, x, y) \geqq \langle p(t, x), y\rangle - c|x|^s - b(t), \quad \forall t, x, y$$

for some $c \in R$, $b(t) \in L_1$;

(ii) *if* $x_k(\cdot)$ *norm converges in* L_s, $y_k(\cdot)$ *weakly converges in* L_q *and* $I(x_k(\cdot), y_k(\cdot)) \leqq a < \infty$ *for all* $k = 1, 2, \cdots$, *then the sequence of functions*

$$t \to |p(t, x_k(t))|^{q'}$$

is weakly precompact in L_1.

Here q' is defined by $1/q + 1/q' = 1$, and $|\cdot|$ and $\langle \cdot, \cdot \rangle$ denote the Euclidean norm and the inner product respectively.

Remark 1. For $s = \infty$ two changes should be made. Firstly, the beginning of the second sentence in the Theorem should be as follows: for $I(\cdot, \cdot))$ to be everywhere greater than $-\infty$ on $L_s \times L_q$ and l.s.c. . . . . Secondly, the term $c|x|^s$ in the right-hand part of (i) should be replaced by $r(|x|)$, where $r(\lambda)$ is a nondecreasing real-valued function on the positive half-line.

Remark 2. The theorem remains valid if (i) holds with $c = 0$ and norm convergence of $x(\cdot)$-components is replaced in the statement by convergence in measure.

The proof of the theorem is contained in the following two sections. In the concluding section we prove, under an additional assumption, another criterion which is more convenient to verify.

2. Two basic lemmas. Let $g(y)$ be a convex function on R^n satisfying

$$g(y) \geqq -|y|^q/q, \quad \forall y \in R^n. \tag{2.1}$$

Then there are $p \in R^n$ and $\beta \in R$ such that

$$g(y) \geqq \langle p, y\rangle + \beta \geqq -|y|^q/q \quad \forall y \in R^n. \tag{2.2}$$

Denote by $\Pi(g)$ the set of all $p \in R^n$ satisfying (2.2) together with some $\beta \in R$ (depending on p). It is easy to see that $p \in \Pi(g)$ if and only if

$$|p|^{q'}/q' + g^*(p) \leqq 0, \tag{2.3}$$

where $g^*(p)$ is the Fenchel conjugate to g. (Here and below we use the standard terminology of convex analysis without explanations.) Obviously, $\Pi(g)$ is convex and compact unless $g(y) \equiv \infty$ while in the latter case $\Pi(g)$ contains the origin. In either case there exists a unique vector $\pi(g) \in \Pi(g)$ such that

$$|\pi(g)| = \min\{|p| \mid p \in \Pi(g)\}.$$

In other words, $\pi(g)$ is the unique solution to the problem

$$\text{minimize } |p| \tag{2.4}$$

subject to (2.3).

LEMMA 1. *Let g satisfy* (2.1). *Denote for brevity* $p_0 = \pi(g)$. *If* $p_0 \neq 0$, *then there is* $y_0 \in R^n$ *such that*

$$g(y_0) \leqq -\frac{\sqrt{q'}-1}{q'}|p_0|\,|y_0|, \tag{2.5}$$

$$|y_0|^q \geqq |p_0|^{q'}. \tag{2.6}$$

Proof. Let $r = |p_0|^{q'}/q'$, $z = |p_0|^{q'-1}(p_0/|p_0|)$, the maximal value and the maximum point of the function $y \mapsto \langle p_0, y\rangle - |y|^q/q$. To prove the lemma, it is sufficient to find $y_0 \in R^n$ which satisfies

$$g(y_0) \leqq -(1/\sqrt{q'})|p_0|\,|y_0| + r, \tag{2.7}$$

$$|y_0| \geqq |z|. \tag{2.8}$$

Indeed, (2.8) is the same as (2.6). On the other hand, (2.7) and (2.8) imply together

$$g(y_0) \leqq -\frac{1}{\sqrt{q'}}\left(1-\frac{1}{\sqrt{q'}}\right)|p_0|\,|y_0| - \frac{1}{q'}|p_0|\,|y_0| + r$$

$$\leqq -\frac{\sqrt{q'}-1}{q'}|p_0|\,|y_0| - \frac{1}{q'}|p_0|\,|z| + r = -\frac{\sqrt{q'}-1}{q'}|p_0|\,|y_0|.$$

We shall consider two possible situations and show that in either of them (2.7), (2.8) hold.

1) Assume that

$$|p_0|^{q'}/q' + g^*(p_0) = 0. \tag{2.9}$$

Since p_0 is a solution to (2.3), (2.4) and $p_0 \neq 0$, there is $\lambda \geqq 0$ such that

$$0 \in \partial(\lambda|p_0| + |p_0|^{q'}/q' + g^*(p_0)).$$

Two of the three bracketed functions are continuous. Hence summation and subdifferentiation operations commutate and there is a $y \in \partial g^*(p_0)$ such that

$$y_0 = -(\lambda p_0)/|p_0| - z. \tag{2.10}$$

By definition, z is positively proportional to p_0 and hence

$$\langle p_0, y_0\rangle = -|p_0|\,|y_0|. \tag{2.11}$$

Thus (2.10) implies (2.8). On the other hand, using (2.9)–(2.11) we get

$$g(y_0) = \langle p_0, y_0\rangle - g^*(p_0) = -|p_0|\,|y_0| + r$$

which implies (2.7).

2) Assume that

$$|p_0|^{q'}/q' + g^*(p_0) < 0.$$

This may happen only if p_0 is the unique point of dom g^* nearest to the origin, which means in particular that

$$\langle -p_0, p_0 - p\rangle \geqq 0, \quad \forall p \in \operatorname{dom} g^*$$

because of convexity of dom g^*.

Note that $g(y)$ is not identically equal to infinity, since $p_0 \neq 0$ by the assumptions. Hence for any $y \in \operatorname{dom} g$ and any $\lambda > 0$ the following inequality holds (see [6, Thms. 8.5 and 13.3]):

$$\begin{aligned} g(y - \lambda p_0) &\leqq g(y) + \max\{-\langle \lambda p_0, p\rangle \mid p \in \operatorname{dom} g^*\} \\ &= g(y) - \lambda\langle p_0, p_0\rangle \\ &= g(y) - \langle p_0, y\rangle + \langle y - \lambda p_0, p_0\rangle. \end{aligned}$$

Since $p_0 \neq 0$ and $q' > 1$, the latter inequality shows that (2.7), (2.8) will be satisfied for $y_0 = y - \lambda p_0$ if λ is sufficiently large.

The following proposition, though playing a subsidiary role here, seems to us very useful in itself (cf. [10] where a similar result was proved under much stronger assumptions).

PROPOSITION. *Let $g(t, x, y)$ be an extended-real-valued function on $G \times R^m \times R^n$ which is almost B-measurable and l.s.c. in y. Let*

$$g^*(t, x, p) = \sup_y (\langle p, y\rangle - g(t, x, y))$$

be the Fenchel conjugate to $g(t, x, \cdot)$. Then $g^*(t, x, p)$ is also almost B-measurable.

Proof. Take a Borel set $G' \subset G$ such that $\mu G' = \mu G$ and $g(t, x, y)$ is B-measurable on $G' \times R^m \times R^n$. Define the following multifunction from $G' \times R^m$ into $R^n \times R$:

$$\Gamma(t, x) = \operatorname{epi} g(t, x, \cdot) = \{(y, \alpha) \mid \alpha \geqq g(t, x, y)\}.$$

According to the choice of G', the graph of Γ is a Borel set. On the other hand, Γ is closed-valued since g is l.s.c. in y. It follows from the Novikov projection theorem that Γ is B-measurable, which is to say that every set

$$\{(t, x) \mid t \in G', x \in R^m, \Gamma(t, x) \cap C \neq \varnothing\}$$

is Borel whenever $C \subset R^n \times R$ is closed (see, for instance, [9, Thm. 1.6]). This in turn implies that $g^*(t, x, p)$ is B-measurable on $G' \times R^m \times R^n$ ([8, Proposition 2S]).

LEMMA 2. *Let $g(t, x, y)$ be an extended-real-valued function on $G \times R^m \times R^n$ which is almost B-measurable, convex and l.s.c. in y and satisfies*

$$g(t, x, y) \geqq -|y|^q/q.$$

Let $p(t, x) = \pi(g(t, x, \cdot))$. Then the mapping $p(t, x)$: $G \times R^m \to R^n$ is almost B-measurable.

Proof. Let G' be the same as in the above proof. According to Proposition, the set

$$\{(t, x, p) \mid t \in G', x \in R^m, g^*(t, x, p) + |p|^{q'}/q' \leqq 0\}$$

is Borel.

Denote $P(t, x) = \Pi(g(t, x, \cdot))$. Then P is a closed-valued multifunction and it follows from what we have just established that the graph of the restriction of P on $G' \times R^m$ is a Borel set. Therefore (again, according to [9, Thm. 1.6]) P is a B-measurable multifunction on $G' \times R^m$ and

$$p(t, x) = \operatorname{prox} P(t, x)$$

is B-measurable on $G' \times R^m$ (see [7]) which implies the required result.

3. Proof of the theorem. The sufficiency part of the theorem follows immediately from [1]. Hence we need only to verify necessity.

Assume that $I(\cdot, \cdot)$ is sequentially l.s.c. relative to the above specified type of convergence. Since $I(\cdot, \cdot)$ is not everywhere on $L_s \times L_q$ equal to $\pm\infty$,

$$f(t, x, y) \geqq -c|x|^s - c|y|^q - b(t)$$

for some $c \in R$, $b(\cdot) \in L_1$. For f satisfying the Carathéodory condition, this was proved by Poljak [5]. In our case the proof needs no changes. Let

$$g(t, x, y) = (q|c|)^{-1}(f(t, x, y) + |b(t)|) + q^{-1}|x|^s.$$

Then

$$g(t, x, y) \geqq -|y|^q/q. \tag{3.1}$$

Clearly, g is almost B-measurable, l.s.c. in (x, y) and convex in y. Likewise, the functional

$$J(x(\cdot), y(\cdot)) = \int_G g(t, x(t), y(t))\, d\mu$$

is l.s.c. in the same sense as $I(\cdot, \cdot)$ and if $I(x_k(\cdot), y_k(\cdot))$ $(k = 1, 2, \cdots)$ are upper bounded and $x_k(\cdot)$ are norm bounded in L_s, then $J(x_k(\cdot), y_k(\cdot))$ are also upper bounded. Therefore we may prove the theorem for g and J instead of f, I.

Let $p(t, x) = \pi(g(t, x, \cdot))$. Then $p(t, x)$ is almost B-measurable and (i) is satisfied by definition. Hence we have only to prove that $p(t, x)$ satisfies the condition (ii) of the theorem.

Let $x_k(\cdot) \to x(\cdot)$ strongly in L_s, $y_k(\cdot) \to y(\cdot)$ weakly in L_q and

$$J(x_k(\cdot), y_k(\cdot)) \leqq a_1 < \infty \quad \forall k = 1, 2, \cdots. \tag{3.2}$$

Let us denote $p_k(t) = p(t, x_k(t))$. We must verify that the functions $|p_k(t)|^{q'}$, $k = 1, 2, \cdots$, form a weakly precompact set in L_1 or, in other words, that they are equi-uniformly summable, that is,

$$\int_T |p_k(t)|^{q'}\, d\mu \to 0 \quad \text{uniformly in } k$$

as $\mu T \to 0$.

To prove this, we shall assume the contrary and come to a contradiction. First we note that

$$\|y_k(\cdot)\|_q \leqq a_2 < \infty, \quad \forall k = 1, 2, \cdots, \tag{3.3}$$

due to the fact that $y_k(\cdot)$ converge weakly. This along with (2.1), (3.3) implies that

$$\int_G |g(t, x_k(t), y_k(t))|\, d\mu \leqq a_3 < \infty, \quad \forall k = 1, 2, \cdots. \tag{3.4}$$

If $|p_k(t)|^{q'}$ are not equi-uniformly summable, then there are $\delta > 0$ and a sequence $\{T_k\}$ of measurable subsets of G such that

$$\mu T_k \to 0, \qquad \limsup_{k\to\infty} \int_{T_k} |p_k(t)|^{q'}\, d\mu > \delta.$$

With no loss of generality, we may assume that

$$\int_{T_k} |p_k(t)|^{q'}\, d\mu \geqq \delta, \qquad k = 1, 2, \cdots, \tag{3.5}$$

and

$$p_k(t) \neq 0 \quad \forall t \in T_k, \quad \forall k = 1, 2, \cdots. \tag{3.6}$$

By Lemma 1, for any k and any $t \in T_k$, the set

$$B_k(t) = \{y \mid g(t, x_k(t), y) \leqq -\xi |p_k(t)|\, |y|,\ |y|^q \geqq |p_k(t)|^{q'}\}$$

where $\xi = (1/q')\,(\sqrt{q'} - 1)$, is nonempty and closed. Furthermore, the multifunction $t \to B_k(t)$ is Lebesgue measurable (because its graph is obviously almost Borel and (G, μ) is a complete measure space [8]). Hence (see [3], [8]) for any k, we can find a measurable mapping $z_k(\cdot)\colon G \to R^n$ such that $z_k(t) \in B_k(t)$ a.e. on T_k. In other words, a.e. on T_k the following inequalities hold:

$$g(t, x_k(t), z_k(t)) \leqq -\xi |p_k(t)|\, |z_k(t)|, \tag{3.7}$$

$$|z_k(t)|^q \geqq |p_k(t)|^{q'}. \tag{3.8}$$

Let

$$T'_{kN} = \{t \in T_k \mid |z_k(t)|^q \leqq N |p_k(t)|^{q'}\},$$

$$T''_{kN} = \{t \in T_k \mid |z_k(t)|^q > N |p_k(t)|^{q'}\},$$

$$F_{kN} = \int_{T'_{kN}} |z_k(t)|^q\, d\mu + N \int_{T''_{kN}} |p_k(t)|^{q'}\, d\mu.$$

We shall consider three cases which together cover all possible situations and show that each case is contradictory.

Case 1.

$$\liminf_{k,N\to\infty} F_{kN} < \infty.$$

No loss of generality will follow if we assume that $F_{kN} \leqq b < \infty$ for all k, N. Fix some N such that $b/N < \delta/2$. Then

$$\int_{T'_{kN}} |z_k(t)|^q\, d\mu \leqq b, \qquad \int_{T'_{kN}} |p_k(t)|^{q'}\, d\mu \geqq \delta/2. \tag{3.9}$$

Choose some $x_0(\cdot)\in L_s$, $y_0(\cdot)\in L_q$ such that $|J(x_0(\cdot), y_0(\cdot))|<\infty$, and let

$$w_k(t)=x_0(t)+\chi_{T'_{kN}}(t)(x_k(t)-x_0(t)),$$
$$u_k(t)=y_0(t)+\chi_{T'_{kN}}(t)(z_k(t)-y_0(t)),$$

where $\chi_T(t)$ denotes the characteristic function of T. Then $w_k(\cdot)\in L_s$ and converge strongly to $x_0(\cdot)$ because $\mu T_k\to 0$ and $x_k(\cdot)$ converge strongly to $x_0(\cdot)$. Likewise, the first inequality in (3.9) together with $\mu T'_{kN}\to 0$ shows that $u_k(\cdot)$ belong to L_q and converge weakly to $y_0(\cdot)$. At the same time, by (3.7)–(3.9),

$$\begin{aligned}
&J(w_k(\cdot), u_k(\cdot))-J(x_0(\cdot), y_0(\cdot)\\
&\quad=\int_{T'_{kN}}(g(t, x_k(t), z_k(t))-g(t, x_0(t), y_0(t)))\,d\mu\\
&\quad\leqq-\xi\int_{T'_{kN}}|p_k(t)|\,|z_k(t)|\,d\mu+b_k\\
&\quad\leqq-\xi\int_{T'_{kN}}|p_k(t)|^{q'}\,d\mu+b_k\\
&\quad\leqq-(1/2)\xi\delta+b_k,
\end{aligned}$$

where

$$b_k=-\int_{T'_{kN}}g(t, x_0(t), y_0(t))\,d\mu\to 0,\quad \text{if } k\to\infty.$$

This, however, contradicts to the fact that $J(\cdot,\cdot)$ is l.s.c.

Case 2.

$$\liminf_{k,N\to\infty}F_{kN}=\infty;\qquad \liminf_{k,N\to\infty}\int_{T''_{kN}}|p_k(t)|^{q'}\,d\mu<\infty.$$

In this case we may assume that

$$\liminf_{k,N\to\infty}\int_{T'_{kN}}|z_k(t)|^q\,d\mu=\infty$$

and hence

$$\lim_{k,N\to\infty}\int_{T'_{kN}}|p_k(t)|^{q'}\,d\mu=\infty.$$

It follows that we can choose $N>0$ such that

$$N\int_{T'_{kN}}|p_k(t)|^{q'}\,d\mu\geqq 1,\quad \forall k=1, 2, \cdots.$$

We can also find measurable sets $E_k\subset T'_{kN}$ such that

$$N\int_{E_k}|p_k(t)|^{q'}\,d\mu=1.$$

Then

$$\int_{E_k} |z_k(t)|^q \, d\mu \leqq 1.$$

The two last relations are analogous to (3.9) and hence they are also contradictory.

Case 3.

$$\liminf_{k,N\to\infty} \int_{T''_{kN}} |p_k(t)|^{q'} \, d\mu = \infty.$$

As above, we can choose $N>0$ and $E_k \subset T''_{kN}$ such that

$$\xi(N^{1/q}) \int_{E_k} |p_k(t)|^{q'} \, d\mu = a_3 + 1, \quad \forall k = 1, 2, \cdots \tag{3.10}$$

(a_3 being the same as in (3.4)).

Define functions $\alpha_k(t)$ on E_k by

$$|\alpha_k(t) z_k(t)|^q = N |p_k(t)|^{q'}. \tag{3.11}$$

Then $\alpha_k(\cdot)$ are measurable and $0 \leqq \alpha_k(t) \leqq 1$ a.e. on E_k according to the definition of T''_{kN}. Let

$$w_k(t) = x_0(t) + \chi_{E_k}(t)(x_k(t) - x_0(t)),$$

$$u_k(t) = y_0(t) + \chi_{E_k}(t)(\alpha_k(t) z_k(t) + (1-\alpha_k(t)) y_k(t) - y_0(t)),$$

where $x_0(t)$ and $y_0(t)$ are the same as above. As in the first case, we see that $w_k(\cdot) \to x_0(\cdot)$ strongly in L_s. On the other hand, by (3.3), (3.10), (3.11),

$$\begin{aligned} 2^{-q} \int_{E_k} |u_k(t)|^q \, d\mu &\leqq \textstyle\int_{E_k} |\alpha_k(t) z_k(t)|^q \, d\mu + \int_{E_k} |y_k(t)|^q \, d\mu \\ &\leqq \xi^{-1}(N^{1/q})(a_3+1) + a_2^q, \quad \forall k = 1, 2, \cdots, \end{aligned}$$

which shows that the norms of $u_k(\cdot)$ in L_q are bounded and hence $u_k(\cdot)$ weakly converge to $y_0(\cdot)$. Finally

$$\begin{aligned} \int_{E_k} g(t, w_k(t), u_k(t)) \, d\mu &\leqq \int_{E_k} \alpha_k(t) g(t, x_k(t), z_k(t)) \, d\mu \\ &\quad + \int_{E_k} (1-\alpha_k(t)) g(t, x_k(t), y_k(t)) \, d\mu \\ &\leqq -\xi \int_{E_k} |p_k(t)| \, |\alpha_k(t) z_k(t)| \, d\mu + \int_{E_k} |g(t, x_k(t), y_k(t))| \, d\mu \\ &\leqq -\xi(N^{1/q}) \int_{E_k} |p_k(t)|^{q'} \, d\mu + a_3 \leqq -1 \end{aligned}$$

because of (3.4), (3.7), (3.10). (3.11) and due to the fact that g is convex in y. As in the first case, this shows that $J(\cdot, \cdot)$ is not l.s.c. This completes the proof of the theorem.

4. A corollary and its proof.

COROLLARY. *In addition to the assumptions of the theorem, suppose that* $|I(x(\cdot), y_0(\cdot))| < \infty$ *for some* $y_0(\cdot) \in L_q$ *and all* $x(\cdot) \in A$, *where* A *is an open set in* $L_s (s < \infty)$. *Then* $I(\cdot, \cdot)$ *is l.s.c. with respect to norm convergence of* $x(\cdot)$*-components in* L_s *and weak convergence of* $y(\cdot)$*-components in* L_q *if and only if there exists an almost* B*-measurable mapping* $p(t, x)$: $G \times R^m \to R^n$ *such that for some* $c > 0$, $b(\cdot) \in L_1$, *inequalities*

(i) $f(t, x, y) \geqq \langle p(t, x), y \rangle - c|x|^s - b(t)$,

(ii) $|p(t, x)|^{q'} \leqq c|x|^s + b(t)$

hold everywhere on $G \times R^m \times R^n$ *up to a set with* μ*-negligible projection on* G.

Proof. In [1, Thm. 8] we have shown that this condition is sufficient. Let $I(\cdot, \cdot)$ be l.s.c. Choose $p(t, x)$ according to the theorem of § 1. Then condition (ii) of the theorem shows that $|p(t, x(t))|^{q'}$ is summable if $x(\cdot) \in A$. We claim that, in fact, this is true for all $x(\cdot) \in L_s$. Indeed, suppose that

$$\int_G |p(t, x(t))|^{q'} \, d\mu = \infty \tag{4.1}$$

for some $x(\cdot) \in L_s$. Choose arbitrarily a $x_0(\cdot) \in A$. Then it is possible to find $\varepsilon > 0$ such that

$$w_T(t) = x_0(t) + \chi_T(t)(x(t) - x_0(t))$$

belongs to A if $\mu T < \varepsilon$. On the other hand, (4.1) shows that for some measurable $T \subset G$ with $\mu T < \varepsilon$,

$$\int_T |p(t, x(t))|^{q'} \, d\mu = \infty.$$

In this case, $|p(t, w_T(t))|^{q'}$ is not summable which contradicts the fact that $w_T(\cdot) \in A$. Hence (4.1) is wrong.

Thus $p(t, x(t))$ belongs to $L_{q'}$ for every $x(\cdot) \in L_s$. But this is the same as (ii). This fact was proved in [2] under the additional assumption that $p(t, x)$ satisfies the Carathéodory condition. But the proof given there demands nothing beyond measurable choice which is possible in our case since $p(t, x)$ is almost B-measurable.

Note. The following condition was introduced by Cesari (L. Cesari, *Closure theorems for orientor fields and weak convergence*, Arch. Rational Mech. Anal., 55 (1974), pp. 332–356): Given a sequence $\{x_k(\cdot), y_k(\cdot)\}$ converging in the desired sense, there is a weakly converging sequence $\{a_k(\cdot)\} \subset L_1$ such that $f(t, x_k(t), y_k(t)) \geqq a_k(t)$. As far as sufficiency (not necessity!) is concerned, this condition is equivalent to the lower compactness property. But in the above-mentioned work as well as in a recently published paper of L. Cesari and M. B. Suryanarayana, *Nemytsky's operators and lower closure theorems*, J. Optimization Theory Appl., 19 (1976), pp. 165–183, this condition is accompanied with other assumptions such as property (Q) or its weakened versions, the Carathéodory condition and certain others which are needless for lower semicontinuity purposes.

Acknowledgment. I am thankful to C. Olech and to the referees for pointing out this condition of Cesari and for some other helpful remarks.

REFERENCES

[1] A. D. IOFFE, *On lower semicontinuity of integral functionals. I*, this Journal, 15 (1977), pp. 521–538.

[2] M. A. KRASNOSEL'SKII, P. P. ZABREIKO, E. I. PUSTYL'NIK AND P. E. SOBOLEVSKII, *Linear Operators in Spaces of Summable Functions*, Nauka, Moscow, 1966. (In Russian.)

[3] K. KURATOWSKI AND C. RYLL-NARDZEWSKI, *A general theorem on selectors*, Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys., 13 (1965), pp. 273–411.

[4] C. OLECH, *Weak lower semicontinuity of integral functionals I*, J. Optimization Theory Appl., 19 (1976).

[5] B. T. POLJAK, *Semicontinuity of integral functionals and existence theorem on extremal problems*, Mat. Sb. 78, (1969), pp. 65–84 = Math. USSR Sb., 7 (1969), pp. 59–77.

[6] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[7] ———, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

[8] ———, *Integral functionals, normal integrands and measurable selections*, Nonlinear operators and the Calculus of Variations, Lucien Waelbroeck, ed., Springer-Verlag, New York, 1976.

[9] V. I. ARKIN AND V. L. LEVIN, *Convexity of ranges of vector integrals, measurable choice theorems and variational problems*, Uspehi Mat. Nauk., 27 (1972), no. 3, pp. 21–77.

[10] R. T. ROCKAFELLAR, *Existence theorems for general control problems of Bolza and Lagrange*, Advances in Math., 15 (1975), pp. 312–333.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

ENVELOPE CORRESPONDENCE AND ITS APPLICATION TO THE DESIGN OF GUARANTEED ERROR CONTROLLERS*

B. ROSS BARMISH†

Abstract. In this paper, various problems of minimax error regulation are solved via the method of "envelope correspondence." Using the Fenchel–Rockafellar system of conjugacy correspondence, we define a transformation *e and proceed to transform Problem (P), the original problem, into (P_e^*), the so-called envelope problem. The derivation of (P_e^*) from (P) exploits convexity rather than differentiability of the system error norm. The solvability of (P_e^*) is then considered. To meet this end, we develop a subdifferential description of the dual objective function associated with (P_e^*). When the output error norm is polyhedral, it is shown how one can get a practical hold on (P_e^*), i.e., a linear programming approach becomes feasible for computation of approximate numerical solutions.

1. Introduction. An incompletely or inaccurately identified dynamical system may often be characterized by a vector q of uncertain parameters. Problems of guaranteed performance (G.P.) arise when control is attempted for such systems. Basically, the problem one faces is that of finding a control law u_0 to guarantee some upper bound V_0 for a performance index $J(q, x, u)$. We insist a priori that this bound must hold for all excursions of q within some prespecified uncertainty set Q.

In the literature of control theory, various assumptions (compactness, differentiability, etc.) are made about Q and $J(q, x, u)$ above. A selection from this extensive literature includes the minimax control problems of Witsenhausen [1], Salmon [2] and Wilson [3], the sensitivity approaches of Sobral [4] and Dorato and Kesterbaum [5], the adaptive G.P. controllers of Chang and Peng [6], the norm-uncertain systems of Donati [7] and Negro [8], and the fuzzy set formulation of Chang [9].

In this paper, we propose a new approach to G.P. control which does not depend on the differentiability of the cost functional. Underlying the theoretical developments of §§ 2–5 is the following rather simple geometric notion:

> If a function f is "well-behaved", then we may describe the convex hull of f, conv f,[1] as the upper envelope of the tangent hyperplanes to the graph of f.

This so-called envelope operation was recently used in a convex programming context by White [10]. In [1], Witsenhausen considers G.P. control of sampled systems subjected to an additive disturbance. The envelope operation is used in deriving algorithms dual to dynamic programming. In contrast to [1], we shall consider the envelope operation within the framework of initial state uncertainty. McLinden [11] and Rockafellar [12] have related this operation to Fenchel's theory of conjugate duality [13]. It is this relationship that is instrumental to the

* Received by the editors January 28, 1976, and in revised form January 24, 1977.

† Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520.

[1] conv f is the largest convex function majorized by f.

G.P. control schemes presented here. The plan for the remainder of this paper is as follows:

Section 2. We describe the *envelope correspondence* *e. A typical object in the domain of *e is a family $\{J_q: q \in Q\}$ of real-valued (performance) functionals.

Section 3. We define Problem (P)—a G.P. output regulation problem. Our goal is to minimize the worst-case output error subject to variations in the initial state x_0 within prescribed bounds.

Section 4. Under *e, we obtain an envelope-type dual objective functional. Hence, we consider Problem (P_e^*) in lieu of Problem (P). (P_e^*) is simpler than (P) in the sense that its solution can be found in Euclidean n-space whereas Problem (P) is infinite-dimensional. Furthermore, we can obtain necessary conditions on the solution of (P) from the solution of (P_e^*). It is seen that a strengthening of the results of § 4 is possible when U is a ball (amplitude constraints) and the output norm has a polyhedral structure.

Section 5. The structure of (P_e^*) is examined and an approach is suggested for computation of numerical solutions. Under the strengthened hypothesis of § 4, we can characterize the dual objective as the pointwise maximum of finitely many "preferred" affine functions.

Section 6. A numerical example is used to illustrate the implementation of the computational approach of § 5.

Section 7. Conclusion.

The appendices contain proofs of all results presented herein.

2. Envelope correspondence *e.

(i) *Envelope operation.*[2] Let $F = \{f_i: i \in I\}$ be a nonempty collection of extended real-valued functionals on a normed vector space X. Define the set

$$\Theta(F) = \{g: X \to R^1: g \text{ is affine; } g \leqq \inf f_i \text{ pointwise}\}. \tag{1}$$

Then env $F: X \to R^1 \cup \{\pm\infty\}$ is defined by

$$(\text{env } F)(x) = \sup \{g(x): g \in \Theta(F)\} \tag{2}$$

where the supremum over the empty set is taken as $-\infty$.

(ii) *Fenchel's conjugate.* For f_i as above, we define f_i^*, the *conjugate of* f_i, by

$$f_i^*(x^*) = \sup_{x \in X} [x^*(x) - f_i(x)] \tag{3}$$

where $x^* \in X^*$, the dual of X. The following fundamental lemma relates Fenchel's conjugate in R^n to the envelope operation.

LEMMA (See [12] for proof). *Let $F = \{f_i: i \in I\}$ be a nonempty indexed collection of proper*[3] *lower semicontinuous convex functions on R^n having a*

[2] Geometrically, we are constructing env F from the closure of the convex hull of the epigraph of $f = \inf f_i$. Consequently, env F is the pointwise supremum over all affine functions majorized by f.

[3] f is *proper* if it never assumes the value $-\infty$ and $f \not\equiv +\infty$.

common effective domain. Then

$$(\max_i f_i)^* = \operatorname{env} F^* \tag{4}$$

where $F^* = \{f_i^* : i \in I\}$.

In light of this lemma, we define the action of *e by

$$\{f_i : i \in I\} \overset{*e}{\mapsto} \operatorname{env} F^*.$$

3. Guaranteed error performance. We describe below a linear differential system S having uncertain starting state x_0. For a given command signal y_d, we seek a control u_0 providing the smallest possible worst-case output error $y - y_d$, the worst case taken with respect to *all* possible variations of x_0 within given bounds.

(i) *The given data.*

1. A nonempty *admissible control set* U. We assume that U is a prescribed closed and bounded convex subset of $L_\infty^m[0, T]$—the space of m-dimensional essentially bounded real-valued measurable functions on $[0, T]$.

2. A nonempty compact set $X_0 \subset R^n$. Points $x_0 \in X_0$ are *possible initial states.*

3. A controllable linear differential system S described by

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \\ y(t) &= C(t)x(t) \end{aligned} \tag{5}$$

where $\dim x = n$ and $\dim y = r$. The matrices $A(\cdot)$, $B(\cdot)$ and $C(\cdot)$ are assumed continuous in their arguments.

4. A *desired output trajectory* $y_d(t)$—an r-dimensional, real, measurable function on $[0, T]$.

(ii) *Problem* (P). We let $y(x_0, u, t)$ denote the output at time $t > 0$ induced by the input pair $x_0 \in X_0$, $u \in U$. Also, $\|\cdot\|$ will be a specified norm on R^r. Then the *terminal error* for fixed (x_0, u) is

$$E(x_0, u) = \|y_d(T) - y(x_0, u, T)\|. \tag{6}$$

$$V_0 = \inf_{u \in U} \sup_{x_0 \in X_0} E(x_0, u). \tag{P}$$

(iii) *Special cases.* In many engineering applications, $E(x_0, u)$, U and X_0 are endowed with additional structure. As special cases of the results to follow, we shall consider the assumptions

A1. $E(x_0, u)$ measures the largest weighted component of the terminal error, i.e.,

$$E(x_0, u) = \max_i w_i |y_d^i(T) - y^i(x_0, u, T)|, \qquad w_i \geqq 0.$$

For simplicity, we take all $w_i = 1$ and note that our results can be easily modified to handle $w_i \neq 1$ as well.

A2. The admissible controls are amplitude bounded by $M>0$, i.e.,

$$U=\{(u^1, u^2, \cdots, u^m)\in L_\infty^m[0, T]: |u^i(t)|\leqq M<\infty \text{ a.e. for } i=1, 2, \cdots, m\}.$$

4. The envelope problem (P_e^*).

(i) *Notational preliminaries.* We take $\Phi(t, \tau)$ as the state transition matrix for S; $h_1(\tau), h_2(\tau), \cdots, h_m(\tau)$ will be the columns of $H(\tau)=C(T)\Phi(T, \tau)B(\tau)$; $H^*(\tau)$ will be the transpose of $H(\tau)$; $\|\cdot\|_*$ will denote the dual norm on R^r. (E.g. if $\|y\|=(\sum_{i=1}^r |y^i|^p)^{1/p}$, $1<p<\infty$, then $\|y\|_*=(\sum_{i=1}^r |y^i|^q)^{1/q}$ where $1/p+1/q=1$.) With this notation, the dual unit ball is

$$b^*=\{y^*\in R^r: \|y^*\|_*\leqq 1\}.$$

The indicator on b^* will be $\delta(y^*|b^*)=0$ if $y^*\in b^*$; $\delta(y^*|b^*)=+\infty$ if $y^*\notin b^*$. We also define the set

$$\bar{Y}=\{C(T)\Phi(T, 0)x_0: x_0\in X_0\}.$$

Finally, $h(\cdot|K)$ will denote the support function of a (convex) subset K of a normed vector space X, i.e., if $x^*\in X^*$, then $h(x^*|K)=\sup_{x\in K} x^*(x)$.

(ii) *Problem* (P_e^*). Using the theory of conjugacy correspondence [12], [16], we can describe Problem (P_e^*), a finite-dimensional dualized version of (P). The manner in which (P) and (P_e^*) are related is given below in Theorem 1 and Corollaries 1 and 2.

We seek

$$(P_e^*) \qquad V_0^*=\min\{\bar{F}^*(y^*)+\bar{G}^*(y^*): y^*\in R^r\}$$

where

$$\bar{F}^*=\operatorname{env} F^*, \qquad F^*=\{f^*(\bar{y}, \cdot): \bar{y}\in\bar{Y}\};$$
$$f^*(\bar{y}, y^*)=\langle y_d(T)-\bar{y}, y^*\rangle+\delta(y^*|b^*);$$
$$\bar{G}^*(y^*)=h(-H^*(\cdot)y^*|U).$$

THEOREM 1 (see Appendix A for proof) $V_0+V_0^*=0$. *Furthermore, an optimal element* $y_0^*\in R^r$ *solving* (P_e^*) *always exists and any* $u_0(\cdot)\in U$ *solving* (P) *satisfies the necessary condition*

$$-h(-H^*(\cdot)y_0^*|U)=\int_0^T \langle u_0(\tau), H^*(\tau)y_0^*\rangle\, d\tau. \tag{7}$$

COROLLARY 1. (See Appendix A for proof). *Under* A1, *there is a finite subset* $\bar{Y}_p=\{\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_p\}$ *of* $\bar{Y}$ *such that Theorem* 1 *holds with* $\bar{Y}_p$ *replacing* $\bar{Y}$; $p\leqq 2r$.

COROLLARY 2. (See Appendix A for proof). *Under* A2, *the necessary condition in* (7) *is equivalent to*

$$\begin{aligned} &u_0^i(t)=M \quad \text{if } \langle y_0^*, h_i(t)\rangle<0;\\ &u_0^i(t)=-M \quad \text{if } \langle y_0^*, h_i(t)\rangle>0;\\ &u_0^i(t)\in[-M, M] \quad \text{if } \langle y_0^*, h_i(t)\rangle=0. \end{aligned} \tag{8}$$

(iii) *Remarks.* The points $\bar{y}_i$ above can be constructed from an extremal basis for (P), i.e., there are points $\bar{x}_{01}, \bar{x}_{02}, \cdots, \bar{x}_{0p}$ in X_0 such that

$$V_0 = \inf_{u \in U} \max_i E(\bar{x}_{0i}, u).$$

Then, we may take $\bar{y}_i = y_d(T) - C(T)\Phi(T, 0)\bar{x}_{0i}$ (see Appendix A).

Corollary 1 also holds as is if A1 is replaced by

A1′. X_0 is a convex polytope.

We note that (8) is simply an alignment requirement between the Banach spaces $L_\infty^m[0, T]$ and $L_1^m[0, T]$.

(iv) *Trajectory-weighted problems.* The results developed in this paper can be easily modified to handle the *trajectory-weighted version* of Problem (P), i.e., we seek

$$(\mathrm{P_t}) \qquad V_{0t} = \inf_{u \in U} \sup_{x_0 \in X_0} E_t(x_0, u)$$

where

$$E_t(x_0, u) = \max_j \| y_d(t_j) - y(x_0, u, t_j) \|$$

and $0 < t_1 < t_2 < \cdots < t_N = T$ is a set of sample times. Now, it can be shown (using an argument similar to that in Appendix A) that Theorem 1 remains valid subject to the following substitutions:

Replace V_0 by V_{0t};
Replace V_0^* by V_{0t}^*;
Replace $(\mathrm{P_e^*})$ by $(\mathrm{P_{et}^*})$;
Replace (P) by $(\mathrm{P_t})$;
Replace R^r by R^{Nr};

Replace $\bar{Y}$ by $\left\{ \begin{bmatrix} C(t_1)\Phi(t_1, 0)x_0 \\ C(t_2)\Phi(t_2, 0)x_0 \\ \vdots \\ C(t_N)\Phi(t_N, 0)x_0 \end{bmatrix} : x_0 \in X_0 \right\}$;

Replace $y_d(T)$ by $\begin{bmatrix} y_d(t_1) \\ y_d(t_2) \\ \vdots \\ y_d(t_N) \end{bmatrix}$;

Replace $H(\tau)$ by $\begin{bmatrix} C(t_1)\Phi(t_1, \tau)B(\tau)I(\tau|[0, t_1]) \\ C(t_2)\Phi(t_2, \tau)B(\tau)I(\tau|[0, t_2]) \\ \vdots \qquad\qquad \vdots \\ C(t_N)\Phi(t_N, \tau)B(\tau)I(\tau|[0, t_N]) \end{bmatrix}$

where

$$I(\tau|[0, t_i]) = 1 \quad \text{if } \tau \in [0, t_i],$$
$$= 0 \quad \text{otherwise.}$$

5. The structure of ($\mathbf{P}_e^*$). In this section, we investigate

1. the dependence of $\bar{F}^* = \text{env } F^*$ on its argument $y^* \in R^r$;
2. the structure of the nonlinear convex function $\bar{G}^*(y^*)$.

(i) *The envelope.* In general, $\bar{F}^*(y^*)$ is given by (cf. Corollary 17.1.3 of [12] and Theorem 1 above)

$$\bar{F}^*(y^*) = \min \sum_i \lambda_i \langle y_d(T) - \bar{y}_i, y_i^* \rangle, \qquad \bar{y}_i \in \bar{Y},$$

where the minimum is taken over all expressions of y^* as a convex combination of vectors $y_i^* \in b^*$. Furthermore, we can restrict attention to those convex combinations in which at most $(n+1)$ of the λ_i are nonzero.

Such a scheme, however, has one obvious drawback: To evaluate $\bar{F}^*$ at a single point, we must solve an entire optimization problem involving, perhaps, many parameters. With the inclusion of A1, however, we shall obtain a closed form expression for $\bar{F}^*(y^*)$.

(ii) *Algorithm for finding* $\bar{F}^*(y^*)$. We first define the polytope E in R^{r+1} as the convex hull of all points of the form $e \oplus \min_i \langle \bar{y}_i, e \rangle$ where e is a vertex of b^* and $\bar{y}_i \in \bar{Y}_p$ (of Corollary 1). Now, $\bar{E}$ is said to be a *preferred subset* of E if

P1. aff $\bar{E}$, the smallest linear manifold containing $\bar{E}$, is a nonvertical hyperplane.

P2. The affine linear function $\bar{F}$ on R^r generated by aff $\bar{E}$ contains E in its epigraph, i.e.,

$$E \subseteq \{y^* \oplus z : z \geqq \bar{F}(y^*)\}. \tag{9}$$

The class $\{F_j^* : j \in J\}$ is now taken to be those (*preferred*) affine functions generated by the preferred subsets of E. Then we have

(iii) THEOREM 2 (See Appendix B for proof). *Under* A1,

$$\bar{F}^*(y^*) = \max_j F_j^*(y^*) + \delta(y^*|b^*). \tag{10}$$

(iv) *Remarks.* The fact that $\bar{F}^*$ turns out to be polyhedral when b^* is a polytope (A1) is not surprising in light of Theorem 19.2 of [12]. For the purpose of numerical computation, we require an algorithm telling us *how* to generate $\bar{F}^*$. Consequently, Theorem 2 cannot be proven simply by invoking Theorem 19.2 of [12].

(v) *The nonlinear convex function* $\bar{G}^*(y^*)$. Without assumption A2 (or some other assumption about the structure of U), $\bar{G}^*(y^*)$ is found from the formula for the support function (See Theorem 1)

$$\bar{G}^*(y^*) = -\min_{u \in U} \int_0^T \langle u(\tau), H^*(\tau) y^* \rangle \, d\tau.$$

Under A2, more can be said, i.e., if we substitute for U in the expression for $\bar{G}^*(y^*)$ above, we obtain

$$\bar{G}^*(y^*) = M \int_0^T \sum_{i=1}^m |\langle y^*, h_i(t)\rangle| \, dt. \tag{11}$$

From the point of view of computing numerical solutions to (P_e^*), the following lemma will be useful.

LEMMA 1. (See Appendix C for proof). *Let $\bar{y}^* \in R^r$. Then the subdifferential of $\bar{G}^*$ at $\bar{y}^*$, denoted $\partial\bar{G}^*(\bar{y}^*)$, contains all vectors of the form*

$$s(\bar{y}^*) = M \int_0^T \sum_{i=1}^m h_i(t)\sigma_i(t) \, dt \tag{12}$$

where each $\sigma_i(\cdot)$ is a measurable function satisfying

$$\begin{aligned} \sigma_i(t) &= 1 \quad \text{if } \langle \bar{y}^*, h_i(t)\rangle > 0; \\ \sigma_i(t) &= -1 \quad \text{if } \langle \bar{y}^*, h_i(t)\rangle < 0; \\ -1 \leqq \sigma_i(t) &= 1 \quad \text{if } \langle \bar{y}^*, h_i(t)\rangle = 0. \end{aligned}$$

(vi) *Remarks.* A vector $s(\bar{y}^*)$ in $\partial\bar{G}^*(\bar{y}^*)$ can be used to generate a hyperplane which supports the epigraph of $\bar{G}^*$ at the point $\bar{y}^*$, i.e., the affine linear function (of y^*)

$$\bar{G}^*(\bar{y}^*, y^*) = \langle s(\bar{y}^*), y^*\rangle + [\bar{G}^*(\bar{y}^*) - \langle s(\bar{y}^*), \bar{y}^*\rangle] \tag{13}$$

is majorized by $\bar{G}^*$ and agrees with $\bar{G}^*$ at $\bar{y}^*$. Using this fact we can "approximate" $\bar{G}^*$ (from below) as the pointwise maximum of a finite collection of affine linear functions, i.e., if $s(\bar{y}_i^*) \in \partial\bar{G}^*(\bar{y}_i^*)$ for $i = 1, 2, \cdots, k$, then we might approximate $\bar{G}^*(y^*)$ by

$$\max\{\bar{G}^*(\bar{y}_i^*, y^*): i = 1, 2, \cdots, k\}.$$

6. Numerical solutions. In light of the preceeding remarks, we propose the following "heuristic algorithm" for computation of "candidate" G.P. controllers (under A1, A2).

Step 0 (Initialization). Construct the conjugate functions f_i^* using the points $\bar{y}_i$ of Corollary 1. Then use the vertices of the polytope E to generate the preferred affine functions F_j^* in accordance with P1 and P2. Replace $\bar{G}^*(y^*)$ in (P_e^*) by the *zeroeth approximation* $\bar{G}_0^*(y^*) \equiv 0$. Also, replace $\bar{F}^*$ in (P_e^*) by the right hand side of (10). Set $n = 0$.

Step 1. Solve the following linear programming problem for

$$V_n^* = \min\{\bar{F}^*(y^*) + \bar{G}_n^*(y^*): y^* \in R^r\}. \tag{LP$_n$}$$

Let y_n^* denote one solution to (LP$_n$).

Step 2. Compute $\varepsilon_n = \bar{G}^*(y_n^*) - \bar{G}_n^*(y_n^*)$. (Note: ε_n is an upper bound for $V_0^* - V_n^*$.) If ε_n is "sufficiently small," proceed to Step 5. Otherwise,

Step 3. Compute $s(y_n^*)$, a vector in $\partial\bar{G}^*(y_n^*)$, using the formula in Lemma 1.

Step 4. Let

$$\bar{G}^*_{n+1}(y^*) = \max\{\bar{G}^*_n(y^*), \bar{G}^*(y^*_n, y^*)\}$$

where $\bar{G}(y^*_n, \cdot)$ is given by (13). Replace n by $n+1$ and return to Step 1.

Step 5 (Termination). Use the "approximate solution" y^*_n to generate a candidate G.P. controller $u_n(\cdot)$ via (8). (Note that $u_n(\cdot)$ will be feasible (in U) despite the fact that y^*_n is not necessarily the exact solution of (P^*_e).)

(ii) *Remarks.* Step 4 defines a progressive rule for generating a well defined sequence of linear programs. Arguing as in [18], it can be shown that any limit point of the sequence $\langle y^*_n\rangle_{n=1}^{\infty}$ is a solution to (P^*_e). We call the algorithm "heuristic", however, because we can provide no guarantee that the corresponding sequence of controls $\langle u_n(\cdot)\rangle_{n=1}^{\infty}$ will converge to a solution of (P).

(iii) *Unconstrained version of* (P). We may also wish to consider the case $U = L^m_\infty[0, T]$. Now, we seek

$$(P') \qquad V'_0 = \inf_{u \in L^m_\infty} \sup_{x_0 \in X_0} E'(x_0, u)$$

where

$$(14) \qquad E'(x_0, u) = \|u\| + E(x_0, u).$$

Using an argument similar to that in the proof of Theorem 1, it can be shown that $V'_0 + V^{*\prime}_0 = 0$ where

$$(P^{*\prime}_e) \qquad V^{*\prime}_0 = \min\left\{\bar{F}^*(y^*): \int_0^T \sum_{i=1}^m |\langle y^*, h_i(t)\rangle|\, dt \leqq 1\right\}.$$

Furthermore, if $\bar{y}^*$ is a boundary point of the (convex) "dual constraint region"

$$\left\{y^* \in R^r: \int_0^T \sum_{i=1}^m |\langle y^*, h_i(t)\rangle|\, dt \leqq 1\right\},$$

then Lemma 1 can be re-interpreted (with $M = 1$) as being a description of vectors in the normal cone of this region at $\bar{y}^*$. Consequently, we can develop an algorithm (analogous to the one above) to generate candidate solutions to (P').

(iv) *An example.* We consider an unstable, 3-dimensional plant S described by

$$\begin{aligned}
\dot{x}_1(t) &= x_2(t) + u_1(t),\\
\dot{x}_2(t) &= x_3(t) + u_2(t),\\
\dot{x}_3(t) &= 48x_1(t) - 16x_2(t) + 3x_3(t) + u_1(t),\\
y(t) &= x_1(t) + x_2(t) + x_3(t), \qquad t \in [0, 0.75].
\end{aligned}$$

The set X_0 of possible initial states is the tetrahedron having vertices (0, 0, 0), (0.1, 0, 0), (0, 1, 0) and (0, 0, 1), and the desired output $y_d(t)$ is the triangle wave of Fig. 1. Terminal errors are measured at the sample times $t_{1,2,3} = 0.25, 0.50, 0.75$ and we regulate the quantity

$$E(x_0, u) = \max_{i,j} |y^i(x_0, u, t_j) - y^i_d(t_j)|.$$

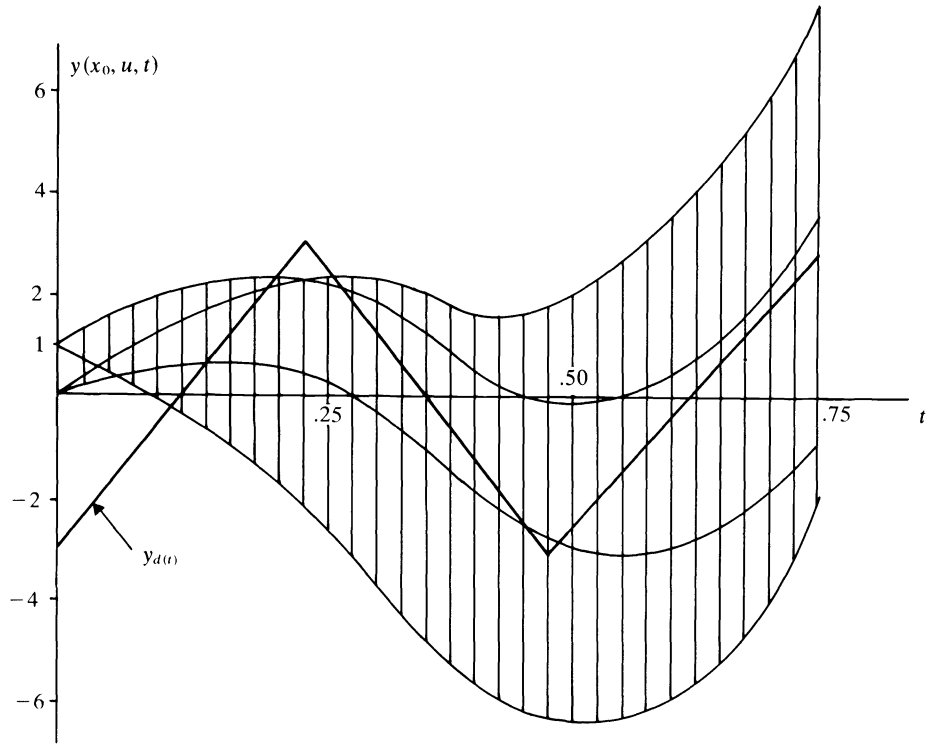


FIG. 1. *Possible output trajectories subject to G.P. control*

We also include a term penalizing control effort as in (14). The conjugate functions are computed to be

$$f_1^*(y^*) = 1.14y^{1*} - 7.56y^{2*} - 5.70y^{3*} + \delta(y^*|b^*),$$
$$f_2^*(y^*) = 5.62y^{1*} + 0.82y^{2*} + 4.52y^{3*} + \delta(y^*|b^*).$$

The preferred affine functions are

$$F_1^*(y^*) = 6.25y^{1*} - 2.45y^{2*} - 0.59y^{3*} - 5.11,$$
$$F_2^*(y^*) = 6.25y^{1*} - 4.29^{2*} - 0.59y^{3*} - 4.19,$$
$$F_3^*(y^*) = 0.51y^{1*} - 2.45y^{2*} - 0.59y^{3*} - 5.11,$$
$$F_4^*(y^*) = 0.51y^{1*} - 4.29y^{2*} - 0.59y^{3*} - 5.11.$$

Using 9 supporting hyperplanes to approximate the dual constraint region, we obtain an "approximate" solution to Problem ($P_e^{*\prime}$):

$$y_9^* = (-0.51, 0.38, -0.11),$$
$$V_9^{*\prime} = 6.24.$$

Using (8) of Corollary 2, we generate the candidate control

$$u_9^1(t) = 1 \quad \text{if } t \in [0.0, 0.1] \cup [0.38, 0.75],$$
$$= -1 \quad \text{otherwise.}$$
$$u_9^2(t) = 1 \quad \text{if } t \in [0.0, 0.38] \cup [0.68, 0.75],$$
$$= -1 \quad \text{otherwise.}$$

In Fig. 1, the cross-hatched region represents the set of possible output trajectories $y(x_0, u_9, t)$ that arise as the initial state x_0 varies over X_0.

7. Conclusion. Theorem 1 presupposes the existence of a G.P. controller $u_0 \in U$. In [14], a weak-star lower semicontinuity argument is used to guarantee the existence of u_0 above. The situation may arise wherein the quantity $\langle y_0^*, h_i(t) \rangle$ is zero on a time set of nonzero measure. In such a case, a G.P. controller is nonunique and we may consider the inclusion of a secondary performance criterion. Finally, we mention two G.P. control problems currently being investigated via the method of envelope correspondence. Under *e, we hope to obtain *new* envelope problems analogous to (P_e^*).

(i) *Optimal control subject to multiple or conflicting objectives.* In this situation we consider a *set* Y_d of desired paths for the differential system S. $J(y_d; x; u)$ is a performance index for S and we seek the infimum

$$V_0 = \inf_{u \in U} \sup_{y_d \in Y_d} J(y_d; x; u).$$

(ii) *Inaccurately modeled dynamic systems.* $M(u)$ is the response of a model M of S to an input u. $S(u)$, the output of S, is known only within prescribed bounds, i.e., it is known a priori that the system-model error

$$e(u) = S(u) - M(u)$$

is bounded in norm by some constant β. The G.P. control problem is that of finding the best performance level $J(e(u); u)$ that can be guaranteed independently of possible error excursions between the system and the model, i.e., find

$$V_0 = \inf_{u \in U} \sup_{e} J(e(u); u).$$

Negro [8] and others have recently considered a class of such problems in a Hilbert space setting.

Appendix A. Proof of Theorem 1, Corollaries 1 and 2. First, we define the function $f: R^r \to R$ by

$$f(y) = \sup_{\bar{y} \in \bar{Y}} f(\bar{y}, y)$$

where $f(\bar{y}, \cdot): R^r \to R$ is given by

$$f(\bar{y}, y) = \|y_d(T) - \bar{y} - y\|.$$

Step 1. *Convexity of* $f(\cdot)$. Clearly, $f(\bar{y}, \cdot)$ can be expressed as the composition of $E_2(y) = \|y\|$ and $E_1(y) = y_d - \bar{y} - y$, i.e., $f(\bar{y}, y) = (E_2 \circ E_1)(y)$. Since $E_1(\cdot)$ is

affine linear and $E_2(\cdot)$ is convex, it follows that $f(\bar{y}, y)$ is convex. Now, we may view $f(\cdot)$ as the pointwise supremum over the family $\{f(\bar{y}, \cdot): \bar{y} \in \bar{Y}\}$ of convex functions. Hence $f(\cdot)$ must be convex (cf. [12, Thm. 5.5]).

Step 2. *Application of Rockafellar's duality theorem.* We are going to put Problem (P) into the standard form required for applying Rockafellar's extension of Fenchel's duality theorem (cf. Theorem 3 of [16]). To meet this end, we define the linear operator $\Lambda: L_m^\infty[0, T] \to R^r$ by

$$\Lambda u = \int_0^T C(T)\Phi(T, \tau)B(\tau)u(\tau)\, d\tau;$$

and the concave function $g: L_m^\infty[0, T] \to R$ by

$$g(u) = -\delta(u|U).$$

Now, it is easily shown that

$$\text{(P)} \qquad V_0 = \inf\{f(\Lambda u) - g(u): u \in L_m^\infty[0, T]\}.$$

Applying Theorem 3 of [16], we obtain[4]

$$\text{(P*)} \qquad V_0 = \max\{-g^*(-\Lambda^* y^*) - f^*(y^*): y^* \in R^r\}.$$

Step 3. *We show that* $f^*(y^*) = \bar{F}^*(y^*)$. By definition,

$$f^*(y^*) = \sup_y [\langle y, y^*\rangle - \sup_{\bar{y} \in \bar{Y}} f(\bar{y}, y)]$$

$$= \sup_{\bar{y} \in \bar{Y}} (f(\bar{y}, \cdot))^*(y^*).$$

Now, by Lemma 1, it follows that

$$f^*(y^*) = (\operatorname{env} F^*)(y^*)$$

where $F^* = \{f^*(\bar{y}, \cdot): \bar{y} \in \bar{Y}\}$. The calculation

$$f(\bar{y}, y^*) = \langle y_d(T) - \bar{y}, y^*\rangle + \delta(y^*|b^*)$$

is straightforward.

Step 4. *Completion of proof.* It is easily verified that

$$g^*(-\Lambda^* y^*) = -h(-H^*(\cdot)y^*|U)$$

$$= -\bar{G}^*(y^*).$$

Substituting for f^* and g^* in (P*) yields

$$V_0 = \max\{-\bar{F}^*(y^*) - \bar{G}^*(y^*): y^* \in R^r\}$$

$$= -V_0^*.$$

Solutions to (P) and (P_e^*) must satisfy the so-called "extremality conditions" (cf. [16, p. 184]). Hence, we require $-\Lambda^* y_0^* \in \partial g(u_0)$. This says that $-\Lambda^* y_0^*$ must

[4] It is easily shown (using Theorem 1 of [16]) that (P) is "stably set",—a technical precondition for Theorem 3.

be in the normal cone of U at u_0, i.e.,

$$\langle -\Lambda^* y_0^*, u_0\rangle = \sup_{u\in U}\langle -\Lambda^* y_0^*, u\rangle.$$

Clearly, this condition is the same as (7).

Proof of Corollary 1. We construct a set of points $\{\bar{x}_{01}, \bar{x}_{02}, \cdots, \bar{x}_{0p}\}\subseteq R^n$ such that the equality

$$\sup_{x_0\in X_0} E(x_0, u) = \max_i E(\bar{x}_{0i}, u)$$

holds for all $u\in U$ and $p\leqq 2r$. Define $P_j: R^r \to R^1$ to be the projection map onto the jth coordinate, i.e., $P_j y = y^j$. Now for fixed j, it follows from the continuity of P_j and the compactness of $\bar{Y}$ that there are (at least) two points $\bar{y}_j$, $\underline{y}_j$ in $\bar{Y}$ such that

$$\inf_{y\in\bar{Y}} P_j y = P_j \underline{y}_j, \qquad \sup_{y\in\bar{Y}} P_j \underline{y}_j = P_j \bar{y}_j.$$

Pick any two points $\bar{x}_j$, $\underline{x}_j \in R^n$ such that $C(T)\Phi(T, 0)\bar{x}_j = \bar{y}_j$, $C(T)\Phi(T, 0)\underline{x}_j = \underline{y}_j$. Letting j vary from 1 to r, we obtain at most $2r$ distinct points $\underline{x}_1$, $\bar{x}_1$, $\underline{x}_2$, $\bar{x}_2, \cdots, \underline{x}_r, \bar{x}_r$ in this manner. Relabel these points $\bar{x}_{01}, \bar{x}_{02}, \cdots, \bar{x}_{0p}$. Now we fix any $x_0\in X_0$, $u\in U$ and observe that for some $k\leqq r$,

$$\begin{aligned} E(x_0, u) &= |P_k(y_d(T) - C(T)\Phi(T, 0)x_0 - \Lambda u)| \\ &\leqq \max\{E(\bar{x}_k, u), E(\underline{x}_k, u)\} \\ &\leqq \max_i E(\bar{x}_{0i}, u). \end{aligned}$$

The inequality

$$\sup_{x_0\in X_0} E(x_0, u) \leqq \max_i E(\bar{x}_{0i}, u) \tag{A.1}$$

now follows by taking the supremum over $x_0\in X_0$. Furthermore, (A.1) can be reversed because $\bar{x}_{0i}\in X_0$.

Proof of Corollary 2. When U is a closed ball of radius M in $L_m^\infty[0, T]$, we have

$$h(-H^*(\cdot)y^*|U) = M\|H^*(\cdot)y^*\|_*$$

where $\|\cdot\|_*$ is the norm on $L_m^1[0, T]$. Substituting (7) above, we obtain the necessary condition

$$\int_0^T \sum_{i=1}^m u_0^i(\tau)\langle y_0^*, h_i(\tau)\rangle\, d\tau = M\int_0^T \sum_{i=1}^m |\langle y_0^*, h_i(\tau)\rangle|\, d\tau$$

from which (8) easily follows.

Appendix B. Proof of Theorem 2.

Preliminaries. The point $y^*\oplus z^*$ in R^{r+1} is called a *lower point of the polytope* E if

$$z^* = \inf\{z: y^*\oplus z\in E\}. \tag{B.1}$$

We denote the set of lower points of E by L and call conv $\bar{E}$ a *lower r-face of* E when $\bar{E}$ is a preferred subset of E.

Proofs of the following "geometrically obvious" facts are left to the reader.

Fact 1. L is the nonempty union of the lower r-faces of E.

Fact 2. L shadows b^* in the following sense: For every $y^* \in b^*$, there is some $z^* \in R^1$ such that $y^* \oplus z^* \in L$.

We also require one lemma. (See Witsenhausen [1] for proof.)

LEMMA. *Suppose* $f: R^r \to R^1 \cup \{\pm\infty\}$ *is bounded from below and* $f(x) = +\infty$ *outside a convex set* K. *Then* $(\operatorname{env} f)(x) = +\infty$ *for* x *outside the closure of* K.

Proof of Theorem. For y^* outside b^*, we have

$$\min_i f_i^*(y^*) = +\infty.$$

The preceding lemma necessitates

$$(\operatorname{env} F^*)(y^*) = +\infty$$

outside b^*. Hence we have the term $\delta(y^* | b^*)$ in (14).

For $y^* \in b^*$, we first establish the inequality

$$(\operatorname{env} F^*)(y^*) \leqq \max_j F_j^*(y^*). \tag{B.2}$$

To do this, pick any $\theta \in \Theta(F^*)$ and fix $y^* \in b^*$. By Fact 2, we may find $z^* \in R^1$ such that $y^* \oplus z^* \in L$. Using Fact 1, we may now select some lower r-face $\operatorname{conv} \bar{E}_k$ containing $y^* \oplus z^*$. Let F_k^* denote the preferred affine function generated by $\bar{E}_k$. We now proceed to show that

$$\theta(y^*) \leqq F_k^*(y^*). \tag{B.3}$$

First we expand $y^* \oplus z^*$ as a convex combination of the vertices of $\bar{E}_k$, i.e.,

$$y^* \oplus z^* = \sum_i \alpha_i (\bar{e}_i \oplus \min_j \langle \bar{y}_j, \bar{e}_i \rangle), \qquad \alpha_i \geqq 0, \qquad \sum_i \alpha_i = 1$$

$$= \sum_i \alpha_i (\bar{e}_i \oplus F_k^*(\bar{e}_i)).$$

By the nonnegativity of the α_i and the affine linearity of θ and F_k^*, it follows that

$$\theta(y^*) = \theta\Big(\sum_i \alpha_i \bar{e}_i\Big) = \sum_i \alpha_i \theta(\bar{e}_i)$$

$$\leqq \sum_i \alpha_i \min_j \langle \bar{y}_j, \bar{e}_i \rangle$$

$$= F_k^*\Big(\sum_i \alpha_i \bar{e}_i\Big) = F_k^*(y^*).$$

Consequently,

$$\theta(y^*) \leqq \max_j F_j^*(y^*)$$

Inequality (B.2) now follows by taking the supremum over $\theta \in \Theta(F^*)$.

To establish the reverse inequality

$$\max_j F_j^*(y^*) \leqq (\operatorname{env} F^*)(y^*), \tag{B.4}$$

we first show that

$$\max_j F_j^*(y^*) \leqq \min_j f_j^*(y^*). \tag{B.5}$$

To do this, we exploit the convexity of $\max_j F_j^*$, i.e.,

$$\max_j F_j^*(y^*) \leqq \sum_i \alpha_i \max_j F_j^*(\bar{e}_i). \tag{B.6}$$

The condition $E \subseteq$ epigraph F_j^* for all j then requires

$$\sum_i \alpha_i \max_j F_j^*(\bar{e}_i) \leqq \sum_i \alpha_i \min_j f_j^*(\bar{e}_i). \tag{B.7}$$

Now, by the concavity of $\min_j f_j^*$, we have

$$\sum_i \alpha_i \min_j f_j^*(\bar{e}_i) \leqq \min_j f_j^*(y^*). \tag{B.8}$$

Inequality (B.5) now follows from the chain of inequalities (B.6)–(B.8). Inequality (B.4) is now a trivial consequence of (B.5).

Appendix C. Proof of Lemma 1. Suppose $\bar{y}^* \in R^r$ and $s(\bar{y}^*)$ has the structure given by (12). Then, we must show that

$$\bar{G}^*(y^*) \geqq \bar{G}^*(\bar{y}^*) + \langle y^* - \bar{y}^*, s(\bar{y}^*)\rangle \quad \text{for } \textit{all } y^* \in R^r.$$

For arbitrary $y^* \in R^r$, we have

$$\begin{aligned}
\bar{G}^*(y^*) &= \int_0^T \sum_{i=1}^m |\langle y^*, h_i(t)\rangle| \, dt \\
&\geqq \int_0^T \sum_{i=1}^m \langle y^*, h_i(t)\rangle \sigma_i(t) \, dt \\
&= \int_0^T \sum_{i=1}^m \langle y^* - \bar{y}^*, h_i(t)\rangle \sigma_i(t) \, dt + \bar{G}^*(\bar{y}^*) \\
&= \bar{G}^*(\bar{y}^*) + \int_0^T \langle y^* - \bar{y}^*, \sum_{i=1}^m h_i(t)\sigma_i(t)\rangle \, dt \\
&= \bar{G}^*(\bar{y}^*) + \langle y^* - \bar{y}^*, s(\bar{y}^*)\rangle.
\end{aligned}$$

Acknowledgment. The author is grateful to Professor James S. Thorp for a number of helpful consultations held during the course of research. He also wishes to express his thanks to the reviewers whose suggestions were incorporated into the revised manuscript.

REFERENCES

[1] H. S. WITSENHAUSEN, *Min-max control for sampled linear systems*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 5–21.

[2] D. SALMON, *Minimax Controller Design*, Ibid., AC-13 (1968), pp. 369–375.

[3] D. J. WILSON, *Min-max control of quadratic systems*, Proc. IEEE Conf. on Decision and Control, Phoenix, AZ, 1974.

[4] M. SOBRAL, *Sensitivity in optimal control systems*, Proc. IEEE 56, Institute of Electrical and Electronics Engineers, New York, 1968, pp. 1644–1652.
[5] P. DORATO AND A. KESTERBAUM, *Application of game theory to sensitivity design of optimal systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 85–87.
[6] S. S. L. CHANG AND T. K. C. PENG, *Adaptive guaranteed cost control of systems with uncertain parameters*, Ibid., AC-17 (1972), pp. 474–483.
[7] F. DONATI, *Approssimazione di sistemi lineari in spazi normati*, Proc. XI Inter. Automation and Instrumentation Conf., FAST, Milan, 1970.
[8] A. NEGRO, *Min-max optimal control of systems approximated by finite-dimensional models*, J. Optimization Theory and Appl., 12 (1973), pp. 182–203.
[9] S. S. L. CHANG, *Control and estimation of fuzzy systems*, Proc. IEEE Conf. on Decision and Control, Phoenix, AZ, 1974.
[10] D. J. WHITE, *Envelope programming and a minimax theorem*, J. Math. Anal. Appl., 40 (1972), pp. 1–11.
[11] L. MCLINDEN, *Envelope programming and conjugate duality*, Ibid., 47 (1974), pp. 256–268.
[12] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.
[13] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[14] B. R. BARMISH, *On a class of perturbation-invariant Chebyshev regulators*, Ph.D. dissertation, Cornell Univ., Ithaca, NY, 1975.
[15] ———, *Guaranteed error performance for uncertain linear systems*, Proc. Conf. Systems and Information Sciences, Johns Hopkins Univ., Baltimore, 1975.
[16] R. T. ROCKAFELLAR, *Duality and stability in extremum problems involving convex functions*, Pacific J. Math., 21 (1967), pp. 167–187.
[17] B. R. BARMISH AND A. BRANDES, *Transformed control problems over adjoint constraint regions*, in preparation.
[18] A. F. VEINOTT, *The supporting hyperplane method for unimodal programming*, J. Operations Res., 15 (1967), pp. 147–152.
[19] W. FENCHEL, *On conjugate convex functions*, Canad. J. Math., 1 (1949), pp. 73–77.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

OPTIMAL CONTROL OF GENERATING POLICIES IN A POWER SYSTEM GOVERNED BY A SECOND ORDER HYPERBOLIC PARTIAL DIFFERENTIAL EQUATION*

N. U. AHMED†

Abstract. In this paper we present necessary and sufficient conditions for determining the optimum generating policies for each of a system of power stations feeding into a primary transmission line. The system is modeled as a linear second order partial differential equation of hyperbolic type with appropriate initial and boundary conditions. Further, existence of optimal control policies for the system is proved. A computational algorithm for determining the optimal policies is also presented.

1. Introduction. In this paper we consider the problem of optimal control of generating policies of power stations feeding into a primary transmission line. For economic operation of a power system it is necessary to adjust the generating levels of each of the power sources (subject to the limitation imposed by its maximum generating capacity) with the change of the distribution of load with time. In § 2 we present the model for a power system consisting of a transmission line, two or more generating stations and load distributed in space and time. Assuming the generating stations are located at the terminals of the line, we develop a model consisting of a second order partial differential equation of hyperbolic type with Dirichlet boundary conditions. In § 3 we present certain fundamental results that assure the existence of solution to the Dirichlet (first boundary) problems. In § 4 necessary and sufficient conditions of optimality are presented for power systems with two generating stations located at the terminals of the line (Theorems 4.1, 4.3) and also for power systems having generating stations located at arbitrary points on the line including those at the terminals (Theorem 4.7, Corollary 4.8). In § 5 we present a theorem on the existence of optimal control policies, and in § 6 a computational algorithm, based on the necessary conditions of optimality of § 4, is given.

In the knowledge of the author there appears to be no known results on power systems optimization employing partial differential equations as the systems model. Static optimization techniques and classical variational methods have been widely used for power systems optimization without taking into account the dynamic constraints of the transmission line (Dommel [1, (1971)]; see also the references thereof). The idea of "transposition" used to prove the existence of solutions of nonhomogeneous boundary value problems (§ 3) is taken from Lions and Magenes [4, vol. 1, p. 283] and Lions [2, p. 291]. Necessary conditions of optimality for systems governed by hyperbolic partial differential equations of the type $\partial^2 y/\partial t^2 + Ay = g$, A elliptic, with Dirichlet or Neumann boundary controls are available in the literature (Lions [2, Chap. 4, p. 272 ff]). However these results are not directly applicable to the problems considered in this paper since the damping term $(\partial y/\partial t)$ is missing from the above model and also because the function g, in some of the results given in our paper, is actually a distribution in the

* Received by the editors October 26, 1976, and in revised form February 7, 1977.

† Department of Electrical Engineering, University of Ottawa, Ottawa, Ontario, Canada K1N 6N5.

sense of Schwartz. Russell [5], [6], [7] has considered optimal control of systems governed by a system of hyperbolic equations in symmetric form. In [5] Russell obtains the optimal control in feedback form; however, he does not impose any control constraint. Further, Russell's models do not contain distributional forcing terms.

2. Modeling and formulation of the optimal control problem. A transmission line at power frequencies is described by a pair of first order hyperbolic equations relating currents (i) and voltages (e)

$$i_x = -Ge - Ce_t, \qquad e_x = -Ri - Li_t, \tag{2.1}$$

where i_x, e_x and i_t, e_t are first partials of i, e with respect to space and time variables. Equation (2.1) represents a line without load. A loaded line is represented by

$$i_x = -Ge - Ce_t - f, \qquad e_x = -Ri - Li_t, \tag{2.2}$$

where $f(t, x)$ is the distribution of load current (in time and space) so that the total current drawn by the load from the entire line at any time t is given by

$$I(t) = \int_\Omega f(t, \xi)\, d\xi, \quad \Omega = (0, L_0), \quad L_0 = \text{length of the line.} \tag{2.3}$$

The system (2.2) can be reduced to a single second order hyperbolic equation involving only current distribution

$$i_{tt} + \alpha i_t + \beta i - \gamma i_{xx} = \gamma f_x \tag{2.4}$$

where

$$\alpha = (LG + RC)/(LC), \quad \beta = RG/(LC), \quad \gamma = 1/(LC).$$

and L, C, R, G are the fundamental line parameters (series inductance, shunt capacitance, series resistance and shunt conductance per unit length). Since the equation (2.4) is second order in t it is required to specify the initial distribution of current and its time rate of change before one can consider its solution. In general we have a Cauchy problem:

$$\left.\begin{aligned} &i_{tt} + \alpha i_t + \beta i - \gamma i_{xx} = \gamma f_x, \\ &i(0, x) = i_0(x) \\ &i_t(0, x) = i_1(x) \end{aligned}\right\} \text{given for } x \in \Omega = (0, L_0), \tag{2.5}$$

We consider two different systems, one in which the generating stations are located at the terminals of the line and the other with generating stations located at arbitrary points of the line.

Case (i). We consider that the transmission line is supplied with power from two generating stations located at the terminals. Further we will assume that they are able to generate power at different rates up to a maximum limited by the capacity of the plants. For a given generating voltage $e_0(t)$, the power (active and reactive combined) supplied by the sources is determined by the current supply. Under these assumptions the Cauchy problem (2.5) is converted into a first

boundary value problem called the Dirichlet problem with boundary conditions

$$i(t, 0) = u_1(t), \qquad i(t, L_0) = u_2(t),$$

where $u_1(t)$ and $u_2(t)$ are the currents supplied by the sources located at the terminals at time t. Let L_0 be the length of the line and $\Omega = (0, L_0)$ its span and $I = (0, T)$ the time interval of interest and $Q = I \times \Omega$. Let $\partial\Omega = \{0, L_0\}$ denote the endpoints of the line and $\Sigma = I \times \partial\Omega$. Let $u = (u_1, u_2)$ denote the control vector representing the generating policies of the two sources during the time interval I. With these notations we can rewrite our system model as

$$\begin{aligned} & i_{tt} + \alpha i_t + \beta i - \gamma i_{xx} = \gamma f_x, \\ & i(0, \cdot) = i_0, \\ & i_t(0, \cdot) = i_1, \qquad (t, x) \in Q, \\ & i/\Sigma = u. \end{aligned} \tag{2.6}$$

Let $\mathcal{U}_{ad}$, a closed convex subset of $L_2(I) \times L_2(I)$, denote the class of admissible controls or the admissible generating policies over the period of operation $I = (0, T)$. In the sequel we will consider several performance functions $J(u)$ representing the cost of operation of the system. The problem is to find a control $u \in \mathcal{U}_{ad}$, subject to the dynamic constraint (2.6), so that $J(u) \leqq J(v)$ for all $v \in \mathcal{U}_{ad}$. We consider in this paper a distributed cost function

$$\begin{aligned} J(u) = & \frac{\lambda_0}{2} \int_Q (i(t, x) - i_d(t, x))^2 \, dx \, dt + \frac{1}{2} \int_0^T (\lambda_1 u_1^2 + \lambda_2 u_2^2) \, dt \\ & + \frac{R}{2} \int_Q (i(t, x))^2 \, dx \, dt, \end{aligned} \tag{2.7}$$

and a terminal cost function

$$J(u) = \frac{\lambda_0}{2} \int_\Omega (i(T, x) - i_d(x))^2 \, dx + \frac{1}{2} \int_0^T (\lambda_1 u_1^2 + \lambda_2 u_2^2) \, dt \tag{2.8}$$

and present the corresponding necessary conditions of optimality from which optimal generating policies can be computed. The function i_d ($i_d = i_d(t, x)$ or $i_d(x)$) represents the desired current distribution and is assumed known. The first term in the cost functions, with $\lambda_0 > 0$, represents losses due to deviation from the expected demand and the second term, with $\lambda_1, \lambda_2 > 0$, represents cost of generation. The last term in (2.7) represents resistive losses dissipated in the form of heat.

Case (ii). A power system consisting of a primary transmission line with generating stations located at intermediate points in addition to those at the terminals can be modeled in a similar way. We point out the modifications necessary. Suppose, in addition to the terminal generating stations, there are additional $(n-2)$ generating sources located at points $x_2, x_3, \cdots, x_{n-1} \in \Omega$ with $0 = x_1$ and $L_0 = x_n$. Then the system (2.2) takes the form

$$\begin{aligned} & i_x = -Ge - Ce_t - f + \sum_{s=2}^{n-1} u_s(t) \delta(x - x_s), \\ & e_x = -Ri - Li_t, \end{aligned} \tag{2.2$'$}$$

where u_i, $i = 2, \cdots, n-1$ are the generating levels of the sources located at the points x_i, $i = 2, \cdots, n-1$. Consequently the system (2.4) takes the form

$$i_{tt} + \alpha i_t + \beta i - \gamma i_{xx} = \gamma f_x - \gamma \sum_{s=2}^{n-1} u_s(t)\delta'(x - x_s), \tag{2.4}'$$

where δ' is the distributional derivative of the Dirac measure δ. The complete model in this case is then given by

$$\begin{aligned} & i_{tt} + \alpha i_t + \beta i - \gamma i_{xx} = \gamma f_x - \gamma \sum_{s=2}^{n-1} u_s(t)\delta'(x - x_s), \\ & i(0, \cdot) = i_0, \\ & i_t(0, \cdot) = i_1, \qquad (t, x) \in Q, \\ & i(t, 0) = u_1, \qquad i(t, L_0) = u_n. \end{aligned} \tag{2.6}'$$

The necessary conditions of optimality for both the cases are presented in § 4. The existence theorem is given in § 5 and a computational algorithm is presented in § 6.

3. Some preparatory results. Let B be a real Banach space with the norm $|\cdot|$ and I an open interval of the real line R. Denote by $L_p(I, B)$, $1 \leqq p \leqq \infty$, the equivalence classes of strongly measurable functions on I with values in B and equipped with the norm $\|\cdot\|$ where

$$\|\cdot\|^p = \int_I |\cdot|^p \, dt.$$

For $B = H$, a Hilbert space and $p = 2$, $L_2(I, H)$ is a Hilbert space. Let H denote a real Hilbert space of functions defined on $\Omega = (0, L_0)$ with scalar product denoted by $\langle \cdot, \cdot \rangle$ and norm by $|\cdot|$. We introduce the following Sobolev spaces:

$$H^1(\Omega) = \{e \in H = L_2(\Omega) : e, e_x \in L_2(\Omega)\}$$

with scalar product

$$(e, f) = \langle e, f \rangle + \langle e_x, f_x \rangle$$

and norm $\|e\|_1$ with

$$\|e\|_1^2 = |e|^2 + |e_x|^2;$$

$$H^{1,1}(Q) = \{e \in L_2(Q) : e, e_t, e_x \in L_2(Q)\}$$

with the scalar product

$$(e, f) = \int_0^T \langle e, f \rangle \, dt + \int_0^T \langle e_t, f_t \rangle \, dt + \int_0^T \langle e_x, f_x \rangle \, dt$$

and norm $\|e\|_{1,1}$ where

$$\|e\|_{1,1}^2 = \int_0^T \{|e|^2 + |e_t|^2 + |e_x|^2\} \, dt;$$

$$H^{1,0} = \{e \in L_2(Q) : e, e_t \in L_2(Q)\}$$

with scalar product

$$(e, f) = \int_0^T \langle e, f \rangle \, dt + \int_0^T \langle e_t, f_t \rangle \, dt$$

and norm $\|e\|_{1,0}$ where

$$\|e\|_{1,0}^2 = \int_0^T |e|^2 dt + \int_0^T |e_t|^2 dt.$$

Similarly $H^{0,1}$ is defined. Clearly $H^{0,0}(Q) = L_2(Q)$ and $H^{0,1}(Q) = L_2(I, H^1)$. By H^{-1} we will denote the dual of H_0^1 where H_0^1 is the closure in H^1 topology of C^∞ functions with compact support. Thus $L_2(I, H^{-1}) = (L_2(I, H_0^1))^*$.

There are several techniques for solving nonhomogeneous boundary value problems (Lions and Magenes [4, vol. 1, Chap. 3]). In this paper we will use the method of transposition. Towards this end let us define the map F

$$\varphi \to \varphi_{tt} + \alpha\varphi_t + \beta\varphi - \gamma\varphi_{xx}$$

and F^* its formal adjoint given by

$$\varphi \to \varphi_{tt} - \alpha\varphi_t + \beta\varphi - \gamma\varphi_{xx}.$$

We then consider the homogenous boundary value problem

$$\begin{aligned} &F^*(\varphi) = h \quad \text{for } (t, x) \in Q \equiv I \times \Omega = (0, T) \times \Omega, \\ &\varphi(T, \cdot) = 0, \\ &\varphi_t(T, \cdot) = 0, \\ &\varphi|_\Sigma = 0, \quad \text{where } \Sigma = (0, T) \times \{0, L_0\}. \end{aligned} \tag{3.1}$$

For the problem (3.1) we have the following result.

LEMMA 3.1. *Let* $\alpha, \beta, \gamma > 0$, $h \in L_2(I, H) = L_2(Q)$ *and* $0 < T < \infty$. *Then the system* (3.1) *has a unique solution* $\varphi \in H^{1,1}(Q)$.

Proof. For the proof we use Galerkin's approach after an a priori bound is established. Multiplying the first equation in (3.1) by φ_t and integrating over the set Ω we obtain

$$\frac{d}{dt}|Y(t)|^2 - 2\alpha|\varphi_t|^2 = 2\langle h, \varphi_t \rangle \tag{3.2}$$

where, since $\beta, \gamma > 0$,

$$|Y(t)|^2 \equiv |\varphi_t|^2 + \beta|\varphi|^2 + \gamma|\varphi_x|^2 \geqq 0$$

and

$$|\cdot|^2 = \int_\Omega (\cdot)^2 \, dx.$$

Integrating once with respect to t we have from (3.2) that

$$|Y(t)|^2 - 2\alpha \int_0^t |\varphi_t|^2 \, d\theta = |Y(0)|^2 + 2 \int_0^t \langle h(\theta, \cdot), \varphi_t(\theta, \cdot) \rangle \, d\theta \tag{3.3}$$

for each $t \in [0, T]$. Subtracting (3.3) from the same for $t = T$ one obtains

$$|Y(t)|^2 + 2\alpha \int_t^T |\varphi_t|^2 \, d\theta = |Y(T)|^2 - 2 \int_t^T \langle h, \varphi_t \rangle \, d\theta. \tag{3.4}$$

Since $\varphi(T, \cdot) = \varphi_t(T, \cdot) \equiv 0$ on Ω it follows from the definition of Y that $|Y(T)|^2 = 0$. Thus

$$|Y(t)|^2 + 2\alpha \int_t^T |\varphi_t|^2 \, d\theta \leqq 2\left(\int_t^T |h|^2 \, d\theta \right)^{1/2} \left(\int_t^T |\varphi_t|^2 \, d\theta \right)^{1/2}. \tag{3.5}$$

Using the elementary inequality

$$|a|\,|b| \leqq \frac{a^2}{2\varepsilon^2} + \frac{\varepsilon^2}{2} b^2 \quad \text{for } \varepsilon > 0$$

in (3.5) we obtain

$$|Y(t)|^2 + 2\alpha \int_t^T |\varphi_t|^2 \, d\theta \leqq \frac{1}{\varepsilon^2} \int_t^T |h|^2 \, d\theta + \varepsilon^2 \int_t^T |\varphi_t|^2 \, d\theta$$

or equivalently

$$|Y(t)|^2 + (2\alpha - \varepsilon^2) \int_t^T |\varphi_t|^2 \, d\theta \leqq \frac{1}{\varepsilon^2} \int_t^T |h|^2 \, d\theta. \tag{3.6}$$

Since $\alpha > 0$, it follows from (3.6) by choice of $\varepsilon > 0$ so that $\alpha - \varepsilon^2 > 0$ that

$$|Y(t)|^2 \leqq \frac{1}{\varepsilon^2} \int_0^T |h|^2 \, d\theta \tag{3.7}$$

and since $0 \leqq t < T < \infty$ we have $\int_0^T |Y(t)|^2 \, dt < \infty$. Thus we conclude that if the problem (3.1) has a solution φ it is bounded in $H^{1,1}(Q)$ or equivalently $\varphi \in L_2(I, H^1(\Omega))$ and $\varphi_t \in L_2(I, H)$. Denote by H_0^1 the class of functions in H^1 that vanish on the boundary $\partial\Omega$ and let $\{w_i\}$ be an orthonormal system on Ω complete in the class H_0^1. Following Galerkin's procedure, the solution of the problem (3.1) is approximated by the solution of the finite dimensional problem

$$\begin{aligned} &\sum_{j=1}^m \ddot{z}_j^{(m)} \langle w_j, w_i \rangle - \alpha \sum_{j=1}^m \dot{z}_j^{(m)} \langle w_j, w_i \rangle + \beta \sum_{j=1}^m z_j^{(m)} \langle w_j, w_i \rangle \\ &\qquad + \gamma \sum_{j=1}^m z_j^{(m)} \langle (w_j)_x, (w_i)_x \rangle = \langle h, w_i \rangle, \\ &z_i^{(m)}(T) = 0, \\ &\dot{z}_i^{(m)}(T) = 0, \qquad 1 \leqq i \leqq m. \end{aligned} \tag{3.8}$$

Since, for each positive integer m, the system in (3.8) is finite dimensional and linear it has a unique solution satisfying the terminal conditions. By defining

$$\varphi^m = \sum_{i=1}^m z_i^{(m)} w_i \tag{3.9}$$

and using the a priori bound established above it is easily shown that $\{\varphi^m\}$ is a

bounded sequence from $H^{1,1}(Q)$. Thus there exists a subsequence again denoted by $\{\varphi^m\}$ and an element $\varphi \in H^{1,1}(Q)$ so that $\varphi^m \to \varphi$ weakly in $H^{1,1}(Q)$. Equivalently $\{\varphi^m\}$, $\{\varphi_t^m\}$ and $\{\varphi_x^m\}$ are bounded in $L_2(I, H)$ and have weak limits φ, φ_t and φ_x respectively with $\varphi^m(T, \cdot) \to 0$ and $\varphi_t^m(T, \cdot) \to 0$ strongly in H^1 and H respectively. From these facts it follows that φ is a weak solution of the problem (3.1) in the sense that

$$(3.10) \quad -\int_0^T \langle \varphi_t, g_t + \alpha g \rangle \, dt + \beta \int_0^T \langle \varphi, g \rangle \, dt + \gamma \int_0^T \langle \varphi_x g_x \rangle \, dt = \int_0^T \langle h, g \rangle \, dt$$

for all $g \in H^{1,1}(Q)$ with $g(0, \cdot) = 0$ and $g|_\Sigma = 0$. Choosing g with compact support in Q, it follows from (3.10) that φ solves problem (3.1) in the sense of distributions. For uniqueness it follows from the inequality (3.7), with $h = 0$, that $|Y(t)|^2 = 0$ and consequently $\varphi \equiv 0$. This implies uniqueness. This completes the proof.

Remark 3.2. Note that, since $\{w_i\}$ is orthonormal the system (3.8) can be written as

$$\dot{Y}^{(m)} = A^{(m)} Y^{(m)} + H^{(m)}$$

where

$$Y^{(m)} = (Y_1^{(m)}, Y_2^{(m)})', \quad Y_1^{(m)} = z^{(m)} \quad \text{and} \quad Y_2^{(m)} = \dot{z}^{(m)},$$

$$A^{(m)} = \begin{bmatrix} 0 & I \\ -(\beta I + \gamma A), & \alpha I \end{bmatrix}, \qquad z^{(m)} = (z_1^{(m)} \cdots z_m^{(m)}),$$

$I = (m \times m)$ identify matrix,

$A = \{\langle (w_i)_x, (w_j)_x \rangle, i, j = 1, 2, \cdots, m\}$, an $m \times m$ matrix,

and

$$H^{(m)} = [0, h^{(m)}]' \quad \text{with } h^{(m)} = \{\langle h, w_i \rangle i = 1, 2, \cdots, m\}.$$

Here "$'$" denotes transpose of a vector.

Remark 3.3. In fact the above result remains valid if $-\gamma\psi_{xx}$ is replaced by $B(t)\psi$ where $B(t)$ is any self adjoint elliptic operator from H^1 to $(H^1)^*$ and $\langle B(t)\psi, \psi \rangle \geqq b|\psi|_{H^1}$ with $b > 0$ and $h \in (H^1)^*$ where we have used $*$ to denote the dual and $\langle \cdot, \cdot \rangle$ to denote the duality bracket in H^1 and $(H^1)^*$ and $|\cdot|_{H^1}$ to denote the norm in H^1. For general results see Lions [2, p. 272–327].

With this preparation we can now solve the nonhomogeneous boundary value problem (2.6).

Note. We will, from now on, denote the response of the system (2.6) or (2.6′) corresponding to the control u by $i(u)$ and its values by $i(u)(t, x)$, $(t, x) \in Q$.

LEMMA 3.4. *Consider the system* (2.6) *with data* $i_0 \in H^1$, $i_1 \in H$ *and* $f \in L_2(I, H^1)$ *and suppose the assumptions of Lemma* 3.1 *hold. Then for each control* $u \in \mathcal{U}_{ad}$ *there exists a unique solution* $i(u) \in L_2(I, H)$ *for the problem* (2.6) *and the map* $u \to i(u)$ *is affine continuous.*

Proof. We will transpose Lemma 3.1. By Lemma 3.1, for each $h \in L_2(I, H)$ there exists a unique $\varphi \in H^{1,1}(Q)$ with $\varphi(T, \cdot) = \varphi_t(T, \cdot) = 0$ and $\varphi|_\Sigma = 0$. Define

$$X \equiv \{\varphi \in H^{1,1}(Q) : F^*(\varphi) \in L_2(Q), \varphi(T, \cdot) = \varphi_t(T, \cdot) = 0, \varphi|_\Sigma = \varphi_t|_\Sigma = 0\}$$

and suppose it is endowed with the topology induced by the norm $\|\cdot\|_X$ where

$$\|\varphi\|_X = \|F^*(\varphi)\|_{L_2(I,H)} = \|h\|_{L_2(I,H)} \tag{3.11}$$

with φ the solution of the homogeneous boundary value problem (3.1) corresponding to $h \in L_2(I, H)$. Endowed with this norm, X is a Hilbert space and the operator F^* is an isomorphism of X onto $L_2(I, H)$. With φ from X, the system (2.6) can be written in the following variational form:

$$\int_0^T \langle i(u), F^*(\varphi)\rangle \, dt = l(\varphi) \tag{3.12}$$

where

$$\begin{aligned} l(\varphi) \equiv & \int_0^T \langle \gamma f_x \varphi\rangle \, dt + \int_\Omega i_0(x)[\alpha\varphi(0,x) - \varphi_t(0,x)] \, dx + \int_\Omega i_1(x)\varphi(0,x) \, dx \\ & + \gamma \int_0^T [u_1(t)\varphi_x(t,0) - u_2(t)\varphi_x(t,L_0)] \, dt. \end{aligned} \tag{3.13}$$

The system (3.12) is equivalent to (2.6) and a formal interpretation of (3.12) is obtained by choice of a $\varphi \in X$ with compact support in Q. This leads to the equality

$$\int_0^T \langle i(u), F^*(\varphi)\rangle \, dt = \int_0^T \langle \gamma f_x, \varphi\rangle \, dt. \tag{3.14}$$

Since (3.14) holds for every $\varphi \in X$, with compact support in Q, it follows that

$$i_{tt}(u) + \alpha i_t(u) + \beta i(u) - \gamma i_{xx}(u) = \gamma f_x \quad \text{in } Q. \tag{3.15}$$

Multiplying (3.15) by $\varphi \in X$ and integrating by parts over Q we obtain

$$\begin{aligned} \int_0^T \langle i(u), F^*(\varphi)\rangle \, dt = & \int_0^T \langle \gamma f_x, \varphi\rangle \, dt + \int_\Omega i(u)(0,x)[\alpha\varphi(0,x) - \varphi_t(0,x)] \, dx \\ & + \int_\Omega i_t(u)(0,x)\varphi(0,x) \, dx \\ & + \gamma \int_\Omega [i(u)(t,0)\varphi_x(t,0) - i(u)(t,L_0)\varphi_x(t,L_0)] \, dt. \end{aligned} \tag{3.16}$$

Clearly this shows that $i(u)(0,\cdot) = i_0$, $i_t(u)(0,\cdot) = i_1$, $i(u)(t,0) = u_1(t)$ and $i(u)(t,L_0) = u_2(t)$. Hence the equivalence. Since $f \in L_2(I, H^1)$, $i_0 \in H^1$, $i_1 \in H$ and $u \in \mathcal{U}_{ad} \subset L_2 \times L_2$ it follows that l is a continuous linear functional on X, that is, $l \in X^*$ the dual of X. Further F^* is an isomorphism of X onto $L_2(I, H)$; thus it follows from the principle of transposition (Lions and Magenes [4, vol 1, p. 283]; Lions, [2, p. 291]) that there exists a unique $i(u) \in L_2(I, H)$ so that

$$\int_0^T \langle i(u), F^*(\varphi)\rangle \, dt = l(\varphi) \quad \text{for all } \varphi \in X.$$

Further the map $l \to i$ from X^* into $L_2(I, H)$ is continuous and in particular the map $u \to i(u)$ from $\mathcal{U}_{ad} \subset L_2 \times L_2$ into $L_2(I, H)$ is affine and continuous. This completes the proof of the lemma.

In fact the Lemma 3.4 is true under more relaxed condition as stated below.

LEMMA 3.5. *Consider the system* (2.6) *with* $i_0 \in H^1$, $i_1 \in H$ *and* $f \in L_2(I, H^{-1})$ *and suppose the assumptions of Lemma* 3.1 *hold. Then for each control* $u \in \mathcal{U}_{ad}$ *there exists a unique solution* $i(u) \in L_2(I, H)$.

Proof. The only difference in the proof is in the definition of the functional l. In the present case we take

$$l(\varphi) = -\int_0^T \langle \gamma f, \varphi_x \rangle \, dt + \int_\Omega i_0(x)[\alpha\varphi(0,x) - \varphi_t(0,x)] \, dx + \int_\Omega i_1(x)\varphi(0,x) \, dx$$
$$+ \gamma \int_0^T [u_1(t)\varphi_x(t,0) - u_2(t)\varphi_x(t,L_0)] \, dt. \tag{3.17}$$

Since $\varphi|_\Sigma = 0$, the two functionals (3.13) and (3.17) are identical. The integrand in the first term of (3.17) is the duality product between H^{-1} and H^1.

For the proof of the necessary conditions presented in the following section we need one more lemma. This result asserts the existence of the Gateaux differential of $i(u)$ with respect to the control u. Define

$$\hat{i}(v) = \text{w. } \lim_{\varepsilon \to 0} \frac{i(u^0 + \varepsilon v) - i(u^0)}{\varepsilon}$$

to be the weak Gateaux differential of i at u^0 in the direction v.

LEMMA 3.6. *The solution* $i(u)$ *of the first boundary value problem* (2.6) *corresponding to the control* u *has a linear (weak) Gateaux differential at every point* $u \in \mathcal{U}_{ad}$. *This differential is unique, independent of* u *and is given by the solution* $\hat{i}(v) \in L_2(I, H)$ *of the problem*

$$\begin{aligned} &\hat{i}_{tt} + \alpha\hat{i}_t + \beta\hat{i} - \gamma\hat{i}_{xx} = 0, \\ &\hat{i}(0,\cdot) = 0, \qquad \hat{i}_t(0,\cdot) = 0, \\ &\hat{i}|_\Sigma = v \in \mathcal{U}_{ad}. \end{aligned} \tag{3.18}$$

Further, the transformation $v \to \hat{i}(v)$ *is linear.*

Proof. The proof is similar to that as given in Lemma 3.4 with

$$l(\varphi) = \gamma \int_0^T [v_1(t)\varphi_x(t,0) - v_2(t)\varphi_x(t,L_0)] \, dt. \tag{3.19}$$

That $\hat{i}(v)$ is the Gateaux differential of $i(u)$ in the direction v follows from subtracting (3.12) corresponding to u from that corresponding to $u + \varepsilon v$ and then taking the limit $\varepsilon \to 0$. Linearity of the transformation $v \to \hat{i}(v)$ is obvious.

4. Necessary conditions of optimality. With the help of the above results we will be able to prove several necessary conditions of optimality.

THEOREM 4.1. *Consider the system* (2.6) *in its variational form* (3.12) *with* l *defined by either* (3.13) *or* (3.17). *Let the hypotheses of Lemma* 3.4 *or Lemma* 3.5 *hold and suppose the cost function* J *is given by* (2.7) *with* $i_d \in L_2(I, H)$ *and* $\lambda_0, \lambda_1, \lambda_2 > 0$. *Then the optimal control* u *is determined by the simultaneous solution of the*

system of equations

$$\int_0^T \langle i(u), F^*(\psi)\rangle\, dt = l(\psi) \tag{4.1}$$

and

$$\begin{aligned} &\psi_{tt} - \alpha\psi_t + \beta\psi - \gamma\psi_{xx} = \lambda_0(i(u) - i_d) + Ri(u), \\ &\psi(T, \cdot) = \psi_t(T, \cdot) = 0, \\ &\psi|_\Sigma = \psi_t|_\Sigma = 0, \end{aligned} \tag{4.2}$$

and the inequality

$$\int_0^T \{[\lambda_1 u_1 + \gamma\psi_x(t, 0)](w_1 - u_1) + [\lambda_2 u_2 - \gamma\psi_x(t, L_0)](w_2 - u_2)\}\, dt \geqq 0 \tag{4.3}$$

for all $w = (w_1, w_2) \in \mathcal{U}_{\text{ad}}$.

Proof. Let $u \in \mathcal{U}_{\text{ad}}$ be the optimal control (existence proved in Theorem 5.2). Then, for the given data i_0, i_1 and f and the control u, the functional l is well defined through the expressions (3.13) or (3.17). Thus, by Lemma 3.4 or Lemma 3.5, there exists a unique solution $i(u) \in L_2(I, H)$ for the problem (4.1). Therefore $Ri(u) + \lambda_0(i(u) - i_d) \in L_2(I, H)$ and consequently by Lemma 3.1 the system (4.2) has a unique solution $\psi(u) \in X$. Since for arbitrary $v \in \mathcal{U}_{\text{ad}}$, (3.18) has a unique solution $i(v) \in L_2(I, H)$ (Lemma 3.6), we can scalar multiply the first equation in (4.2) by $\hat{i}(v)$ in the Hilbert space $L_2(I, H)$ giving

$$\int_0^T \langle \hat{i}(v), F^*(\psi)\rangle\, dt = \lambda_0 \int_0^T \left\langle \hat{i}(v), \left(1 + \frac{R}{\lambda_0}\right) i(u) - i_d \right\rangle dt. \tag{4.4}$$

It follows from Lemma 3.6 that

$$\int_0^T \langle \hat{i}(v), F^*(\psi)\rangle\, dt = \gamma \int_0^T [v_1\psi_x(t, 0) - v_2\psi_x(t, L_0)]\, dt. \tag{4.5}$$

Since i has a Gateaux differential on $\mathcal{U}_{\text{ad}}$ (Lemma 3.6) and J is defined by (2.7) it follows that J also has a Gateaux differential. Further J is quadratic in u and therefore convex. Since by hypothesis $\mathcal{U}_{\text{ad}}$ is a closed convex set, the necessary and sufficient condition that J attains its minimum at $u \in \mathcal{U}_{\text{ad}}$ is that the Gateaux differential

$$J'_u(w - u) \geqq 0 \quad \text{for all } w \in \mathcal{U}_{\text{ad}}. \tag{4.6}$$

This gives us the inequality

$$\begin{aligned} J'_u(w - u) = \lambda_0 &\int_0^T \left\langle \hat{i}(w - u), \left(1 + \frac{R}{\lambda_0}\right) i(u) - i_d \right\rangle dt \\ &+ \int_0^T [\lambda_1 u_1(w_1 - u_1) + \lambda_2 u_2(w_2 - u_2)]\, dt \geqq 0 \quad \text{for } w \in \mathcal{U}_{\text{ad}}, \end{aligned} \tag{4.7}$$

where $\hat{i}(w - u)$ is the solution of the problem (3.18) corresponding to the control $v = (w - u)$.

Using (4.4) and (4.5) for $v=(w-u)$ in the inequality (4.7) we obtain the desired inequality (4.3). That is, (4.3) is a necessary and sufficient condition for optimality. This completes the proof.

Remark 4.2. In case $\mathcal{U}_{ad}=L_2\times L_2$ (no control constraint) the optimal control is given by

$$u=\begin{cases} u_1(t)=-\dfrac{\gamma}{\lambda_1}\psi_x(t,0), \\ u_2(t)=\dfrac{\gamma}{\lambda_2}\psi_x(t,L_0). \end{cases}$$

In this case the optimal control is given by the solution of the two point boundary value problems

$$\begin{aligned} &i_{tt}+\alpha i_t+\beta i-\gamma i_{xx}=\gamma f_x, \\ &i(0,x)=i_0(x), \qquad i_t(0,x)=i_1(x) \\ &i(t,0)=-\frac{\gamma}{\lambda_1}\psi_x(t,0), \qquad i(t,L_0)=\frac{\gamma}{\lambda_2}\psi_x(t,L_0), \end{aligned} \tag{4.8}$$

and

$$\begin{aligned} &\psi_{tt}-\alpha\psi_t+\beta\psi-\gamma\psi_{xx}=\lambda_0(i-i_d)+Ri, \\ &\psi(T,x)=\psi_t(T,x)=0, \\ &\psi(t,0)=\psi(t,L_0)=\psi_t(t,0)=\psi_t(t,L_0)=0. \end{aligned} \tag{4.9}$$

The system (4.8) is solved in the variational form as defined by (4.1).

In problems where power demand changes fast and it is required to adjust the generating level to meet the predicted demand by a given time, it is natural to use a terminal cost function. Denoting the period of transition by the same interval $(0, T)$ as in Theorem 4.1, the cost function is given by the expression (2.8). For this problem we have the following result.

THEOREM 4.3. *Consider the system* (2.6) *in its variational form* (3.12) *with l given by either* (3.13) *or* (3.17). *Let the hypotheses of Lemma* 3.4 *or Lemma* 3.5 *hold and suppose the cost function is given by* (2.8) *with* $i_d\in H$ *and* $\lambda_0, \lambda_1, \lambda_2>0$. *Then the optimal control u is determined by the simultaneous solution of the system of equations* (4.1),

$$\begin{aligned} &\psi_{tt}-\alpha\psi_t+\beta\psi-\gamma\psi_{xx}=0, \\ &\psi(T,x)=0, \\ &\psi_t(T,x)=\lambda_0[i(u)(T,x)-i_d(x)], \\ &\psi|_\Sigma=\psi_t|_\Sigma=0, \end{aligned} \tag{4.10}$$

and the inequality

$$\int_0^T \{[\lambda_1 u_1-\gamma\psi_x(t,0)](w_1-u_1)+[\lambda_2 u_2+\gamma\psi_x(t,L_0)](w_2-u_2)\}\,dt\geqq 0 \tag{4.11}$$

for all $w\in\mathcal{U}_{ad}$.

Proof. By Lemma 3.5 the system (4.1) has a unique solution $i(u) \in L_2(I, H)$ for $u \in \mathcal{U}_{ad}$. Thus $i(u)$ has continuous representative. Choosing the continuous representative we have $\lambda_0[i(u)(T, \cdot) - i_d(\cdot)] \in H$. Multiplying the first equation of (4.10) by $i(v)$ of Lemma 3.6 and using Green's formula we obtain formally

$$\begin{aligned}\int_0^T \langle \hat{i}(v), F^*(\psi)\rangle\, dt &= \int_0^T \langle \psi, F(\hat{i}(v))\rangle\, dt \\ &+ \lambda_0 \int_\Omega \hat{i}(v)(T, x)[i(u)(T, x) - i_d(x)]\, dx \\ &+ \gamma \int_0^T [v_1 \psi_x(t, 0) - v_2 \psi_x(t, L_0)]\, dt.\end{aligned} \tag{4.12}$$

From (4.10) $\int_0^T \langle \hat{i}(v), F^*(\psi)\rangle\, dt = 0$ and from (3.18) of Lemma 3.6 $\int_0^T \langle \psi, F(\hat{i}(v))\rangle\, dt = 0$ and consequently

$$\lambda_0 \int_\Omega \hat{i}(v)(T, x)[i(u)(T, x) - i_d(x)]\, dx = \gamma \int_0^T [v_2 \psi_x(t, L_0) - v_1 \psi_x(t, 0)]\, dt. \tag{4.13}$$

As in Theorem 4.1 we have for all $w \in \mathcal{U}_{ad}$,

$$\begin{aligned}J_u'(w - u) &= \lambda_0 \int_\Omega \hat{i}(w - u)(T, x)[i(u)(T, x) - i_d(x)]\, dx \\ &+ \int_0^T [\lambda_1 u_1(w_1 - u_1) + \lambda_2 u_2(w_2 - u_2)]\, dt \geqq 0\end{aligned} \tag{4.14}$$

where $w = (w_1, w_2)$ and $u = (u_1, u_2)$. Substituting (4.13), corresponding to $v = (w - u)$, into (4.14) we have

$$\int_0^T \{[\lambda_1 u_1 - \gamma \psi_x(t, 0)](w_1 - u_1) + [\lambda_2 u_2 + \gamma \psi_x(t, L_0)](w_2 - u_2)\}\, dt \geqq 0$$

for all $w \in \mathcal{U}_{ad}$. Formal application of Green's formula in (4.12) is justified by approximating $i(u)$ and $\hat{i}(v)$ by a sequence of regular functions (for example integral averages) and then passing to the limit.

Remark 4.4. Again for the unconstrained problem, $\mathcal{U}_{ad} = L_2 \times L_2$, the optimal control is given by

$$u_1(t) = \frac{\gamma}{\lambda_1} \psi_x(t, 0),$$

$$u_2(t) = -\frac{\gamma}{\lambda_2} \psi_x(t, L_0).$$

A power system usually consists of a network of transmission and distribution lines carrying power from generating stations located at several geographical positions. We consider here a system consisting of a transmission line with power supplied by generating stations located at n points $x_1, \cdots, x_n \in \bar{\Omega} = [0, L_0]$ with $x_1 (= 0)$ and $x_n (= L_0)$ being the end points and x_i, $i = 2, \cdots, n - 1$, the interior points of $\Omega = (0, L_0)$. The output (current) of the generating sources is denoted by

$u=(u_1, \cdots, u_n)$. Let L_2^n denote n-copies of the space $L_2(0, T)$ and $\mathcal{U}_{ad}$ a closed convex subset of L_2^n representing the class of admissible generating policies or controls. The system in this case is described by the first boundary value problem

$$
\begin{aligned}
& i_{tt}+\alpha i_t+\beta i-\gamma i_{xx}=\gamma f_x-\gamma \sum_{s=2}^{n-1} u_s(t)\delta'(x-x_s) \\
& i(0, x)=i_0(x), \quad i_t(0, x)=i_1(x), \qquad (t, x)\in Q, \\
& i(t, 0)=u_1(t), \qquad i(t, L_0)=u_n(t),
\end{aligned} \tag{4.15}
$$

where $\delta(x-x_s)$ is the Dirac mass located at $x=x_s$ and δ' its distributional derivative, which itself is a distribution (in the sense of Schwartz). Due to the presence of distributions in the equation it is now absolutely essential to interpret the system (4.15) in the sense of distribution and consider it in the variational form (4.1), here,

$$
\int_0^T \langle i(u), F^*(\varphi)\rangle\, dt=\tilde{l}(\varphi) \quad \text{for all } \varphi\in X \tag{4.16}
$$

where

$$
\begin{aligned}
\tilde{l}(\varphi)\equiv & -\gamma\int_0^T \langle f, \varphi_x\rangle\, dt+\int_\Omega \{i_0(x)[\alpha\varphi(0, x)-\varphi_t(0, x)]+i_1(x)\varphi(0, x)\}\, dx \\
& +\gamma\int_0^T \left\{\sum_{s=1}^{n-1} u_s(t)\varphi_x(t, x_s)-u_n(t)\varphi_x(t, x_n)\right\} dt, \qquad x_1=0, \quad x_n=L_0,
\end{aligned} \tag{4.17}
$$

and the set X and the operator F^* are as defined in § 3. The system (4.16) is equivalent to the system (4.15) as discussed in the proof of Lemma 3.4. The proof of the following lemmas is similar to that of Lemma 3.5 and 3.6.

LEMMA 4.5. *Consider the system* (4.15) *with* $\alpha, \beta, \gamma>0$, $f\in L_2(I, H^{-1})$, $i_0\in H^1$, *and* $i_1\in H$. *Then for every control* $u\in \mathcal{U}_{ad}\subseteq L_2^n$ *the problem* (4.15) *has a unique solution* $i(u)\in L_2(I, H)$ *in the sense that*

$$
\int_0^T \langle i(u), F^*(\varphi)\rangle\, dt=\tilde{l}(\varphi) \quad \textit{for all } \varphi\in X.
$$

For the Gateaux differential of $i(u)$ we have the following result.

LEMMA 4.6. *The solution* $i(u)$ *of the problem* (4.15) *corresponding to the control* u *has a linear Gateaux differential at every point in* $\mathcal{U}_{ad}\subseteq L_2^n$. *The differential is unique, independent of* u *and is given by the solution* $\hat{i}(v)\in L_2(I, H)$ *of the problem*

$$
\begin{aligned}
& \hat{i}_{tt}+\alpha\hat{i}_t+\beta\hat{i}-\gamma\hat{i}_{xx}=-\gamma\sum_{s=2}^{n-1} v_s\delta'(x-x_s) \\
& \hat{i}(0, x)=0, \qquad \hat{i}_t(0, x)=0, \\
& \hat{i}(t, 0)=v_1, \quad \hat{i}(t, L_0)=v_n, \quad v\in\mathcal{U}_{ad}, \quad (t, x)\in Q,
\end{aligned} \tag{4.18}
$$

in the sense that

$$\int_0^T \langle \hat{i}(v), F^*(\varphi)\rangle\, dt = \gamma \int_0^T \sum_{s=1}^{n-1} v_s(t)\varphi_x(t, x_s)\, dt - \gamma \int_0^T v_n(t)\varphi_x(t, x_n)\, dt \tag{4.19}$$

for all $\varphi \in X$ *where* $x_1 \equiv 0$ *and* $x_n \equiv L_0$.

With the help of the above results we can prove the following necessary and sufficient conditions for optimality.

THEOREM 4.7. *Consider the system* (4.15) *in its variational form* (4.16) *with* $\tilde{I}$ *as defined by* (4.17). *Let the hypotheses of Lemma* 4.5 *hold and suppose the cost function J is given by*

$$J(v) = \frac{\lambda_0}{2}\int_0^T \langle i(v) - i_d, i(v) - i_d\rangle\, dt + \frac{1}{2}\int_0^T \sum_{s=1}^{n} \lambda_s (v_s(t))^2\, dt + \frac{R}{2}\int_0^T \langle i(v), i(v)\rangle\, dt \tag{4.20}$$

with $i_d \in L_2(I, H)$, λ_0, λ_s, $s = 1, 2, \cdots, n$, *positive and* $v \in \mathcal{U}_{ad}$. *Then in order that* $u \in \mathcal{U}_{ad}$ *be the optimal control it is necessary and sufficient that the equalities*

$$\int_0^T \langle i(u), F^*(\psi)\rangle\, dt = \tilde{I}(\psi), \tag{4.21}$$

$$\psi_{tt} - \alpha\psi_t + \beta\psi - \gamma\psi_{xx} = \lambda_0(i(u) - i_d) + Ri(u),$$
$$\psi(T, \cdot) = \psi_t(T, \cdot) = 0 \qquad \text{for } (t, x) \in Q, \tag{4.22}$$
$$\psi|_\Sigma = \psi_t|_\Sigma = 0,$$

and the inequality

$$\int_0^T \sum_{s=1}^{n-1} [\lambda_s u_s(t) + \gamma\psi_x(t, x_s)](w_s(t) - u_s(t))\, dt + \int_0^T [\lambda_n u_n(t) - \gamma\psi_x(t, x_n)][w_n(t) - u_n(t)]\, dt \geqq 0 \tag{4.23}$$

for all $w \in \mathcal{U}_{ad}$ *hold simultaneously.*

Proof. The proof is similar to that given for Theorem 4.1.

Next we consider a pointwise necessary condition of optimality. Let U be a compact and convex subset of R^n and suppose $\mathcal{U}_{ad}$ consists of measurable functions on I with values in U. Clearly $\mathcal{U}_{ad} \subset L_\infty^n \subset L_2^n$ and the Theorem 4.7 holds. We present a pointwise necessary condition of optimality for this case.

COROLLARY 4.8. *Suppose the hypotheses of Theorem* 4.7 *hold and let* $\mathcal{U}_{ad} \subset L_\infty^n$, *as defined above, be the class of admissible controls. Then the necessary and sufficient conditions of optimality consist of the equalities* (4.21) *and* (4.22) *and*

the inequality

$$(4.24)\quad \sum_{s=1}^{n-1}[\lambda_s u_s(t)+\gamma\psi_x(t,x_s)][v_s-u_s(t)]+[\lambda_n u_n(t)-\gamma\psi_x(t,x_n)][v_n-u_n(t)]\geqq 0$$

for almost all $t\in I$ *and all* $v\in U$.

Proof. It suffices to demonstrate the inequality (4.24). Let E be any measurable set containing the point t and contracting to the one point set $\{t\}$ where t is an arbitrary point in $(0, T)$. Let $v\in U$ be arbitrary and define

$$\tilde{u}(\theta)=\begin{cases} v & \text{for } \theta\in E,\\ u(\theta) & \text{for } \theta\in I\backslash E, \quad I=(0,T),\end{cases}$$

which is an admissible control. Substituting $\tilde{u}$ for w in (4.23) and dividing the resulting expression by $\mu(E)$, the Lebesgue measure of the set E, we obtain

$$(4.25)\quad \frac{1}{\mu(E)}\int_E \Big\{\sum_{s=1}^{n-1}[\lambda_s u_s(\theta)+\gamma\psi_x(\theta,x_s)][v_s-u_s(\theta)] +[\lambda_n u_n(\theta)-\gamma\psi_x(\theta,x_n)]\,.\,[v_n-u_n(\theta)]\Big\}\,d\theta\geqq 0.$$

Since u and $\psi_x(\,\cdot\,, x_s)$ are measurable functions and consequently almost all points of I are Lebesgue density points with respect to these functions we obtain the inequality (4.24) by letting $\mu(E)\to 0$. This completes the proof.

Remark. In practice the current capacities of the generating stations are limited in amplitude and in that case the control range space is given by a hypercube

$$U=\{v\in R^n:\ |v_s|\leqq b_s,\ b_s>0,\ s=1,2,\cdots,n\}.$$

Therefore it follows from the above corollary that the optimal control $u=(u_1\cdots u_n)$ has the form

$$u_s(t)=\nu_s\Big(-\frac{\gamma}{2\lambda_s}\psi_x(t,x_s)\Big), \qquad s=1,2,\cdots,n-1,$$

$$u_n(t)=\nu_n\Big(+\frac{\gamma}{2\lambda_n}\psi_x(t,x_n)\Big),$$

where

$$\nu_s(z)=\begin{cases} b_s & \text{for } z\geqq b_s,\\ z & \text{for } |z|<b_s, \quad s=1,2,\cdots,n.\\ -b_s & \text{for } z\leqq -b_s.\end{cases}$$

5. Existence of optimal control policies. In the necessary (and sufficient) conditions developed in § 4 it was assumed that optimal control policies exist. In this section we give a proof of this fact. For this we will need the following lemma.

LEMMA 5.1. *Consider the system* (4.15) *in its variational form* (4.16) *with* $\tilde{l}$ *as defined by* (4.17) *and the cost function* J *given by* (4.20) *with* λ_0, λ_s $(s=1,2,\cdots,n)$ *positive,* $i_d\in L_2(I,H)$ *and* $\mathcal{U}_{\text{ad}}$ *a closed convex subset of* L_2^n. *Suppose*

Lemma 4.5 *holds. Then the cost function J is weakly lower semicontinuous on* $\mathcal{U}_{ad}$.

Proof. Let $\{u^k\}$ be a sequence from $\mathcal{U}_{ad}$ so that $u^k \to u^0$ weakly in L_2^n. We must show that $J(u^0) \leqq \liminf_k J(u^k)$. By Lemma 4.5 there exists a sequence of solutions $\{i(u^k)\}$ and $i(u^0) \in L_2(I, H)$ for the problem (4.15) so that for each k

$$\int_0^T \langle i(u^k), F^*(\varphi)\rangle \, dt = \tilde{l}^k(\varphi)$$

and

$$\int_0^T \langle i(u^0), F^*(\varphi)\rangle \, dt = \tilde{l}^0(\varphi)$$

for all $\varphi \in X$ with $\tilde{l}^k(\varphi)$ and $\tilde{l}^0(\varphi)$ given by (4.17) with u replaced by u^k and u^0 respectively. Since $u^k \to u^0$ weakly in L_2^n and $\varphi \in X$ it is clear from the expressions for $\tilde{l}^k$ and $\tilde{l}^0$ that $\tilde{l}^k \to \tilde{l}^0$ in the weak star topology of X^* (dual of X). Therefore

$$\int_0^T \langle i(u^k) - i(u^0), F^*(\varphi)\rangle \, dt \to 0$$

for all $\varphi \in X$. Since F^* is an isomorphism of X onto $L_2(I, H)$ it follows that $i(u^k) \to i(u^0)$ weakly in $L_2(I, H)$. Further it is easily verified that

$$\begin{aligned} J(u^k) \geqq J(u^0) + \lambda_0 \int_0^T \left\langle i(u^k) - i(u^0), \left(1 + \frac{R}{\lambda_0}\right) i(u^0) - i_d \right\rangle dt \\ + \int_0^T \left\{ \sum_{s=1}^n \lambda_s (u_s^k - u_s^0) u_s^0 \right\} dt. \end{aligned} \tag{5.1}$$

Since $u^k \to u^0$ weakly in L_2^n and $i(u^k) \to i(u^0)$ weakly in $L_2(I, H)$ it follows from application of limit inferior on either side of (5.1) that

$$J(u^0) \leqq \liminf_k J(u^k). \tag{5.2}$$

This completes the proof.

THEOREM 5.2 (existence theorem). *Consider the system* (4.15) *in its variational form* (4.16) *with* $\tilde{l}$ *as defined by* (4.17) *and cost function* (4.20). *Suppose the hypotheses of Lemma* 5.1 *are satisfied. Then there exists a unique optimal control* $u^0 \in \mathcal{U}_{ad}$ *so that* $J(u^0) \leqq J(v)$ *for all* $v \in \mathcal{U}_{ad}$.

Proof. Define

$$\inf \{J(v) : v \in \mathcal{U}_{ad}\} = \gamma. \tag{5.3}$$

Since by definition $J \geqq 0$ and for each u, $J(u) < \infty$, it is clear that $0 \leqq \gamma < \infty$. Let $\{u^k\}$ be a minimizing sequence from $\mathcal{U}_{ad}$ so that

$$\lim_k J(u^k) = \gamma. \tag{5.4}$$

Let $\|u\|$ denote the norm of u in L_2^n. Since $0 \leqq \gamma < \infty$ and $J(u) \to \infty$ as $\|u\| \to \infty$ it is clear that $\{u^k\}$ is a bounded subset of $\mathcal{U}_{ad} \subset L_2^n$. Thus, L_2^n being a reflexive Banach space, there exists a subsequence of the sequence $\{u^k\}$, again denoted by $\{u^k\}$, and an $u^0 \in L_2^n$ so that $u^k \to u^0$ weakly. Since $\mathcal{U}_{ad}$ is closed and convex it is weakly

closed and therefore $u^0 \in \mathcal{U}_{ad}$ and due to (5.3),

$$\gamma \leqq J(u^0). \tag{5.5}$$

By Lemma 5.1 J is weakly lower semicontinuous on $\mathcal{U}_{ad}$ and consequently

$$J(u^0) \leqq \liminf_k J(u^k). \tag{5.6}$$

Combining (5.4), (5.5) and (5.6) we have

$$\gamma \leqq J(u^0) \leqq \liminf_k J(u^k) = \lim_k J(u^k) = \gamma.$$

This shows that u^0 is an optimal control. Since J is quadratic in u it is strictly convex and therefore u^0 is unique. This completes the proof.

Remark 5.3. The result of Theorem 5.2 also holds for $\mathcal{U}_{ad}$ a closed convex subset of L_∞^n endowed with the weak star topology.

6. A computational procedure. An iterative technique based on the necessary conditions of § 4 can be developed for computing the optimal controls. The following discussion is presented with reference to Theorem 4.7 even though it applies to all the necessary conditions of optimality of § 4. The inequality (4.23) gives $J'_u(w-u)$ in terms of the adjoint state. Thus the gradient of the cost function is a linear functional on L_2^n that is for $v \in L_2^n$

$$J'_u(v) = \int_0^T \left\{ \sum_{s=1}^{n-1} [\lambda_s u_s + \gamma \psi_x(u)(t, x_s)] v_s + [\lambda_n u_n - \gamma \psi_x(u)(t, x_n)] v_n \right\} dt \tag{6.1}$$

where $\psi(u)$ is the adjoint state corresponding to $i(u)$ (see (4.22)) and hence the control u. The functional J'_u can be computed for each choice of $u \in \mathcal{U}_{ad}$ by solving the system and the adjoint equations. Since J is quadratic and the system equations are linear, the map $u \to J'_u$ from L_2^n to $(L_2^n)^*$, identified with L_2^n, is an affine map. Therefore there exists a $\eta \in L_2^n$ independent of u, $v \in L_2^n$ so that

$$J'_{u+\alpha v} = J'_u + \alpha(J'_v + \eta). \tag{6.2}$$

Let $u^0 \in L_2^n$ and define $u^1 = u^0 - \rho J'_{u^0}$, $w^0 = J'_{u^0}$ and denote $J'_u(w)$ by (J'_u, w), inner product in L_2^n. Then by Lagrange formula

$$\begin{aligned} J(u^1) &= J(u^0) + \int_0^1 J'_{u^0+\theta(u^1-u^0)}(u^1 - u^0)\, d\theta \\ &= J(u^0) + \int_0^1 (J'_{u^0+\theta(u^1-u^0)}, u^1 - u^0)\, d\theta \\ &= J(u^0) - \rho \int_0^1 (J'_{u^0-\theta\rho w^0}, J'_{u^0})\, d\theta. \end{aligned}$$

Using (6.2) in the above expression we obtain

$$J(u^1) = J(u^0) - \rho \|J'_{u^0}\|^2 + \frac{\rho^2}{2}(J'_{w^0} + \eta, J'_{u^0}). \tag{6.3}$$

Clearly $|(J'_{w^0}+\eta, J'_{u^0})|<\infty$ and consequently it follows from (6.3) that

$$J(u^1)=J(u^0)-\rho\|J'_{u^0}\|^2+o(\rho) \quad \text{where } \lim_{\rho\to 0}\frac{o(\rho)}{\rho}=0. \tag{6.4}$$

Therefore if $\rho>0$ and sufficiently small then $J(u^1)<J(u^0)$ for $u^1=u^0-\rho J'_{u^0}$. The choice of ρ is also dictated by the requirement that $u^1\in\mathcal{U}_{ad}\subset L_2^n$. The above discussion leads us to the following algorithm.

Step 1. Guess $u^0\in\mathcal{U}_{ad}$.

Step 2. Solve the system equation (4.15) in the weak form (4.21) to give $i^0=i(u^0)$.

Step 3. Solve the adjoint system (4.22) to give $\psi^0=\psi(u^0)$.

Step 4. Use the pair $\{u^0, \psi^0\}$ in the necessary condition (4.23) to find J'_{u^0}.

Step 5. Define $u^1=u^0-\rho J'_{u^0}$ with $\rho>0$ but sufficiently small so that $u^1\in\mathcal{U}_{ad}$.

Step 6. Compute $J(u^1)$. If $J(u^1)<J(u^0)$ go to Step 2 with u^1 replacing u^0 and if not reduce ρ by a suitable factor and go to Step 5.

Remark. A suboptimal control is obtained by introducing a stopping criterion $|J(u^{n+1})-J(u^n)|<\varepsilon$ or $\|u^{n+1}-u^n\|<\varepsilon$ for a suitable number ε.

For a discussion of other numerical methods the reader is referred to Lions [3, Chap. 9, p. 296].

REFERENCES

[1] H. W. DOMMEL, *Optimization of power systems*, A. V. Balakrishnan, ed., 4th IFIP Colloquium on Optimization Techniques, Los Angeles, CA, 1971, pp. 487–498.

[2] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971, Chap. IV, pp. 272–327.

[3] ———, *Various Topics in the Theory of Optimal Control of Distributed Systems*, Springer Lecture notes in Economics and Mathematical Systems, part 1, no. 105, Springer-Verlag, New York, 1974, pp. 166–308.

[4] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, New York, 1972.

[5] D. L. RUSSELL, *Quadratic performance criteria in boundary control of linear symmetric hyperbolic systems*, this Journal, 11 (1973), pp. 475–509.

[6] ———, *Optimal regulation of linear symmetric hyperbolic systems with finite dimensional controls*, this Journal, 4 (1966), pp. 276–294.

[7] ———, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, J. Math. Anal. Appl., 40 (1972), pp. 336–368.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

INVERTIBILITY OF CONTROL SYSTEMS ON LIE GROUPS*

RONALD M. HIRSCHORN†

Abstract. This paper gives necessary and sufficient conditions for the invertibility of a class of nonlinear systems which includes matrix bilinear systems. Lie algebraic invertibility criteria are obtained for bilinear systems in R^n which generalize the standard tests for single input linear systems. These results are used to construct nonlinear systems which act as left-inverses for bilinear systems.

1. Introduction. There is a considerable amount of literature dealing with the invertibility of linear control systems (cf. [1], [2], [6]–[8], [11]). The question of invertibility—when the output of a control system uniquely determines the input—is of practical as well as theoretical interest, and is related to functional controllability, problems in coding theory, etc.

The purpose of this paper is to show that the linear results on invertibility can be extended to a much larger class of systems. In particular, we consider the question of invertibility for right-invariant systems, where the state space is a Lie group, and bilinear systems where the control depends linearly on the state (cf. [3], [5], [9]).

The role of Lie theory in the study of right-invariant and bilinear systems is analogous to that of linear algebra in studying linear systems. Many of the standard matrix rank conditions for linear systems have been generalized as Lie algebraic criteria in the right-invariant and bilinear case (cf. [3], [5], [9]). In this paper we find necessary and sufficient conditions for invertibility which are Lie algebraic in nature and generalize known results for linear systems.

In § 2 we introduce notation and basic results. Section 3 examines the invertibility of right-invariant systems, which includes the class of matrix bilinear systems. In § 4 these results are used to construct left-inverses for bilinear systems. A left-inverse is a nonlinear system which, when driven by appropriate derivatives of the output of the original system, produces as an output $u(t)$, the input to the original bilinear system. A number of examples are presented in this section.

2. Notation and preliminary results. In this section we review some basic results and definitions which are used in this paper. We assume that the reader is familiar with the basic notions of differential geometry and Lie theory (cf. [4], [10]).

Let $\mathbf{H}$ be a Lie group. The right multiplication mapping $R_x: y \to yx$ from $\mathbf{H} \to \mathbf{H}$ has differential dR_x. A vector field X on $\mathbf{H}$ is called *right-invariant* if $dR_x X(y) = X(yx)$ for all $y \in \mathbf{H}$. The collection of right-invariant vector fields, $\mathcal{H}$, is called the *Lie algebra* of $\mathbf{H}$.

A *single input-single output bilinear system* is a control system of the form

$$\begin{aligned} &\dot{x}(t) = Ax(t) + u(t)Bx(t); \qquad x(0) = x_0, \\ &y(t) = cx(t) \end{aligned} \tag{1}$$

---

* Received by the editors June 29, 1976, and in revised form January 24, 1977.

† Department of Mathematics, Queen's University, Kingston, Canada K7L 3N6.

where the state $x \in \mathbf{R}^n$; A and B are $n \times n$ matrices over $\mathbf{R}$, c is a $1 \times n$ matrix over $\mathbf{R}$, and $u \in \mathcal{U}$, the class of piecewise real analytic functions on $(0, \infty)$.

It is often convenient to express the solution to (1) as $x(t) = X(t)x_0$ where $X(t)$ is an $n \times n$ matrix valued function of t which is the trajectory of the corresponding matrix bilinear system.

A *single-input matrix bilinear system* is a system of the form

$$\begin{aligned} &\dot{X}(t) = AX(t) + u(t)BX(t); \qquad X(0) = X_0, \\ &Y(t) = CX(t) \end{aligned} \tag{2}$$

where A, B, X are $n \times n$ matrices over $\mathbf{R}$, $u \in \mathcal{U}$, and C is an $r \times n$ matrix over $\mathbf{R}$. We will assume that X_0 is invertible so that $X(t)$ evolves in $GL(n, \mathbf{R})$, the Lie group of invertable $n \times n$ in real matrices (cf. [3], [9]).

The matrix system (2) is a special case of the more general class of right-invariant systems studied in [9]).

DEFINITION. A *right-invariant system* is a system of the form

$$\begin{aligned} &\dot{x}(t) = A(x(t)) + \sum_{i=1}^{m} u_i(t)B_i(x(t)); \qquad x(0) = x_0 \in \mathbf{H}, \\ &y(t) = \mathbf{K}x(t) \end{aligned} \tag{3}$$

where $u_1, \cdots, u_m \in \mathcal{U}$, $\mathbf{H}$ is a Lie group, $\mathbf{K}$ is a Lie subgroup of $\mathbf{H}$ with Lie algebra $\mathcal{K}$, and $A, B_1, \cdots, B_m \in \mathcal{H}$, the Lie algebra of right-invariant vector fields on $\mathbf{H}$.

We remark that the coset output $y(t) = \mathbf{K}x(t)$ generalizes the output $Y(t) = CX(t)$ in (3). In particular one could set $\mathbf{K} = \{X \in GL(n, \mathbf{R})\colon CX = C\}$ and $\mathbf{H} = GL(n, \mathbf{R})$.

A *single-input right invariant system* is a system of the form

$$\begin{aligned} &\dot{x}(t) = A(x(t)) + u(t)B(x(t)); \qquad x(0) = x_0 \in \mathbf{H}, \\ &y(t) = \mathbf{K}x(t) \end{aligned} \tag{4}$$

where A, B, $\mathbf{K}$, u are defined as above.

DEFINITION. The right-invariant system (3) is said to be *invertible* if the output $\tau \to y(\tau)$ on any interval $0 \leqq \tau < t$ uniquely determines the input $\tau \to u(\tau)$ for $0 \leqq \tau < t$. That is, distinct inputs produce distinct outputs. Invertibility for systems (1), (2) and (4) are defined in an analogous manner.

The properties of a right-invariant system are related to the structure of the Lie algebra $\mathcal{H}$. The Lie algebra $\mathcal{H}$ is a vector space with a nonassociative "multiplication" defined as follows:

for $X, Y \in \mathcal{H}$ the *Lie bracket of X and Y is*

$$[X, Y](m) = X(m)Y - Y(m)X$$

(cf. [4], [10]). We define $\mathrm{ad}_X^n Y$ inductively as follows: $\mathrm{ad}_X^0 Y = Y$ and $\mathrm{ad}_X^k Y = [X, \mathrm{ad}_X^{k-1} Y]$. For matrix bilinear systems with $X, Y \in \mathcal{H}$ right-invariance means that $X(M) = XM$ and

$$[X, Y](M) = (YX - XY)M.$$

Let $\mathcal{S}$ be a subset of the Lie algebra $\mathcal{H}$. We define $\{\mathcal{S}\}_{\mathrm{LA}}$ to be the *Lie algebra generated in* $\mathcal{S}$ in $\mathcal{H}$. Thus $\{\mathcal{S}\}_{\mathrm{LA}}$ is the smallest Lie subalgebra of $\mathcal{H}$ containing $\mathcal{S}$. For each $x \in \mathbf{H}$ let $\mathcal{S}(x) = \{L(x)\colon L \in \mathcal{S}\}$.

It is known that the structure of the reachable set for (3) is related to the structure of the Lie algebras:

$$\mathscr{L} = \{A, B_1, B_2, \cdots, B_m\}_{\mathrm{LA}},$$

$$\mathscr{L}_0 = \{\mathrm{ad}_A^k B_i : k = 0, 1, \cdots \text{and } i = 1, \cdots, m\}_{\mathrm{LA}},$$

$$\mathscr{B} = \{B_1, \cdots, B_m\}_{\mathrm{LA}}.$$

Thus each right-invariant system has associated with it the chain of Lie algebras

$$\mathscr{H} \supset \mathscr{L} \supset \mathscr{L}_0 \supset \mathscr{B}.$$

If $\exp: \mathscr{H} \to \mathbf{H}$ is the standard exponential mapping in Lie theory than $\exp \mathscr{L} = \{\exp L : L \in \mathscr{L}\} \subset \mathbf{H}$ and the group generated by $\exp \mathscr{L}$, $\{\exp \mathscr{S}\}_{\mathrm{G}}$, is a Lie subgroup of $\mathbf{H}$ [4], [10]. Set

$$\mathbf{G} = \{\exp \mathscr{L}\}_{\mathrm{G}}, \qquad \mathbf{G}_0 = \{\exp \mathscr{L}_0\}_{\mathrm{G}}$$

and

$$\mathbf{B} = \{\exp \mathscr{B}\}_{\mathrm{G}}.$$

Thus each right-invariant system gives rise to the chain of Lie groups

$$\mathbf{H} \supset \mathbf{G} \supset \mathbf{G}_0 \supset \mathbf{B}.$$

Since $\mathscr{L}_0$ is an ideal in $\mathscr{L}$ (i.e. for each $L_0 \in \mathscr{L}_0$, $L \in \mathscr{L}$, $[L_0, L] \in \mathscr{L}_0$) we know that $\mathbf{G}_0$ is a normal subgroup of $\mathbf{G}$. The following theorem relates the structure of the trajectories of a bilinear system to the above group decompositions.

THEOREM 2.1 (Sussmann and Jurdjevic [9]). *Consider the right-invariant system* (3) *where the state x evolves in the Lie group $\mathbf{H}$ and $A, B_1, \cdots, B_m \in \mathscr{H}$. Associated with this system is the chain of Lie groups $\mathbf{H} \supset \mathbf{G} \supset \mathbf{G}_0 \supset \mathbf{B}$. Then for any set of controls $u_1, \cdots, u_m \in \mathscr{U}$ with corresponding trajectory $t \to x(t)$ we have $x(t) \in (\exp tA) \cdot \mathbf{G}_0 \cdot x_0$ for all $t \geqq 0$, where $(\exp tA) \cdot \mathbf{G}_0 \cdot x_0 = \{\exp tA \cdot g \cdot x_0 : g \in \mathbf{G}_0\}$.*

We conclude this section by presenting two formulae which are used in the next section. The mapping $L_x: y \to xy$ from $\mathbf{H} \to \mathbf{H}$ is called the left multiplication map. Suppose that $x = \exp X$ where $X \in \mathscr{H}$ and $x \in \mathbf{H}$. The mapping $A_x = L_x \circ R_{x^{-1}}: y \to xyx^{-1}$ of $\mathbf{H} \to \mathbf{H}$ has differential $dA_x = \mathrm{Ad}(x): \mathscr{H} \to \mathscr{H}$. The *Campbell–Baker–Hausdorff* formula for right-invariant vector fields asserts that

$$\mathrm{Ad}(x)(Y) = Y - \mathrm{ad}_X Y + \frac{1}{2!}\mathrm{ad}_X^2 Y - \frac{1}{3!}\mathrm{ad}_X^3 Y + \cdots$$

(cf. [4, p. 118]).

The exp mapping of $\mathscr{H} \to \mathbf{H}$ has a differential $X \in \mathscr{H}$, $d\exp_X: \mathscr{H} \to \mathscr{H}$ where

$$d\exp_X Y(e) = (dR_{\exp X})_e \circ \frac{1 - e^{\mathrm{ad}_X}}{-\mathrm{ad}_X} Y(e)$$

$$= Y(\exp X) + \frac{1}{2!}\mathrm{ad}_X Y(\exp X) + \frac{1}{3!}\mathrm{ad}_X^2 Y(\exp X) + \cdots$$

(cf. [4, p. 95]).

3. Invertibility criteria for right-invariant systems. In this section we derive necessary and sufficient conditions for the invertibility of right-invariant systems. The main result in this section is the following theorem:

THEOREM 3.1. *The right-invariant system* (4) *is invertible if and only if* $\mathrm{ad}_A^k B \not\in \mathcal{K}$ *for some* $k \in \{0, 1, \cdots, n-1\}$, *where n is the dimension of* $\mathcal{L}$ *and* $\mathcal{K}$ *is the Lie algebra of* **K**.

COROLLARY 1. *Consider the right-invariant system* (4) *with output* $y(t) = c(x(t))$ *where* c: $\mathbf{H} \to \mathbf{J}$ *is a Lie group homomorphism and* $c_*: \mathcal{H} \to \mathcal{J}$ *is the differential of c.*

This system is invertible if and only if $c_* \mathrm{ad}_A^k B \neq 0$ *for some positive integer* $k \in \{0, 1, \cdots, n-1\}$.

COROLLARY 2. *The matrix bilinear system* (2) *is invertible if and only if*

$$C \mathrm{ad}_A^k B \neq 0$$

for some positive integer $k \in \{0, 1, \cdots, n^2 - 1\}$ *where A and B are* $n \times n$ *matrices.*

COROLLARY 3. *The matrix bilinear system* (2) *fails to be invertible if and only if every control gives rise to the same output function.*

The similarity between the standard linear invertibility results and the above conditions is striking. In [2] Brockett shows that the single-input, single-output linear system $\dot{x} = Ax + bu$; $y = cx$ is invertible if and only if $cA^k b \neq 0$ for some positive integer k. The relative order α of the system is the least positive integer k such that $cA^{k-1}b \neq 0$ (or infinity). Rewriting this system in bilinear form (cf. [3]), we have

$$\dot{z} = A_1 z + uB_1 z,$$
$$y = C_1 z$$

where

$$z = \begin{pmatrix} x \\ 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix}, \quad C_1 = (c \quad 0).$$

The state transition matrix for this system is the state for corresponding matrix bilinear system. Corollary 2 above asserts that this matrix system is invertible if and only if $C_1 \mathrm{ad}_{A_1}^k B_1 \neq 0$ for some k. Since

$$C_1 \mathrm{ad}_{A_1}^k B_1 = \begin{pmatrix} 0 & cA^k b \\ 0 & 0 \end{pmatrix},$$

and invertibility in the linear case is independent of the initial state, the well-known linear result follows from the more general bilinear result, Corollary 2. This motivates the following definition:

DEFINITION. The *relative order*, α, of the matrix bilinear system (2) is the least positive integer k such that $C \mathrm{ad}_A^{k-1} B \neq 0$ or $\alpha = \infty$ if $C \mathrm{ad}_A^k B = 0$ for all $k > 0$.

As in the linear case, a matrix bilinear system is invertible if and only if the relative order $\alpha < \infty$. The remainder of this section will be devoted to proving this result.

In studying the invertibility of matrix bilinear systems one is tempted to repeat the approach which is successful in the linear case—differentiate the output

until the control $u(t)$ appears, and solve for $u(t)$ in terms of the derivatives of the output. Unfortunately the bilinear dependence of the control on the state greatly complicates the situation and little insight is obtained. Instead we will use the fact that the trajectory evolves in a Lie group. We begin by looking for a sufficient condition for invertibility for the right-invariant system (4). Suppose that this system is not invertible. This means that there are two *different* controls u_1 and u_2 which give rise to outputs y_1 and y_2 respectively, where $y_1 \equiv y_2$. Let $t \to x_1(t)$ and $t \to x_2(t)$ denote the trajectories corresponding to u_1 and u_2. Then

$$y_1(t) = \mathbf{K} \cdot x_1(t) \equiv \mathbf{K} \cdot x_2(t) = y_2(t)$$

and

$$\mathbf{K} \cdot x_1(t)x_2(t)^{-1} = \mathbf{K} \quad \text{for all } t \geqq 0.$$

In particular the curve

$$t \to a(t) = x_1(t)x_2(t)^{-1}$$

is contained in the Lie subgroup $\mathbf{K}$ for all $t \geqq 0$, and for each positive time t, the derivative $\dot{a}(t)$ is contained in the tangent space to $\mathbf{K}$ at $a(t)$, $\mathscr{K}(a(t))$. The following lemma establishes some of the basic properties of $a(t)$ and $\dot{a}(t)$.

LEMMA 3.2. *Consider the right-invariant system* (4). *Suppose that* $u_1, u_2 \in \mathscr{U}$ *are controls which give rise to trajectories* $t \to x_1(t)$ *and* $t \to x_2(t)$ *respectively. The curve*

$$t \to a(t) = x_1(t)x_2(t)^{-1}$$

is contained in $\mathbf{G}_0$ *and*

$$\dot{a}(t) = (A + u_1(t)B)(a(t)) + dL_{a(t)}(A + u_2(t)B)(e)$$

where e is the identity element of $\mathbf{G}_0$. *In particular*

$$dR_{a(t)^{-1}}(\dot{a}(t)) = (A + u_1(t)B)(e) - \mathrm{Ad}\,(a(t))(A + u_2(t)B)(e)$$

is contained in $\mathscr{L}_0(e)$ *for all real t.*

Proof. Let $t \to x_1(t)$ and $t \to x_2(t)$ be smooth trajectories corresponding to controls u_1 and u_2. By Theorem 2.1 $x_1(t), x_2(t) \in (\exp tA) \cdot \mathbf{G}_0 \cdot x_0$ for all $t \in \mathbf{R}$. It follows that $x_i(t) = (\exp tA) \cdot P_i(t) \cdot x_0$ where $P_i(t)$ is a smooth curve in $\mathbf{G}_0$, for $i = 1,2$. Thus

$$a(t) = x_1(t)x_2(t)^{-1} = (\exp tA)P_1(t)P_2(t)^{-1} \exp(-tA),$$

and since $\exp tA \in \mathbf{G}$, $P_1(t)P_2(t)^{-1} \in \mathbf{G}_0$, and $\mathbf{G}_0$ is a normal subgroup of $\mathbf{G}$, we see that $a(t) \in \mathbf{G}_0$ for all $t \in \mathbf{R}$.

The product rule for differentiation implies that

$$\dot{a}(t) = \frac{d}{dt}(x_1(t)x_2(t)^{-1}) = \frac{\partial}{\partial t}x_1(t)x_2(s)^{-1}\Big|_{s=t} + \frac{\partial}{\partial s}x_1(t)x_2(s)^{-1}\Big|_{s=t}.$$

Since A and B are right-invariant vector fields and $\dot{x}_i(t) = (A + u_i(t)B)x_i(t)$ for $i = 1, 2$, we have

$$\frac{\partial}{\partial t} x_1(t)x_2^{-1}(s)\Big|_{s=t} = (A + u_1(t)B)(x_1(t)x_2(t)^{-1}) = (A + u_1(t)B)(a(t)).$$

To obtain an expression for $\dot{x}_2(s)^{-1}$ we observe that $x_2(s)x_2(s)^{-1} = e$ for all $s \in \mathbf{R}$. Differentiating both sides of this equality results in the equation

$$dL_{x_2(t)}(\dot{x}_2(t)^{-1}) + dR_{x_2(t)^{-1}}\dot{x}_2(t) = 0$$

or

$$(x_2(\dot{t})^{-1}) = -dL_{x_2(t)^{-1}} \cdot dR_{x_2(t)^{-1}}\dot{x}_2(t) = -dL_{x_2(t)^{-1}}(A + u_2(t)B)(e).$$

Thus

$$\frac{\partial}{\partial s} x_1(t)x_2^{-1}(s) = dL_{x_1(t)} \circ (-dL_{x_2(s)^{-1}}(A + u_2(s)B)(e))$$

and using the chain rule we conclude that

$$\dot{a}(t) = (A + u_1(t)B)(a(t)) - dL_{a(t)}(A + u_2(t)B)(e).$$

To complete the proof we identify $\mathscr{L}_0(g)$ with $T_e(\mathbf{G}_0)$ for all $g \in \mathbf{G}_0$. Then $\dot{a}(t)$ is identified with $dR_{a(t)^{-1}}(\dot{a}(t))$ for all real t. We observe that the mapping $C_x : g \to xgx^{-1}$ of $\mathbf{G}_0 \to \mathbf{G}_0$ can be written as the composition $R_{x^{-1}} \circ L_x$ and thus $dR_{x^{-1}} \circ dL_x = dC_x = \mathrm{Ad}\,(x)$ and $dR_{a(t)^{-1}}(\dot{a}(t)) = dR_{a(t)^{-1}}(A + u_1(t)B)(a(t)) - dR_{a(t)^{-1}} \circ dL_{a(t)}(A + u_2(t)B)(e) = (A + u_1(t)B(e) - \mathrm{Ad}\,(a(t))(A + u_2(t)B)(e)$. This completes the proof.

We have observed that if system (1) fails to be invertible then the curve $t \to a(t)$ is contained in the Lie Group $\mathbf{K}$ and $\dot{a}(t) \in \mathscr{K}(a(t))$. Thus $dR_{a(t)^{-1}}(\dot{a}(t)) \in \mathscr{K}(e)$ for all t in $\mathbf{R}$ and if we set

$$Q(t) = (A + u_1(t)B)(e) - \mathrm{Ad}\,(a(t))(A + u_2(t)B)(e)$$

then the curve $t \to Q(t)$ is contained in $\mathscr{K}(e)$ by Lemma 3.2. If we identify $\mathscr{K}(e)$ with $\mathscr{K}$ then $Q(t)$ and its derivatives with respect to t of all orders are contained in $\mathscr{K}$. In particular $(d^nQ(t)/dt^n)|_{t=0} = Q^{(n)}(0) \in \mathscr{K}$ for $n = 0, 1, \cdots$ and the Lie algebra generated by these tangent vectors is contained in $\mathscr{K}$.

In proving Theorem 3.1 we will show that the Lie algebra generated by the derivatives $Q(0), Q^{(1)}(0), Q^{(2)}(0), \cdots$ is the Lie algebra $\mathscr{L}_0$ of $\mathbf{G}_0$. Thus a sufficient condition for invertibility is that $\mathscr{L}_0 \neq \mathscr{K}$. The next lemma examines the relationship between the curve. $t \to a(t)$ and the Lie algebra $\mathscr{L}_0$.

LEMMA 3.3. *Consider the right-invariant system* (4). *Suppose that* $u_1, u_2 \in \mathscr{U}$ *are distinct controls and* $t \to x_1(t)$, $t \to x_2(t)$ *are the corresponding trajectories. Then there exists* $\varepsilon > 0$ *and a real analytic curve* $t \to L(t)$ *in* $\mathscr{L}_0$, *defined for* $|t| < \varepsilon$, *such that* $x_1(t)x_2(t)^{-1} = \exp L(t)$ *for* $|t| < \varepsilon$ *and* $\mathscr{L}_0 = \{L(t) : |t| < \varepsilon\}_{LA}$.

Proof. The curve $t \to a(t) = x_1(t)x_2(t)^{-1}$ is contained in $\mathbf{G}_0$ as a consequence of Lemma 3.2. It is well known that exp: $\mathscr{L}_0 \to \mathbf{G}_0$ is a local diffeomorphism in some neighborhood $\mathscr{N}$ of 0 in $\mathscr{L}_0$ [4], [10]. Thus there exists an $\varepsilon > 0$ and a real analytic curve $t \to L(t)$ in $\mathscr{L}_0$ such that $\exp L(t) = a(t)$ and $L(t)$ has a Taylor series expansion $L(t) = \sum_{i=0}^{\infty} t^i L_i$ for $|t| < \varepsilon$. Since $a(0) = x_1(0)x_2(0)^{-1} = e\,e^{-1} = e$ and

$a(0) = \exp L(0)$, we have $L(0) = 0$ and $L_0 = 0$. Clearly $L_1, L_2, \cdots \in \mathscr{L}_0$ and $\{L(t): |t| < \varepsilon\}_{\mathrm{LA}} = \{L_i : i = 1, 2, \cdots\}_{\mathrm{LA}}$. We will set $\hat{\mathscr{L}}_0 = \{L_i : 1, 2, \cdots\}_{\mathrm{LA}}$. It follows that $\hat{\mathscr{L}}_0 \subset \mathscr{L}_0$ and the proof will be complete if we can show that $\hat{\mathscr{L}}_0 = \mathscr{L}_0$.

Set $Q(t) = (A + u_1(t)B)(e) - \mathrm{Ad}\,(a(t))(A + u_2(t)B)(e)$ for all real t. Lemma 3.2 asserts that

$$Q(t) = dR_{a(t)^{-1}}(\dot{a}(t)) = dR_{a(t)^{-1}}\left(\frac{d}{dt}\exp L(t)\right).$$

We will use this equality to study $\hat{\mathscr{L}}_0$. We begin by noting that $\mathrm{Ad}\,(a(t)) = \mathrm{Ad}\,(\exp L(t))$. Using the Campbell–Baker–Hausdorff formula we have

$$\mathrm{Ad}\,(a(t))(A + u_2(t)B)(e) = \sum_{k=0}^{\infty} ((-1)^k/k!)\,\mathrm{ad}^k_{L(t)}(A + u_2(t)B)(e),$$

and so

$$Q(t) = (u_1(t) - u_2(t))B(e) - \sum_{k=1}^{\infty} ((-1)^k/k!)\,\mathrm{ad}^k_{L(t)}(A + u_2(t)B)(e).$$

Choosing ε smaller if necessary we can assume that $u_1(t) = \sum_{i=0}^{\infty} a_i t^i$ and $u_2(t) = \sum_{i=0}^{\infty} b_i t^i$ for $|t| < \varepsilon$. Setting $c_i = a_i - b_i$ we have

$$Q(t) = \sum_{i=0}^{\infty} c_i t^i B(e) - \sum_{k=1}^{\infty} ((-1)^k/k!)\,\mathrm{ad}^k_{L(t)}A(e)$$

$$- \sum_{k=1}^{\infty}\sum_{i=0}^{\infty} ((-1)^k/k!)b_i t^i\,\mathrm{ad}^k_{L(t)}B(e).$$

Expressing $L(t)$ as $\sum_{j=1}^{\infty} t^j L_j$ we can collect like powers of t and write

$$Q(t) = c_0 B(e) + \sum_{k=1}^{\infty} t^k F_k(e),$$

where $F_k \in \mathscr{L}_0$ for all k. A straightforward induction argument shows that

$$F_k = c_k B_k + (-1)^k\,\mathrm{ad}_A L_k + R_k + S_k \tag{5}$$

where R_k is a linear combination of terms of the form $\mathrm{ad}_{L_{k_1}}\mathrm{ad}_{L_{k_2}}\cdots\mathrm{ad}_{L_{k_p}}A$ with $p \geqq 1$, $k_i < k$ for $i = 1, 2, \cdots, p$ and S_k is a linear combination of terms of the form $\mathrm{ad}_{L_{k_1}}\cdots\mathrm{ad}_{L_{k_q}}B$ with $q \geqq 1$, $k_i \leqq k$ for $i = 1, 2, \cdots, q$.

A second expression for $Q(t)$ comes from the identity

$$Q(t) = dR_{a(t)^{-1}}\left(\frac{d}{dt}\exp L(t)\right).$$

Using the formula for dexp and the Taylor series expansion for $L(t)$ it is easy to verify that

$$Q(t) = L_1(e) + \sum_{k=1}^{\infty} t^k((k+1)L_{k+1}(e) + M_k(e))$$

where M_k is contained in $\{L_1, \cdots, L_k\}_{\mathrm{LA}}$. Combining these two expressions for $Q(t)$ we find that $L_1 = c_0 B$ and

$$(k+1)L_{k+1} + M_k = F_k \tag{6}$$

for $k = 1, 2, \cdots$. To complete the proof we will use these relations to show that $\mathrm{ad}_A^k B \in \hat{\mathscr{L}}_0$ for $k = 0, 1, \cdots$, which implies that $\hat{\mathscr{L}}_0 = \mathscr{L}_0$.

Since $u_1 \neq u_2$, $c_k \neq 0$ for some k. Let p be the smallest positive integer k such that $c_k \neq 0$.

Claim. $L_1, L_2, \cdots, L_p = 0$ *and* $(p+1)L_{p+1} = c_p B$: If $p = 0$ then this is the case, since $L_1 = c_0 B$. If $p > 0$ then $c_0 = 0$ and $L_1 = 0$. Suppose that $L_1, L_2, \cdots, L_k = 0$ for $0 \leqq k < p$. Combining (5) and (6) we find that $F_k = (k+1)L_{k+1} + M_k = c_k B + (-1)^k \mathrm{ad}_A L_k + R_k + S_k$. Since $c_1 = c_2 = \cdots = c_k = 0$ and $L_1, L_2, \cdots, L_k = 0$ it follows from the definitions that $M_k = c_k B = \mathrm{ad}_A L_k = R_k = S_k = 0$. This induction argument proves that $L_1, L_2, \cdots, L_p = 0$ and hence $M_p = \mathrm{ad}_A L_p = R_p = S_p = 0$. Thus $(p+1)L_{p+1} = F_p = c_p B_p$, which proves the assertion.

Claim. $\hat{\mathscr{L}}_0$ *is an* ad_A*-invariant subspace of* $\mathscr{L}_0$: Let p be chosen as above. Then $L_1 = L_2 = \cdots = L_p = 0$ and it suffices to show that $(-1)^k \mathrm{ad}_A L_k - (k+1)L_{k+1} \in \{L_i : p < i \leqq k\}_{\mathrm{LA}}$ for all $k > p$, since this implies that $\mathrm{ad}_A L_k \in \hat{\mathscr{L}}_0$ for all k. The proof uses induction on k. We have shown that $(p+1)L_{p+1} = c_p B$, $(p+2)L_{p+2} + M_{p+1} = F_{p+1}$ from (6), and $F_{p+1} = c_{p+1} B + (-1)^{p+1} \mathrm{ad}_A L_{p+1} + R_{p+1} + S_{p+1}$ from (5). Since $L_1 = \cdots = L_p = 0$, we have $R_{p+1} = S_{p+1} = 0$ and since $M_{p+1} \in \{L_1, \cdots, L_{p+1}\}_{\mathrm{LA}}$, $M_{p+1} = \alpha B$ for some real number α. Combining these results we have

$$(p+2)L_{p+2} = (-1)^{p+1}(c_p/(p+1))\,\mathrm{ad}_A B + (c_{p+1} - \alpha)B.$$

If $k = p+1$ then $(-1)^k \mathrm{ad}_A L_k - (k+1)L_{k+1} = (-1)^k \mathrm{ad}_A((c_{k-1}/k)B) - (-1)^k (c_{k-1}/k)\,\mathrm{ad}_A B - (c_k - \alpha)B = (\alpha - c_k)B \in \{L_k\}_{\mathrm{LA}}$. Now assume that $(-1)^k \mathrm{ad}_A L_k - (k+1)L_{k+1} \in \{L_i : p < i \leqq k\}_{\mathrm{LA}}$ for $p < k < n$ where n is a positive integer greater than $p+1$. For $k = n$ we have

$$(n+1)L_{n+1} = F_n - M_n = c_n B + (-1)^n \mathrm{ad}_A L_n + R_n + S_n - M_n$$

from (5) and (6). Thus

$$(-1)^n \mathrm{ad}_A L_n - (n+1)L_{n+1} = M_n - c_n B - R_n - S_n$$

and the induction will be completed if we can show that the right hand side of the above inequality is contained in $\{L_{p+1}, L_{p+2}, \cdots, L_n\}_{\mathrm{LA}}$. Set $\mathscr{K} = \{L_{p+1}, \cdots, L_n\}_{\mathrm{LA}}$. Now $M_n \in \mathscr{K}$ by definition, and since $L_{p+1} = c_p B \in \mathscr{K}$ and $c_p \neq 0$, both B and $c_n B$ are in $\mathscr{K}$. Recall that S_n is a linear combination of terms of the form $\mathrm{ad}_{L_{k_1}} \mathrm{ad}_{L_{k_2}} \cdots \mathrm{ad}_{L_{K_q}} B$ where $k_i \leqq n$, hence $S_n \in \mathscr{K}$. Finally, R_n is a linear combination of terms of the form $\mathrm{ad}_{L_{k_1}} \cdots \mathrm{ad}_{L_{k_q}} A$ where $k_q < n$. By the induction hypothesis $(-1)^n \mathrm{ad}_A L_{n-1} = (-1)^{n-1} \mathrm{ad}_{L_{n-1}} A = -nL_n + L$ where $L \in \{L_{p+1}, \cdots, L_{n-1}\}_{\mathrm{LA}}$ and it follows that R_n is also contained in $\mathscr{K}$. This completes the induction.

Since $c_{p_k} \neq 0$, $B \in \hat{\mathscr{L}}_0$, and $\hat{\mathscr{L}}_0$ is an ad_A-invariant subspace of $\mathscr{L}_0$, $\mathrm{ad}_A^k \in \hat{\mathscr{L}}_0$ for all $k \geqq 0$, which completes the proof of this lemma.

Proof (Theorem 3.1). First we suppose that the system (4) is invertible but $\mathrm{ad}_A^k B \in \mathscr{K}$ for $k = 0, 1, \cdots$. For each control $u \in \mathscr{U}$ the corresponding trajectory is $x(t) = \exp tA \cdot P(t) \cdot x_0$, where $P(t) \in \mathbf{G}_0$, as a consequence of Theorem 2.1. Since

$\mathscr{L}_0$ is the Lie algebra generated by $\{\mathrm{ad}_A^k B: k = 0, 1, \cdots\}$, a subset of the Lie algebra $\mathscr{K}$, we are assuming that $\mathscr{L}_0 \subset \mathscr{K}$, and thus $\mathbf{G}_0 \subset \mathbf{K}$. If $u_1, u_2 \in \mathscr{U}$ are two controls producing trajectories $x_1(t) = \exp tA \cdot P_1(t) \cdot x_0$ and $x_2(t) = \exp tA \cdot P_2(t) \cdot x_0$, then $x_1(t)x_2(t)^{-1} = \exp tA \cdot P_1(t) \cdot x_0 x_0^{-1} P_2^{-1}(t)(\exp tA)^{-1} = (\exp tA)(P_1(t)P_2^{-1}(t))(\exp tA)^{-1}$. Since $\exp tA \in \mathbf{G}$, $P_1(t)P_2^{-1}(t) \in \mathbf{G}_0$, and $\mathbf{G}_0$ is a normal subgroup of $\mathbf{G}$, $x_1(t)x_2(t)^{-1} \in \mathbf{G}_0 \subset \mathbf{K}$. Thus $\mathbf{K}x_1(t)x_2(t)^{-1} = \mathbf{K}$ and $\mathbf{K}x_1(t) = \mathbf{K}x_2(t)$ for all $t > 0$. In other words $u_1(t)$ and $u_2(t)$ produce the same outputs. Clearly this system is not invertible, a contradiction. Thus invertibility implies that $\mathrm{ad}_A^k B \notin \mathscr{K}$ for some positive integer k. Since ad_A is a linear operator on the n dimensional Lie algebra $\mathscr{L}$, a necessary condition for invertibility is that $\mathrm{ad}_A^k B \notin \mathscr{K}$ for some $k \in \{0, 1, \cdots, n-1\}$, by the Cayley–Hamilton theorem.

To show that this Lie algebraic condition implies invertibility it suffices to show that if two different controls result in the same output then $\mathscr{L}_0 \subset \mathscr{K}$. Suppose that $u_1, u_2 \in \mathscr{U}$ are distinct controls producing the same outputs, $y_1(t) = \mathbf{K} \cdot x_1(t) = y_2(t) = \mathbf{K} \cdot x_2(t)$. Then for t sufficiently small the real analytic curve $t \to a(t) = x_1(t)x_2(t)^{-1}$ is contained in $\mathbf{K}$. From Lemma 3.3 we know that there exists an $\varepsilon > 0$ and a real analytic curve $t \to L(t)$ in $\mathscr{L}_0$ such that $a(t) = \exp L(t)$ for $|t| < \varepsilon$ and $\hat{\mathscr{L}}_0 = \{L(t): |t| < \varepsilon\} = \mathscr{L}_0$. Shrinking ε if necessary we can express $L(t)$ by the Taylor expansion $L(t) = \sum_{i=1}^{\infty} t^i L_i$ where $\{L_i = i = 1, 2, \cdots\}_{\mathrm{LA}} = \hat{\mathscr{L}}_0$. Since $a(t) = \exp L(t) \in \mathbf{K}$ for $|t| < \varepsilon$ we know that $L(t) \in \mathscr{K}$ for $|t| < \infty$ (cf. [4] or [10]). Thus $\{L_i: i = 1, 2, \cdots\}_{\mathrm{LA}} \subset \mathscr{K}$ and $\hat{\mathscr{L}}_0 = \mathscr{L}_0 \subset \mathscr{K}$. This completes the proof.

Proof (Corollary 1). Let $\mathbf{K}$ be the kernel of c. Then $\mathscr{K}$ is the kernel of c_* and $X \in \mathscr{K}$ if and only if $c_* X = 0$. Since the outputs $y_1(t) = c(x_1(t))$ and $y_2(t) = c(x_2(t))$ are the same if and only if the coset outputs $\mathbf{K} \cdot x_1(t)$ and $\mathbf{K} \cdot x_2(t)$ agree, Theorem 3.1 applies. Thus the system is invertible if and only if $\mathrm{ad}_A^k B \notin \mathscr{K}$ for some $k \geqq 0$, or $c^* \mathrm{ad}_A^k B \neq 0$. This completes the proof.

Proof (Corollary 2). Set $\mathbf{K} = \{M: M \in GL(n, \mathbf{R})$ and $CM = C\}$. Then two outputs $CX_1(t)$ and $CX_2(t)$ agree if and only if $X_1(t)X_2^{-1}(t) \in \mathbf{K}$ for all $t \geqq 0$, or $\mathbf{K} \cdot X_1(t) = \mathbf{K} \cdot X_2(t)$. Theorem 3.1 asserts that the system is not invertible if and only if $\mathrm{ad}_A^k B \in \mathscr{K}$ for all $k \geqq 0$. Since $\mathscr{K} = \{X: CX = 0$ and $X \in gl(n, \mathbf{R})\}$, the proof is complete.

Proof (Corollary 3). If the system (2) is not invertible then $C \mathrm{ad}_A^k B = 0$ for $k = 0, 1, \cdots$ by Corollary 2. Any trajectory $X(t) = P(t) \cdot \exp tA$ where $P(t) \in \mathbf{G}_0$. Here $C\mathscr{L}_0 = \{0\}$ so $CP(t) = C$ for all t and $Y(t) = CX(t) = C \exp tA$. Thus every control produces the same response. If different controls result in different responses then $C\mathscr{L} \neq \{0\}$, so $C \mathrm{ad}_A^k B \neq 0$ for some k, and the system is invertible, by Corollary 2. This completes the proof.

4. Left-inverses for bilinear systems. Suppose that a given control system is invertible—that is, the output uniquely determines the control. One then faces the practical problem of determining the input given only the output record of the system. In the linear case this problem has been solved in a very elegant manner. A second linear system, called a left-inverse system, can be constructed. This left-inverse system, when driven by appropriate derivatives of the output of the original system, produces as an output $u(t)$, the input to the origunal system [1], [2], [7], [8]. In this section we will construct nonlinear systems which are left-inverses for bilinear systems.

Consider the bilinear system (1). As in the matrix case the *relative order*, α, of this bilinear system the least positive integer k such that $c\ \mathrm{ad}_A^{k-1}B \neq 0$ or $\alpha = \infty$ if $c\ \mathrm{ad}_A^k B = 0$ for all $k > 0$.

In contrast with the linear case it is not yet known whether or not an invertible bilinear system has a bilinear left-inverse system. We will look for a left-inverse in the class of nonlinear systems of the form

$$\begin{aligned} &\hat{x}(t) = a(\hat{x}(t)) + \hat{u}(t)b(\hat{x}(t)), \qquad \hat{x}(0) = \hat{x}_0 \in M, \\ &\hat{y}(t) = d(\hat{x}(t)) + \hat{u}(t)e(\hat{u}(t)) \end{aligned} \tag{7}$$

where $\hat{x} \in M$, a differentiable manifold, $\hat{u} \in \mathcal{U}$, $a(\cdot)$ and $b(\cdot)$ are smooth vector fields on M, and d,e are smooth functions on M.

DEFINITION. The system (7) is called a *left-inverse* for the bilinear system (1) if $\hat{u}(t) = y^{(\alpha)}(t)$ implies that $\hat{y}(t) = u(t)$. The following theorem generalizes the well known linear result on left-inverses [2] to the bilinear case:

THEOREM 4.1. *If the bilinear system* (1) *is invertible then its relative order* $\alpha < \infty$. *If* $\alpha < \infty$ *and* $c\ \mathrm{ad}_A^{\alpha-1}Bx_0 \neq 0$ *then the bilinear system is invertible with left-inverse* (7), *where* $M = \mathbf{R}^n \sim (cA^{\alpha-1}B)^{\perp}$, $\hat{x}_0 = x_0$, *and*

$$\begin{aligned} a(\hat{x}) &= A\hat{x} - (cA^{\alpha}\hat{x}/cA^{\alpha-1}B\hat{x})B\hat{x}, \\ b(\hat{x}) &= (1/cA^{\alpha-1}B\hat{x})B\hat{x}, \\ d(\hat{x}) &= -(cA^{\alpha}\hat{x}/cA^{\alpha-1}B\hat{x}) \end{aligned}$$

and

$$e(\hat{x}) = (1/cA^{\alpha-1}B\hat{x}).$$

If $\hat{u}(t) = y^{(\alpha)}(t)$ *then* $\hat{y}(t) = u(t)$.

Proof. We will begin by showing that an invertible bilinear system has relative order $\alpha < \infty$. If α is infinite then the corresponding matrix bilinear system (2) is not invertible by Theorem 3.1, Corollary 2. Choose distinct controls u_1, $u_2 \in \mathcal{U}$ which produce identical outputs for the matrix bilinear system. Since the output of the bilinear system (1) is $t \to Y(t)x_0$, where $t \to Y(t)$ is the output of the corresponding matrix system, the bilinear system (1) is not invertible. This completes the first part of the proof.

Suppose that $\alpha < \infty$ and $c\ \mathrm{ad}_A^{\alpha-1}Bx_0 \neq 0$. Differentiate the output $y(t) = cx(t)$ to obtain

$$\dot{y}(t) = c\dot{x}(t) = c\,Ax(t) + u(t)cBx(t).$$

If $\alpha > 1$ then $cB = 0$, and differentiating $\dot{y}(t)$ we find that

$$y^{(2)}(t) = cA\dot{x}(t) = cA^2x(t) + u(t)cABx(t).$$

If $\alpha > 2$ then $c\ \mathrm{ad}_A B = 0$ and so $cAB - cBA = 0$. Since $cB = 0$ we have $cAB = 0$ and

$$y^{(3)}(t) = cA^3x(t) + u(t)cA^2Bx(t).$$

Continuing this procedure we find that

$$y^{(\alpha)}(t) = cA^{\alpha}x(t) + u(t)cA^{\alpha-1}Bx(t). \tag{8}$$

Since $cA^{\alpha-1}Bx_0 \neq 0$ by assumption, the scalar function $cA^{\alpha-1}Bx(t)$ is nonzero for t sufficiently small. The set of vectors x in $\mathbf{R}^n$ for which $cA^{\alpha-1}Bx \neq 0$ is the differentiable manifold $M = \mathbf{R}^n \sim (cA^{\alpha-1}B)^{\perp}$. Consider the nonlinear system (7) described in the statement of this theorem, and set $\hat{u}(t) = y^{(\alpha)}(t)$. Then

$$\dot{\hat{x}}(t) = a(\hat{x}(t)) + (cA^{\alpha}x(t) + u(t)cA^{\alpha-1}Bx(t))b(\hat{x}); \qquad \hat{x}(0) = x_0.$$

Claim. $\hat{x}(t) = x(t)$: Since $\hat{x}(0) = x(0)$ it suffices to verify that both $\hat{x}(t)$ and $x(t)$ solve the same differential equation. Replacing $\hat{x}$ by x in the above differential equation in $\hat{x}$, and invoking the definitions for $a(\cdot)$ and $b(\cdot)$, this equation reduces to the differential equation

$$\dot{x}(t) = Ax(t) + u(t)Bx(t).$$

Thus $\hat{x}$ and x satisfy the same differential equation when $\hat{u}(t) = y^{(\alpha)}(t)$.

The corresponding output is

$$\begin{aligned}\hat{y}(t) &= d(x(t)) + \hat{u}(t)e(x(t)) \\ &= -(cA^{\alpha}x(t)/cA^{\alpha-1}Bx(t)) + y^{(\alpha)}(t)(1/cA^{\alpha-1}Bx(t)).\end{aligned}$$

Substituting the expression (8) for $y^{(\alpha)}(t)$ we find that $\hat{y}(t) = u(t)$. Since $x(t)$ involves in M for some interval of time and the controls are piecewise real analytic functions, $u(t)$ is completely determined for all $t > 0$. Thus the bilinear system is invertible and the given nonlinear is a left-inverse system. This completes the proof.

We remark that in the proof of this theorem we show that when $\hat{u}(t) = y^{(\alpha)}(t)$ the state $\hat{x}(t) = x(t)$, the state of the original bilinear system. Thus the left-inverse system acts as a state observer for the bilinear system, a result which itself is of some interest.

We also note that for certain bilinear systems the vector fields $a(x)$, $b(x)$ may not be complete. That is, the integral curves for these vector fields need not be defined for all time. Thus after a finite time has passed the trajectory $x(t)$ could leave M, and in this case $u(t)$ would only be recovered for t in some bounded interval. For a linear system in bilinear form $y(t)$ is defined for all t and the left-inverse system reduces to the standard linear left-inverse (see Example 1).

Theorem 4.1 presents a sufficient condition for inverting vector bilinear systems in case where $\alpha < \infty$ but this condition is far from being necessary. In Example 3, $c\ \mathrm{ad}_A^{\alpha-1}Bx_0 = 0$ but $c\ \mathrm{ad}_A^{\alpha}Bx_0 \neq 0$ and the system is invertible. It seems reasonable to expect that a necessary and sufficient for invertibility must take into account the action of the matrix Lie group $\mathbf{G}$ on the state space $\mathbf{R}^n$.

DEFINITION. The *initialized relative order*, $\alpha(x_0)$, for a bilinear system (1) is the least positive integer k such that $c\ \mathrm{ad}_A^{k-1}Bx_0 \neq 0$ or $\alpha(x_0) = \infty$ if $c\ \mathrm{ad}_A^{k}Bx_0 = 0$ for $k = 0, 1, 2, \cdots, n^2 - 1$.

Note that $\alpha \leqq \alpha(x_0)$ and one could have $\alpha < \infty$ with $\alpha(x_0)$ infinite (see Example 2 with $x_0 = (0, 1, 0, 0)$).

THEOREM 4.2. *Consider the bilinear system* (1) *with associated Lie algebras* $\mathscr{L} \supset \mathscr{L}_0 \supset \mathscr{B}$. *If* $\alpha(x_0) < \infty$ *and*

$$\alpha(x) \geqq \alpha(x_0) \quad \textit{for } x \in \mathbf{G} \cdot x_0 \tag{9}$$

then the system is invertible with left-inverse (7), *where*

$$\hat{x}_0 = x_0, \qquad M = \mathbf{R}^n \sim (cA^{\alpha(x_0)-1}B)^{\perp},$$
$$a(\hat{x}) = A\hat{x} - (cA^{\alpha(x_0)}\hat{x}/cA^{\alpha(x_0)-1}B\hat{x})B\hat{x},$$
$$b(\hat{x}) = (1/cA^{\alpha(x_0)-1}B\hat{x})B\hat{x},$$
$$d(\hat{x}) = -(cA^{\alpha(x_0)}\hat{x}/cA^{\alpha(x_0)-1}B\hat{x}),$$

and

$$e(\hat{x}) = (1/cA^{\alpha(x_0)-1}B\hat{x}).$$

If $\hat{u}(t) = y^{(\alpha(x_0))}(t)$ *then* $\hat{y}(t) = u(t)$.

Remarks. The condition (9) is automatically satisfied if $\alpha = \alpha(x_0)$, since $c\ \mathrm{ad}_A^k B = 0$ for $k < \alpha(x_0) - 1$.

Proof. Suppose $\alpha(x_0) < \infty$ and condition (9) is satisfied. Condition (9) implies that $c\ \mathrm{ad}_A^k Bx = 0$ for $0 \leqq k < \alpha(x_0) - 1$ and $x \in \mathbf{G} \cdot x_0$. In particular $cBx = 0$ and $c\ \mathrm{ad}_A Bx = c(BA - AB)x = cBAx - cABx = 0$ for $\alpha(x_0) > 2$. Now (9) implies that $cB(\exp tA)x = 0$ for all real t, as $\exp tA \in \mathbf{G}$. Differentiating with respect to t and setting $t = 0$ shows that $cBAx = 0$. Combining this with the above expression for $c\ \mathrm{ad}_A Bx$ we see that $cABx = 0$. A similar argument proves that $cA^k Bx = 0$ for $0 \leqq k \leqq \alpha(x_0) - 1$ and for all $x \in \mathbf{G} \cdot x_0$. In particular, if $x(t)$ is the trajectory for the system (1) with $x(0) = x_0$, then $cA^k Bx(t) \equiv 0$ for $0 \leqq k < \alpha(x_0) - 1$ and

$$y^{(\alpha(x_0))}(t) = cA^{\alpha(x_0)}x(t) + u(t)cA^{\alpha(x_0)-1}Bx(t).$$

In the proof of Theorem 4.1 we showed that this implies that $\hat{y}(t) = u(t)$ when $\hat{u}(t) = y^{(\alpha(x_0))}(t)$. This completes the proof.

Example 1. In this example we apply Theorem 4.1 to a linear system in bilinear form. The nonlinear left-inverse reduces to the standard linear left-inverse described in [2].

Consider the linear system $\dot{x}(t) = Ax(t) + bu(t)$; $x(0) = x_0$, with output $y(t) = cx(t)$. In bilinear form

$$\dot{z}(t) = Fz(t) + u(t)Gz(t); \qquad z(0) = z_0,$$
$$y(t) = Hz(t)$$

where $z(t) = (x(t), 1)$, $z(0) = (x_0, 1)$,

$$F = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}, \quad G = \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix}, \mathrm{H} = (c \quad 0).$$

Here $H\ \mathrm{ad}_F^{\alpha-1}G = (0 \quad cA^{\alpha-1}b)$ and $H\ \mathrm{ad}_F^{\alpha-1}Gz_0 = cA^{\alpha-1}b$. Thus Theorem 4.1 asserts that a linear system is invertible if and only if $\alpha < \infty$, which is proved in [2]. If $\alpha < \infty$ the left-inverse described in Theorem 4.1 is of the form

$$\dot{\hat{z}}(t) = a(\hat{z}(t)) + \hat{u}(t)b(\hat{z}(t)); \qquad \hat{z}(0) = z(0),$$
$$\hat{y}(t) = d(\hat{z}(t)) + \hat{u}(t)e(\hat{z}(t))$$

where $\hat{z}(t) = (x(t), \alpha(t))$ with $x \in \mathbf{R}^n$ and $\alpha \in \mathbf{R}$,

$$a(\hat{z}) = (A\hat{x} - (cA^{\alpha}\hat{x}/\alpha \cdot cA^{\alpha-1}b)\alpha \cdot b, 0),$$
$$b(\hat{z}) = ((1/\alpha \cdot cA^{\alpha-1}b)\alpha \cdot b, 0)$$
$$d(\hat{z}) = -(cA^{\alpha}\hat{x}/\alpha \cdot cA^{\alpha-1}b),$$
$$e(\hat{z}) = (1/\alpha \cdot cA^{\alpha-1}b),$$

$M = \{(x, \alpha): x \in \mathbf{R}^n, \alpha \in \mathbf{R} \sim \{0\}\}$ and $z_0 = (x_0, 1)$. With z_0 given it follows that $z(t) = (x(t), 1)$ so that the above system of equations reduces to

$$\dot{\hat{x}}(t) = [A - (bcA^{\alpha}/cA^{\alpha-1}b)]\hat{x}(t) + (1/cA^{\alpha-1}b)b\hat{u}(t),$$
$$\hat{y}(t) = -(cA^{\alpha}/cA^{\alpha-1}b)\hat{x}(t) + (1/cA^{\alpha-1}b)\hat{u}(t),$$

which is the well known linear left-inverse (cf. [2]).

Example 2. In this example we describe a matrix bilinear system which satisfies the hypotheses of Theorem 3.1 Corollary 2. The corresponding vector bilinear system illustrates the construction of the nonlinear left-inverse described in Theorem 4.1.

Consider the matrix bilinear system

$$\dot{X}(t) = AX(t) + u(t)BX(t); \qquad X(0) = I,$$
$$Y(t) = CX(t)$$

with

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix},$$

and $C = (1 \quad 1 \quad 0 \quad 0)$. By direct computation we find that

$$[A, B] = BA - AB = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$[B, \mathrm{ad}_A B] = [A, B]B - B[A, B] = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathrm{ad}_A^2 B = -\mathrm{ad}_A B, \quad \mathrm{ad}_A[B, \mathrm{ad}_A B] = -[B, \mathrm{ad}_A B],$$

and

$$\mathrm{ad}_B^2 \, \mathrm{ad}_A B = [B, \mathrm{ad}_A B].$$

Thus

$$\mathscr{L} \quad \text{has basis} \quad \{B, A, \mathrm{ad}_A B, [B, \mathrm{ad}_A B]\},$$
$$\mathscr{L}_0 \quad \text{has basis} \quad \{B, \mathrm{ad}_A B, [B, \mathrm{ad}_A B]\}$$

and

$$\mathscr{B} \quad \text{has basis} \quad \{B\}.$$

Here $CB = 0$, $C\,\mathrm{ad}_A B = (1, 0, 0, -1)$, hence the relative order $\alpha = 2$, and this system is invertible by Corollary 2 of Theorem 3.1.

Now we consider the corresponding bilinear system

$$\dot{x}(t) = Ax(t) + u(t)Bx(t); \qquad x(0) = x_0,$$
$$y(t) = cx(t)$$

with A, B, c defined above and $x_0 = (1, 0, 0, 0)$. Since $c\,\mathrm{ad}_A Bx \neq 0$ Theorem 4.1 asserts that this system is invertible with left-inverse

$$\dot{\hat{x}} = a(\hat{x}) + \hat{u}b(\hat{x}); \qquad \hat{x}_0 = x_0$$
$$\hat{y} = d(\hat{x}) + \hat{u}e(\hat{x})$$

where

$$a(\hat{x}) = a(\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4) = (\hat{x}_2, 0, 0, \hat{x}_2\hat{x}_3/(\hat{x}_4 - \hat{x}_1)),$$
$$b(x) = (-1, 1, 0, \hat{x}_3/(\hat{x}_1 - \hat{x}_4)),$$
$$d(\hat{x}) = \hat{x}_2/(\hat{x}_1 - \hat{x}_4),$$
$$e(\hat{x}) = 1/(\hat{x}_4 - \hat{x}_1).$$

According to Theorem 4.1 we have $\hat{y}(t) = u(t)$ if $\hat{u}(t) = y^{(2)}(t)$. We now verify this fact directly. We know that

$$\dot{y}(t) = c\dot{x}(t) = cAx(t) + u(t)cBx(t) = cAx(t),$$

since $cB = 0$, and

$$y^{(2)}(t) = cA(Ax(t) + u(t)Bx(t))$$
$$= cA^2x(t) + u(t)cABx(t) = x_2(t) + u(t)(x_4(t) - x_1(t)).$$

Thus if $\hat{u}(t) = y^{(2)}(t)$ then

$$\dot{\hat{x}} = \begin{pmatrix} \hat{x}_2 - x_2 \\ x_2 \\ 0 \\ [\hat{x}_2\hat{x}_3/(\hat{x}_4 - \hat{x}_1)] - [x_2x_3/(x_4 - x_1)] \end{pmatrix} + u \begin{pmatrix} x_1 - x_4 \\ x_4 - x_1 \\ 0 \\ -x_3 \end{pmatrix}.$$

But if we set $\hat{x}(t) = x(t)$ this equation is just

$$\dot{x}(t) = Ax(t) + u(t)Bx(t),$$

so when $\hat{u}(t) = y^{(2)}(t)$ we see that $\hat{x}(t) = x(t)$, and

$$\hat{y}(t) = x_2(t)/(x_1(t) - x_4(t)) + y^{(2)}(t)/(x_4(t) - x_1(t)) = u(t)$$

for all $t \geqq 0$.

Example 3. The following system has $\alpha = 1$ and $c\ \mathrm{ad}_A^{\alpha-1} Bx_0 = 0$, so Theorem 4.1 can't be used. Theorem 4.3 can be used to prove invertibility. Consider the bilinear system

$$\dot{x}(t) = Ax(t) + u(t)Bx(t); \qquad x(0) = x_0,$$
$$y(t) = cx(t)$$

where

$$x_0 = (0, 0, 1), \quad c = (1 \quad 0 \quad 1), \quad A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Here $cB = (1 \quad 0 \quad 0)$, $cBx_0 = 0$, and $c\ \mathrm{ad}_A Bx_0 = -1$. Thus $\alpha = 1$, $\alpha(x_0) = 2$, and Theorem 4.1 does not apply since $c\ \mathrm{ad}_A^{\alpha-1} Bx_0 = 0$.

To apply Theorem 4.3 we must check that $\alpha(x) \geqq \alpha(x_0)$ for all $x \in \mathbf{G} \cdot x_0$. In this case we must verify that $cBx = 0$ for all $x \in \mathbf{G} \cdot x_0$. By direct computation both $\exp tA$ and $\exp tB$ are matrices with first rows of the form $(b \quad 0 \quad 0)$ with b real. Since $\mathbf{G} = \{\exp t_1 A, \exp t_2 B : t_1, t_2 \text{ real}\}_G$, $\mathbf{G} \cdot x_0$ consists of vectors whose first entries are zero. This means that $cBx = (1 \quad 0 \quad 0)x = 0$ for all $x \in \mathbf{G} \cdot x_0$.

Theorem 4.3 states that this system is invertible and provides a left-inverse. Here

$$a(\hat{x}) = A\hat{x}, \quad b(\hat{x}) = \begin{pmatrix} \hat{x}_1/\hat{x}_3 \\ 1 \\ 0 \end{pmatrix}, \quad d(\hat{x}) = 0, \quad e(\hat{x}) = (1/\hat{x}_3),$$

and $M = \{(a, b, c) : a, b, c \in \mathbf{R}, c \neq 0\}$. Since condition (9) is satisfied we know that $cBx(t) \in cB\mathbf{G} \cdot x_0 = \{0\}$. Thus $\dot{y}(t) = cAx(t) + u(t)cBx(t) = cAx(t)$, $y^{(2)}(t) = y^{\alpha(x_0)-1}(t) = cA^2x(t) + u(t)cABx(t) = u(t)x_3(t)$, and when $\hat{u}(t) = y^{(2)}(t)$, $\dot{\hat{x}}(t) = \dot{x}(t)$, $\hat{x}(t) = x(t)$ and

$$\hat{y}(t) = (\hat{u}(t)/\hat{x}_3(t)) = u(t)x_3(t)/x_3(t) = u(t).$$

Of course for certain controls $u(t)$ one could have $x_3(T) = 0$ for some $T > 0$, in which case $x(T) \notin M$ and $u(t)$ is recovered for some interval $0 \leqq t < \varepsilon$ on which $x(t)$ exists.

REFERENCES

[1] R. W. Brockett and M. D. Mesarovic, *The reproducibility of multivariable systems*, J. Math. Anal. Appl., 11 (1965), pp. 548–563.

[2] R. W. Brockett, *Poles, zeros, and feedback: State space interpretation*, IEEE Trans. Automatic Control, AC-10, (1965), pp. 129–135.

[3] ———, *System theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265–284.

[4] S. Helgason, *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962.

[5] R. HIRSCHORN, *Controllibility in nonlinear systems*, J. Differential Equations, 19 (1975), pp. 46–61.
[6] M. K. SAIN AND J. L. MASSEY, *Inverses of linear sequential circuits*, IEEE Trans. Computers, C-17 (1968), pp. 330–337.
[7] ———, *Invertibility of linear time-invariant dynamical systems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 141–149.
[8] L. M. SILVERMAN, *Inversion of multivariable linear systems*, Ibid., AC-14 (1969), pp. 270–276.
[9] H. SUSSMANN AND V. J. JURDJEVIC, *Control systems on Lie groups*, J. Differential Equations, 16 (1972), pp. 313–329.
[10] F. WARNER, *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman and Co., Glenview, Il., 1970.
[11] A. S. WILLSKY, *On the invertibility of linear systems*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 272–274.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 15, No. 6, November 1977

OPTIMAL REGULATION OF NONLINEAR DYNAMICAL SYSTEMS ON A FINITE INTERVAL*

A. P. WILLEMSTEIN†

Abstract. In this paper the optimal control of nonlinear dynamical systems on a finite time interval is considered. The free end-point problem as well as the fixed end-point problem is studied. The existence of a solution is proved and a power series solution of both the problems is constructed.

1. Introduction.

We consider control processes in $\mathbb{R}^n$ of the form

$$\dot{x} = F(x, u, t) \tag{1.1}$$

and investigate the problem of finding a bounded r dimensional feedback control $u(x, t)$ which minimizes the integral

$$J(\tau, b, u) = L(x(T)) + \int_{\tau}^{T} G(x, u, t)\, dt \tag{1.2}$$

for all initial states $x(\tau) = b$ in a neighborhood of the origin in $\mathbb{R}^n$. In § 2 we treat the free end-point problem and in § 3 the fixed end-point problem. More specifically, in § 3 we require the final value $x(T)$ of the state to be zero.

For the situation where F is linear and L and G are quadratic the solution of the optimal control problem is well known (e.g. see [2, § 3.21], [3, § 2.3], [4, § 9.7] for the free end-point problem and [2, § 3.22] for the fixed end-point problem).

Here we consider the situation where the states and controls remain in a neighborhood of a fixed point (for which we without loss of generality take the origin) where the functions F, G and L can be expanded in power series. An analogous problem has been considered by D. L. Lukes [1] (see also [5, § 4.3]) for the infinite horizon case and our treatment will follow this paper to some extent, in particular as far as the free end-point case is concerned. The theory is more complete than the related Hamilton–Jacobi theory since existence and uniqueness proofs of optimal controls are given. For the solution of the fixed end-point problem we introduce a dual problem of (1.1) and (1.2) which we use to reduce the fixed end-point problem to a free end-point problem. Some examples are added to illustrate the theory.

Notation. The inner product of two vectors x and y we shall denote by $x^T y$, the length of a vector x by $|x| = \sqrt{x^T x}$, and the transposed of a matrix M by M^T. The notation $M > 0$ and $M \geqq 0$ means that M represents a (symmetric) positive definite and a nonnegative definite matrix respectively. If $f(x)$ denotes a vector function from $\mathbb{R}^n$ into $\mathbb{R}^m$, the following notation and definition of the functional

*Received by the editors October 14, 1976, and in revised form March 3, 1977.

† Department of Mathematics, Eindhoven University of Technology, Eindhoven, The Netherlands. Now at Faculty of Econometrics, Tilburg University, Tilburg, The Netherlands.

matrix will be used (in agreement with the notation of D. L. Lukes in [1]):

$$f_x := \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_1}{\partial x_n} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

Let N be a neighborhood of the origin in $\mathbb{R}^n$, V a subset of $\mathbb{R}^m$, f a vector function from $N \times V$ into $\mathbb{R}^k$, g a vector function from N into $\mathbb{R}^l$. Then we say that $f(x, y) = O(g(x))$, uniformly for $y \in V$, if the following property holds:

$$\exists_{K \in \mathbb{R}, K>0} \forall_{x \in N} \forall_{y \in V}: |f(x, y)| \leqq K|g(x)|.$$

2. Free end-point problem.

2.1. Assumptions.

(i) $F(x, u, t) = A(t)x + B(t)u + f(x, u, t)$. Here $A(t)$ and $B(t)$ are continuous real matrix functions of dimension $n \times n$ and $n \times r$ respectively. The function $f(x, u, t)$ contains the higher order terms in x and u, and is continuous with respect to t. Furthermore $f(x, u, t)$ is given as a power series in (x, u) which starts with second order terms and converges about the origin, uniformly for $t \in [\tau, T]$.

(ii) $G(x, u, t) = x^T Q(t)x + u^T R(t)u + g(x, u, t)$. Here $Q(t)$ and $R(t)$ are continuous real matrix functions of dimension $n \times n$ and $r \times r$ respectively. The function $g(x, u, t)$ contains the higher order terms in x and u, and is continuous with respect to t. Furthermore $g(x, u, t)$ is given as a power series in (x, u) which starts with third order terms and converges about the origin, uniformly for $t \in [\tau, T]$.

(iii) $L(x) = x^T Mx + l(x)$. Here M is a real matrix of dimension $n \times n$. The function $l(x)$ is given as a power series which starts with third order terms and converges about the origin.

(iv) $Q(t) \geqq 0$ and $R(t) > 0$ for $t \in [\tau, T]$; $M \geqq 0$.

We consider the class of feedback controls which are of the form

$$u(x, t) = D(t)x + h(x, t). \tag{2.1}$$

Here $D(t)$ is a continuous matrix function of dimension $r \times n$. The function $h(x, t)$ contains the higher order terms in x and is continuous with respect to t. Furthermore $h(x, t)$ is given as a power series in x which starts with second order terms and converges about the origin, uniformly for $t \in [\tau, T]$. We shall denote the class of admissible feedback controls by Ω.

DEFINITION (*optimal feedback control*). A feedback control $u_* \in \Omega$ is called *optimal* if there exists an $\varepsilon > 0$ and a neighborhood N_* of the origin in $\mathbb{R}^n$ such that for each $b \in N_*$ the response $x_*(t)$ satisfies $|x_*(t)| \leqq \varepsilon$ and $|u_*(x_*(t), t)| \leqq \varepsilon$ for $t \in [\tau, T]$, and furthermore $J(\tau, b, u_*) \leqq J(\tau, b, u)$ among all feedback controls $u \in \Omega$ generating responses $x(t)$ with $|x(t)| \leqq \varepsilon$ and $|u(x(t), t)| \leqq \varepsilon$ for $t \in [\tau, T]$.

2.2. Statement of the main results.

THEOREM 2.1 (Main Theorem). *For the control process in* $\mathbb{R}^n$

$$\dot{x} = F(x, u, t), \qquad x(\tau) = b$$

with performance index

$$J(\tau, b, u) = L(x(T)) + \int_{\tau}^{T} G(x, u, t)\, dt$$

where x represents the solution (depending upon τ*) of the differential equation, there exists a unique optimal feedback control* $u_*(x, t)$*. This feedback control is the unique solution of the functional equation*

$$(*) \qquad F_u(x, u_*(x, t), t)J_x(t, x, u_*) + G_u(x, u_*(x, t), t) = 0$$

for small $|x|$ *and* $t \in [\tau, T]$*. Furthermore*

$$u_*(x, t) = D_*(t)x + h_*(x, t)$$

and

$$J(\tau, b, u_*) = b^T K_*(\tau)b + j_*(\tau, b),$$

where the matrix functions $D_*(t)$ *and* $K_*(t) \geqq 0$ *depend only on the truncated problem.*

THEOREM 2.2 (Truncated problem). *For the special case in which* $f(x, u, t) = 0$, $g(x, u, t) = 0$ *and* $l(x) = 0$ *the optimal control is given by*

$$u_*(x, t) = D_*(t)x$$

where

$$D_*(t) = -R^{-1}(t)B^T(t)K_*(t).$$

Here $K_*(t) \geqq 0$ *is a solution of the Riccati equation on* $[\tau, T]$*:*

$$\dot{K}(t) + Q(t) + K(t)A(t) + A^T(t)K(t) - K(t)B(t)R^{-1}(t)B^T(t)K(t) = 0,$$

$$K(T) = M.$$

Furthermore $D_*(t)x$ *is a global optimal control in the sense that we can take* $N_* = \mathbb{R}^n$ *and* $\varepsilon = \infty$ *in the definition of optimal feedback control. Finally*

$$J(\tau, b, u_*) = b^T K_*(\tau)b.$$

Remark. Note that for $u \in \Omega$ the property $J(T, b, u) = L(b)$ holds.

2.3. Construction of the optimal feedback control.

LEMMA 2.1. *For each feedback control* $u \in \Omega$, $u(x, t) = D(t)x + h(x, t)$*, there exists a neighborhood* N_u *of the origin in* $\mathbb{R}^n$ *in which*

$$J(\tau, b, u) = b^T \hat{K}(\tau)b + j(\tau, b).$$

Here $j(\tau, b)$ *contains the higher order terms in b. The matrix function* $\hat{K}(\tau) \geqq 0$ *depends only on the truncated problem. Furthermore, the functional equation*

$$F(x, u(x, t), t)^T J_x(t, x, u) + J_t(t, x, u) + G(x, u(x, t), t) = 0$$

holds for each $x \in N_u$, $t \in [\tau, T]$.

Proof. The following differential equation holds:

$$\dot{x} = (A(t)+B(t)D(t))x + B(t)h(x,t) + f(x, u(x,t), t),$$
$$x(\tau) = b.$$

If we define $A_*(t) := A(t)+B(t)D(t)$ and $v(x,t) := B(t)h(x,t)+f(x,u(x,t),t)$ then this equation becomes

$$\dot{x} = A_*(t)x + v(x,t),$$
$$x(\tau) = b.$$

From [6, § 1.7, Thms. 7.1., 7.2] it follows that the solution $x(t)$ of this differential equation exists for b in a neighborhood N_1 of the origin and furthermore that this solution may be differentiated to the initial value b. Hence it is clear that $x(t) = \Psi(t)b + O(|b|^2)$, uniformly for $t \in [\tau, T]$. From the variation of constants formula we conclude that $\Psi(t) = \Phi(t)\Phi^{-1}(\tau)$. So we can write

$$x(t) = \Phi(t)\Phi^{-1}(\tau)b + O(|b|^2),$$

uniformly for $t \in [\tau, T]$. Here $\Phi(t)$ is a *fundamental matrix* of the linear equation $\dot{x} = A_*(t)x$ (i.e. a nonsingular matrix function of dimension $n \times n$ which satisfies $\dot{\Phi}(t) = A_*(t)\Phi(t)$).Hence

$$\begin{aligned} G(x(t), u(x(t),t), t) &= x(t)^T Q(t)x(t) + x(t)^T D(t)^T R(t)D(t)x(t) + O(|x|^3) \\ &= b^T \Phi^{-T}(\tau)\Phi^T(t)\{Q(t)+D(t)^T R(t)D(t)\}\Phi(t)\Phi^{-1}(\tau)b \\ &\quad + O(|b|^3), \end{aligned}$$

uniformly for $t \in [\tau, T]$. Furthermore

$$\begin{aligned} L(x(T)) &= x(T)^T M x(T) + O(|x(T)|^3) \\ &= b^T \Phi^{-T}(\tau)\Phi^T(T)M\Phi(T)\Phi^{-1}(\tau)b + O(|b|^3). \end{aligned}$$

So

$$J(\tau, b, u) = b^T \hat{K}(\tau)b + O(|b|^3),$$

where

$$\begin{aligned} \hat{K}(\tau) &:= \Phi^{-T}(\tau)\Phi^T(T)M\Phi(T)\Phi^{-1}(\tau) \\ &\quad + \int_\tau^T [\Phi^{-T}(\tau)\Phi^T(t)\{Q(t)+D(t)^T R(t)D(t)\}\Phi(t)\Phi^{-1}(\tau)]\, dt. \end{aligned} \tag{2.2}$$

It is easy to verify that $\hat{K}(\tau) \geqq 0$ and $\hat{K}(T) = M$. It is known that there exists a neighborhood N_2 of the origin in $\mathbb{R}^n$ such that for each $s \in [\tau, T]$ and for each $b \in N_2$, the solution of $\dot{x} = F(x, u(x,t), t)$ with $x(s) = b$, exists on $[s, T]$. (See [6, § 1.7, Thm. 7.1]; note that our differential equation meets the requirements.)

Now let $N_u := N_1 \cap N_2$, $s \in [\tau, T]$ and $b \in N_u$. If $x(t, s, b)$ denotes the solution of $\dot{x} = F(x, u(x,t), t)$ with $x(s) = b$ then we can write

$$J(t, x(t,s,b), u) = L(x(T,s,b)) + \int_t^T G(x(\xi,s,b), u(x(\xi,s,b),\xi), \xi)\, d\xi$$

for $t \in [s, T]$ in agreement with the definition of J. One can verify that it is allowed to differentiate this equation with respect to t. Setting $t = s$ afterwards we get the equation

$$F(b, u(b, s), s)^T J_x(s, b, u) + J_t(s, b, u) + G(b, u(b, s), s) = 0.$$

If we finally replace b and s by x and t we get the desired result. □

Remark. From the proof it follows that we even have

$$J(t, x, u) = x^T \hat{K}(t) x + O(|x|^3)$$

uniformly for $t \in [\tau, T]$ and for small $|x|$.

LEMMA 2.2. *The equation*

$$F_u(x, u_*, t)p + G_u(x, u_*, t) = 0$$

has a solution $u_*(x, p, t)$ *near the origin in* $\mathbb{R}^{2n}$ *for which* $u_*(0, 0, t) = 0$ *for* $t \in [\tau, T]$. *Furthermore*

$$u_*(x, p, t) = -\tfrac{1}{2} R^{-1}(t) B^T(t) p + h_*(x, p, t),$$

where $h_*(x, p, t)$ *contains the higher order terms in* (x, p).

Proof. For each $t \in [\tau, T]$ we can use the result in [1, Lemma 2.2]. □

LEMMA 2.3. *There exists a unique solution* $K_*(t)$ *on* $[\tau, T]$ *to the matrix differential equations* (*Riccati equation*)

$$\dot{K}(t) + Q(t) + K(t)A(t) + A^T(t)K(t) - K(t)B(t)R^{-1}(t)B^T(t)K(t) = 0,$$

$$K(T) = M.$$

The property $K_*(t) \geqq 0$ *holds on* $[\tau, T]$.

Proof. See [3, § 2.3]. □

LEMMA 2.4. *Suppose there exists a feedback control* $u_*(x, t) = D_*(t)x + h_*(x, t)$, *which satisfies the nonlinear functional equation*

$$(*) \qquad F_u(x, u_*(x, t), t) J_x(t, x, u_*) + G_u(x, u_*(x, t), t) = 0$$

for small $|x|$ *and* $t \in [\tau, T]$. *Then* u_* *is the* unique optimal *feedback control. Furthermore*

$$D_*(t) = -R^{-1}(t)B^T(t)K_*(t)$$

and

$$J(\tau, b, u_*) = b^T K_*(\tau) b + j_*(\tau, b),$$

where $K_*(t)$ *is defined in Lemma* 2.3. *The function* $j_*(\tau, b)$ *contains the higher terms in* b.

Proof. Consider the following real valued function defined for $t \in [\tau, T]$ and for (x, u) near the origin in $\mathbb{R}^{n+r}$:

$$(2.3) \qquad Q(t, x, u) := F(x, u, t)^T J_x(t, x, u_*) + J_t(t, x, u_*) + G(x, u, t).$$

By Lemma 2.1.

$$Q(t, x, u_*(x, t)) = 0 \quad \text{near } x = 0 \text{ and for } t \in [\tau, T].$$

We have assumed that

$$Q_u(t, x, u_*(x, t)) = 0 \quad \text{near } x = 0 \text{ and for } t \in [\tau, T].$$

Furthermore the Hessian

$$Q_{uu}(t, 0, 0) = 2R(t) \quad \text{is positive definite for } t \in [\tau, T].$$

It follows that

$$Q_{uu}(t, x, u) > 0 \quad \text{for } |x| \text{ small, } |u| \text{ small and } t \in [\tau, T]$$

because $Q(t, x, u)$ is a continuous function. Hence we conclude that there exists an $\varepsilon > 0$ such that

$$0 = Q(t, x, u_*(x, t)) \leqq Q(t, x, u_1)$$

for $t \in [\tau, T]$, $|x| \leqq \varepsilon$ and $|u_1| \leqq \varepsilon$, while strict inequality holds for $u_1 \neq u_*(x, t)$. So

$$0 \leqq F(x, u_1, t)^T J_x(t, x, u_*) + J_t(t, x, u_*) + G(x, u_1, t). \tag{2.4}$$

Now let N_* be a neighborhood of the origin in $\mathbb{R}^n$ such that for each $b \in N_*$ the solution $x_*(t)$ of $\dot{x} = F(x, u_*(x, t), t)$, $x(\tau) = b$, exists for $t \in [\tau, T]$, $|x_*(t)| \leqq \varepsilon$ and $|u_*(x_*(t), t)| \leqq \varepsilon$.

Furthermore let $u_1 \in \Omega$ be an arbitrary feedback control such that the solution $x_1(t)$ of $\dot{x} = F(x, u_1(x, t), t)$, $x(\tau) = b$ is defined on $[\tau, T]$, and satisfies $|x_1(t)| \leqq \varepsilon$ and $|u_1(x_1(t), t)| \leqq \varepsilon$, if $b \in N_*$. Then we can write:

$$0 < \int_\tau^T \{F(x_1(t), u_1(x_1(t), t), t)^T J_x(t, x_1(t), u_*) + J_t(t, x_1(t), u_*) + G(x_1(t), u_1(x_1(t), t), t)\}\, dt,$$

and so

$$0 < \int_\tau^T \left\{\frac{d}{dt} J(t, x_1(t), u_*)\right\} dt + \int_\tau^T G(x_1(t), u_1(x_1(t), t), t)\, dt.$$

This yields the result

$$0 < J(T, x_1(T), u_*) - J(\tau, b, u_*) + \int_\tau^T G(x_1(t), u_1(x_1(t), t), t)\, dt$$

and thus

$$J(\tau, b, u_*) < J(\tau, b, u_1).$$

So $u_*(x, t)$ is the unique optimal feedback control. By Lemma 2.2 we have

$$u_*(x, t) = -\tfrac{1}{2} R^{-1}(t) B^T(t) J_x(t, x, u_*) + O(|x|^2),$$

uniformly for $t \in [\tau, T]$ and in Lemma 2.1 we have

$$J_x(t, x, u_*) = 2\hat{K}(t)x + O(|x|^2).$$

So

$$u_*(x, t) = -R^{-1}(t) B^T(t) \hat{K}(t) x + O(|x|^2), \tag{2.5}$$

uniformly for $t \in [\tau, T]$. By Lemma 2.1 we have

$$F(x, u_*(x, t), t)^T J_x(t, x, u_*) + J_t(t, x, u_*) + G(x, u_*(x, t), t) = 0 \tag{2.6}$$

for $|x|$ small and $t \in [\tau, T]$. Using (2.5) collecting the quadratic terms in x we find that $\hat{K}(t)$ is a solution of the Riccati equation. We also know that $\hat{K}(T) = M$ and by the uniqueness of the solution we have $\hat{K}(t) = K_*(t)$ on $[\tau, T]$. This yields the result

$$u_*(x, t) = -R^{-1}(t)B^T(t)K_*(t)x + O(|x|^2)$$

and

$$J(\tau, b, u_*) = b^T K_*(\tau)b + O(|b|^3). \quad \square$$

Proof of Theorem 2.2. Let $u_*(x, t) = D_*(t)x$, where $D_*(t) = -R^{-1}(t)B^T(t)K_*(t)$ and the matrix $K_*(t)$ satisfies the Riccati equation. Hence

$$x^T\{\dot{K}_*(t) + Q(t) + K_*(t)A(t) + A^T(t)K_*(t) - K_*(t)B(t)R^{-1}(t)B^T(t)K_*(t)\}x = 0$$

for all $x \in \mathbb{R}^n$. So we can write

$$\begin{aligned}[(A(t) - B(t)R^{-1}(t)B^T(t)K_*(t))x]^T 2K_*(t)x + x^T\dot{K}_*(t)x \\ + x^T Q(t)x + x^T K_*(t)B(t)R^{-1}(t)B^T(t)K_*(t)x = 0.\end{aligned}$$

It follows that

$$[(A(t) + B(t)D_*(t))x]^T 2K_*(t)x + x^T\dot{K}_*(t)x + x^T Q(t)x + [D_*(t)x]^T R(t)D_*(t)x = 0.$$

This yields

$$F(x, u_*(x, t), t)^T 2K_*(t)x + x^T\dot{K}_*(t)x + G(x, u_*(x, t), t) = 0.$$

By integrating this equation along the trajectory $\dot{x} = F(x, u_*(x, t), t)$, $x(\tau) = b$, where b is arbitrary in $\mathbb{R}^n$, we obtain the equation

$$J(\tau, b, u_*) = b^T K_*(\tau)b \qquad (b \in \mathbb{R}^n).$$

It is now easy to verify that $u_*(x, t)$ satisfies the functional equation (*) in Lemma 2.4. The global character of $u_*(x, t)$ follows from examination of the proof of Lemma 2.4. $\square$

Before giving the proof of the main theorem, we consider the *Hamiltonian system* in $\mathbb{R}^{2n}$:

$$\begin{aligned}\dot{x} &= F(x, u_*(x, p, t), t), \\ \dot{p} &= -\{F_x(x, u_*(x, p, t), t)p + G_x(x, u_*(x, p, t), t)\}\end{aligned} \tag{2.7}$$

with the boundary values

$$\begin{aligned}x(\tau) &= b, \\ p(T) &= L_x(x(T)).\end{aligned}$$

Here $u_*(x, p, t)$ is defined in Lemma 1.2.

LEMMA 2.5. *For small* $|b|$ *system* (2.7) *has a solution* $(x_*(t), p_*(t))$ *on* $[\tau, T]$ *with the property*

$$p_*(t) = 2K_*(t)x_*(t) + O(|x_*(t)|^2),$$

uniformly for $t \in [\tau, T]$.

Proof. The Hamiltonian system has the form

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} A(t) & -\frac{1}{2}B(t)R^{-1}(t)B^T(t) \\ -2Q(t) & -A^T(t) \end{pmatrix} \begin{pmatrix} x \\ p \end{pmatrix} + h(x, p, t),$$

where the function $h(x, p, t)$ contains the higher order terms. First of all we shall prove that the lemma holds for the case that $h(x, p, t) = 0$. The solvability of the linear system together with the implicit function theorem will be used to obtain a proof for the general case. So we shall first consider the linear Hamiltonian system

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} A(t) & -\frac{1}{2}B(t)R^{-1}(t)B^T(t) \\ -2Q(t) & -A^T(t) \end{pmatrix} \begin{pmatrix} x \\ p \end{pmatrix},$$

with $x(\tau) = b$ and $p(T) = 2Mx(T)$. This system has a solution $(x_*(t), p_*(t))$ with the property $p_*(t) = 2K_*(t)x_*(t)$, which can easily be verified. Note that this solution exists for each $b \in \mathbb{R}^n$. If we now consider this linear system as a final value problem:

$$x(T) = x_T, \qquad p(T) = p_T,$$

then the solution is given by

$$\begin{pmatrix} x \\ p \end{pmatrix}(t) = \Phi(t)\Phi^{-1}(T) \begin{pmatrix} x_T \\ p_T \end{pmatrix}. \tag{2.8}$$

Here $\Phi(t)$ is a fundamental matrix of the problem. If we partition

$$\Phi(t)\Phi^{-1}(T) = \begin{pmatrix} \Theta_{11}(t, T) & \Theta_{12}(t, T) \\ \Theta_{21}(t, T) & \Theta_{22}(t, T) \end{pmatrix},$$

then (2.8) can be written as

$$x(t, x_T, p_T) = \Theta_{11}(t, T)x_T + \Theta_{12}(t, T)p_T,$$
$$p(t, x_t, p_T) = \Theta_{21}(t, T)x_T + \Theta_{22}(t, T)p_T.$$

So

$$x(t, x_T, 2Mx_T) = (\Theta_{11}(t, T) + 2\Theta_{12}(t, T)M)x_T.$$

We saw that for each $b \in \mathbb{R}^n$ there exists a solution on $[\tau, T]$ with $p(T) = 2Mx(T)$. So

$$\forall b \in \mathbb{R}^n \, \exists x_T \in \mathbb{R}^n : (\Theta_{11}(\tau, T) + 2\Theta_{12}(\tau, T)M)x_T = b.$$

Hence the matrix

$$\Theta_{11}(\tau, T) + 2\Theta_{12}(\tau, T)M \tag{2.9}$$

is regular. We shall need this result later. Now consider the nonlinear Hamiltonian

system as a final value problem: $x(T)=x_T$, $p(T)=p_T$. The solution has the form

$$\begin{pmatrix} x \\ p \end{pmatrix}(t)=\Phi(t)\Phi^{-1}(T)\begin{pmatrix} x_T \\ p_T \end{pmatrix}+v(t, x_T, p_T),$$

where $v(t, x_T, p_T)$ contains the second and higher order terms in x_T and p_T. It follows that

$$x(t, x_T, p_T)=\Theta_{11}(t, T)x_T+\Theta_{12}(t, T)p_T+O\left(\left|\begin{pmatrix} x_T \\ p_T \end{pmatrix}\right|^2\right),$$

$$p(t, x_T, p_T)=\Theta_{21}(t, T)x_T+\Theta_{22}(t, T)p_T+O\left(\left|\begin{pmatrix} x_T \\ p_T \end{pmatrix}\right|^2\right),$$

uniformly for $t\in[\tau, T]$. The question is: does there exist for arbitrary $b\in\mathbb{R}^n$, $|b|$ small, a vector $x_T\in\mathbb{R}^n$ such that $x(\tau, x_T, L_x(x_T))=b$? Here the implicit function theorem can help us. Define

$$F(b, x_T) := x(\tau, x_T, L_x(x_T))-b.$$

Then $F(0, 0)=0$ and $F_{x_T}(0, 0)=\Theta_{11}(\tau, T)+2\Theta_{12}(\tau, T)M$. By (2.9) we have that $F_{x_T}(0, 0)$ is regular. Thus there exists a neighborhood $\mathcal{S}$ of the origin in $\mathbb{R}^n$ and a function $\tilde{x}_T:\mathcal{S}\to\mathbb{R}^n$ such that

(i) $$\tilde{x}_T(0)=0,$$

(ii) $$F(b, \tilde{x}_T(b))=0 \quad \text{for } b\in\mathcal{S}.$$

So $x(\tau, \tilde{x}_T(b), L_x(\tilde{x}_T(b)))=b$. Hence the Hamiltonian system (2.7) has a solution on $[\tau, T]$ for small $|b|$. From the considerations of the linear system we have

$$p_*(t)=2K_*(t)x_*(t)+O(|x_*(t)|^2),$$

uniformly for $t\in[\tau, T]$. □

Proof of the Main Theorem. It is sufficient to establish the existence of a feedback control $u_*\in\Omega$ which satisfies the functional equation (*). Define

(2.10) $$u_*(x, t) := u_*(x, p_*(x, t), t),$$

where $p_*(x, t)$ represents the solution of (2.7) and $u_*(x, p, t)$ is defined as in Lemma 2.2. Then

$$\begin{aligned} u_*(x, t) &= -\tfrac{1}{2}R^{-1}(t)B^T(t)p_*(x, t)+O(|x|^2) \\ &= -R^{-1}(t)B^T(t)K_*(t)x+O(|x|^2) \end{aligned}$$

uniformly for $t\in[\tau, T]$. Thus we can conclude that $u_*\in\Omega$. Now let $s\in[\tau, T]$ fixed and choose $y\in\mathbb{R}^n$ so small that the solution of $\dot{x}=F(x, u_*(x, t), t)$, with $x(s)=y$, exists on $[\tau, T]$, and $x(\tau)=:b$ is so small that the solution of (2.7) exists. By the continuity and analyticity of $G(x, u_*(x, t), t)$ the following differentiation of the

integral is allowed:

$$\frac{\partial J(s, y, u_*)}{\partial y} = \int_s^T \frac{\partial}{\partial y} G(x, u_*(x, t), t)\, dt + \frac{\partial}{\partial y} L(x(T))$$

$$= \int_s^T \left\{ \frac{\partial x}{\partial y} \frac{\partial G(x, u_*(x, t), t)}{\partial x} + \frac{\partial u_*}{\partial y} \frac{\partial G(x, u_*(x, t), t)}{\partial u_*} \right\} dt + \frac{\partial}{\partial y} L(x(T))$$

$$= \int_s^T \left\{ \frac{\partial x}{\partial y} \left[-\dot{p}_*(x, t) - \frac{\partial F(x, u_*(x, t), t)}{\partial x} p_*(x, t) \right] \right.$$

$$\left. + \frac{\partial u_*}{\partial y} \frac{\partial G(x, u_*(x, t), t)}{\partial u_*} \right\} dt + \frac{\partial}{\partial y} L(x(T))$$

$$= -\int_s^T \left\{ \frac{\partial x}{\partial y} \dot{p}_*(x, t) \right\} dt + \frac{\partial}{\partial y} L(x(T))$$

$$+ \int_s^T \left\{ \frac{\partial u_*}{\partial y} \left[-\frac{\partial F(x, u_*(x, t), t)}{\partial u_*} p_*(x, t) \right] \right.$$

$$\left. - \frac{\partial x}{\partial y} \frac{\partial F(x, u_*(x, t), t)}{\partial x} p_*(x, t) \right\} dt$$

$$= -\frac{\partial x}{\partial y} p_*(x, t) \Big|_s^T + \int_s^T \left\{ \frac{d}{dt} \frac{\partial x}{\partial y} p_*(x, t) \right\} dt + \frac{\partial}{\partial y} L(x(T))$$

$$- \int_s^T \left[\frac{\partial}{\partial y} F(x, u_*(x, t), t) \right] p_*(x, t)\, dt$$

$$= p_*(y, s) - \frac{\partial x(T)}{\partial y} L_x(x(T)) + \frac{\partial}{\partial y} L(x(T)) = p_*(y, s).$$

So $J_y(s, y, u_*) = p_*(y, s)$ for small $|y|$ and $s \in [\tau, T]$. If we now replace s by t and y by x, and if we use Lemma 2.2, we obtain

$$F_u(x, u_*(x, t), t) J_x(t, x, u_*) + G_u(x, u_*(x, t), t) = 0$$

for $|x|$ small and $t \in [\tau, T]$. So $u_*(x, t)$ satisfies (*). □

2.4. A method for calculating $u_*(x, t)$ and $J(t, x, u_*)$. In this section we shall use the following notation: if $t(x)$ is a power series in x then the kth order term will be denoted by $t^{(k)}(x)$ or $[t(x)]^{(k)}$.

$u_*(x, t)$ and $J_*(x, t) := J(t, x, u_*)$ can be expanded in power series:

$$u_*(x, t) = u_*^{(1)}(x, t) + u_*^{(2)}(x, t) + \cdots,$$

$$J_*(x, t) = J_*^{(2)}(x, t) + J_*^{(3)}(x, t) + \cdots.$$

We have seen that the lowest order terms are given by

$$u_*^{(1)}(x, t) = D_*(t)x \quad \text{and} \quad J_*^{(2)}(x, t) = x^T K_*(t) x,$$

where

$$D_*(t) = -R^{-1}(t) B^T(t) K_*(t)$$

and $K_*(t)$ is the solution of the Riccati equation. We indicate a method for computing the higher order terms analogous to the method followed in [1]. This method is based on the fact that $u_*(x, t)$ is a solution of the following two functional equations:

$$F(x, u_*(x, t), t)^T J_x(t, x, u_*) + J_t(t, x, u_*) + G(x, u_*(x, t), t) = 0,$$

$$F_u(x, u_*(x, t), t) J_x(t, x, u_*) + G_u(x, u_*(x, t), t) = 0.$$

In contrast to [1] where one has to solve linear algebraic equations, the problem defined here reduces to solving successively a set of linear differential equations. We shall now give the result in the form of two equations:

$$\begin{aligned}(A_*(t)x)^T[J_*^{(m)}(x, t)]_x + [J_*^{(m)}(x, t)]_t \\ = -\sum_{k=3}^{m-1} [B(t)u_*^{(m-k+1)}(x, t)]^T [J_*^{(k)}(x, t)]_x \\ -\sum_{k=2}^{m-1} f^{(m-k+1)}(x, u_*(x, t), t)^T [J_*^{(k)}(x, t)]_x \\ -2\sum_{k=2}^{[(m-1)/2]} u_*^{(k)}(x, t)^T R(t) u_*^{(m-k)}(x, t) \\ -u_*^{(m/2)}(x, t)^T R(t) u_*^{(m/2)}(x, t) - g^{(m)}(x, u_*(x, t), t)\end{aligned} \tag{A}$$

$$(m = 3, 4, \cdots);$$

$$\begin{aligned}u_*^{(k)}(x, t) = -\tfrac{1}{2}R^{-1}(t)\Big\{ B^T(t)[J_*^{(k+1)}(x, t)]_x \\ + \sum_{j=1}^{k-1} [f_u(x, u_*(x, t), t)]^{(j)} [J_*^{(k-j+1)}(x, t)]_x \\ + [g_u(x, u_*(x, t), t)]^{(k)}\Big\}\end{aligned} \tag{B}$$

$$(k = 2, 3, \cdots).$$

Here $A_*(t) := A(t) + B(t)D_*(t)$; $[k]$ denotes the integer part of k. Furthermore the term with $u_*^{(m/2)}$ is to be omitted for odd values of m.

With the values $J_*^{(2)}(x, t)$ and $u_*^{(1)}(x, t)$ to start with, the higher order terms can be calculated from (A) and (B) in the sequence

$$J_*^{(3)}(x, t), u_*^{(2)}(x, t), J_*^{(4)}(x, t), u_*^{(3)}(x, t), \cdots.$$

The sequence of terms $\{u_*^{(1)}, \cdots, u_*^{(m-2)}; J_*^{(2)}, \cdots, J_*^{(m-1)}\}$ determines $J_*^{(m)}$ in (A) by solving a partial differential equation with boundary value $J_*^{(m)}(x, T) = L^{(m)}(x)$. The sequence of terms $\{u_*^{(1)}, \cdots, u_*^{(k-1)}; J_*^{(2)}, \cdots . J_*^{(k+1)}\}$ determines $u_*^{(k)}$ in (B).

Example.

$$\dot{x} = x^3 + u, \qquad x(0) = x_0,$$

$$\min \int_0^T (x^2 + u^2)\, dt.$$

Here $A(t) = 0$, $B(t) = 1$, $Q(t) = 1$ and $R(t) = 1$. Furthermore $f(x, u, t) = x^3$,

$g(x, u, t) = 0$ and $L(x) = 0$. We have the Riccati equation

$$\dot{K} + 1 - K^2 = 0,$$

$$K(T) = 0,$$

and the solution is given by $K_*(t) = \tanh(T-t)$. Hence

$$J_*^{(2)}(x, t) = x^T K_*(t) x = x^2 \tanh(T-t)$$

and

$$u_*^{(1)}(x, t) = -R^{-1}(t) B^T(t) K_*(t) x = -x \tanh(T-t).$$

Furthermore

$$A_*(t) = A(t) - B(t) R^{-1}(t) B^T(t) K_*(t) = -\tanh(T-t).$$

For $m = 3$, equation (A) reads as follows:

$$(-x \tanh(T-t))[J_*^{(3)}(x, t)]_x + [J_*^{(3)}(x, t)]_t = 0.$$

If we set

$$J_*^{(3)}(x, t) = \alpha(t) x^3$$

then this equation becomes

$$-3x^3 \alpha(t) \tanh(T-t) + \dot{\alpha}(t) x^3 = 0$$

or

$$\dot{\alpha}(t) - 3\alpha(t) \tanh(T-t) = 0$$

with the boundary value $\alpha(T) = 0$. This yields the solution $\alpha(t) = 0$ on $[\tau, T]$. So $J_*^{(3)}(x, t) = 0$ and (B) give for $k = 2$: $u_*^{(2)}(x, t) = 0$.

For $m = 4$, equation (A) becomes

$$(-x \tanh(T-t))[J_*^{(4)}(x, t)]_x + [J_*^{(4)}(x, t)]_t = -f^{(3)}(x, u_*, t)[J_*^{(2)}(x, t)]_x.$$

Setting $J_*^{(4)}(x, t) = \alpha(t) x^4$ we have

$$\{-4\alpha(t) \tanh(T-t) + \dot{\alpha}(t)\} x^4 = -2 \tanh(T-t) x^4$$

or

$$\dot{\alpha}(t) - 4\alpha(t) \tanh(T-t) + 2 \tanh(T-t) = 0$$

with the boundary value $\alpha(T) = 0$. The solution of this differential equation is

$$\alpha(t) = \tfrac{1}{2} - \tfrac{1}{2}(\cosh(T-t))^{-4}.$$

Thus

$$J_*^{(4)}(x, t) = \{\tfrac{1}{2} - \tfrac{1}{2}(\cosh(T-t))^{-4}\} x^4.$$

Formula (B) gives for $k = 3$:

$$u_*^{(3)}(x, t) = -\tfrac{1}{2} R^{-1}(t) B^T(t) [J_*^{(4)}(x, t)]_x,$$

so

$$u_*^{(3)}(x, t) = \{-1 + (\cosh(T-t))^{-4}\}x^3.$$

The higher order terms can be computed in a similar manner.

3. Fixed end-point problem.

3.1. Assumptions. In this section we consider a problem similar to the problem discussed in § 2. The difference being that now we require the final value of the state to be zero: $x(T) = 0$. As a matter of course we can take now $L(x) = 0$. The basic assumptions made in § 2 remain. A new assumption is the *controllability to the zero state* of the linear system $\dot{x} = A(t)x + B(t)u$. Furthermore we restrict ourselves to feedback controls $u(x, t)$ with the following properties:

1. $u(x, t) = D(t)x + h(x, t)$. Here $D(t)$ is a continuous matrix function for $t \in [\tau, T)$. The function $h(x, t)$ contains the higher order terms in x and is continuous with respect to $t \in [\tau, T)$. Furthermore $h(x, t)$ is given as a power series in x which starts with second order terms and converges about the origin.

2. There exists a neighborhood N_u of the origin in $\mathbb{R}^n$ such that for $b \in N_u$ the solution $x(t, \tau, b)$ of (1.1) is defined on $[\tau, T)$ and in addition $\lim_{t \to T} x(t, \tau, b) = 0$.

3. The function $t \to u(x(t, \tau, b), t)$ is bounded on $[\tau, T]$ for all $b \in N_u$.

We shall denote again the class of admissible feedback controls by Ω. If $u \in \Omega$ then it is clear that $u(x, t)$ has a singularity in $t = T$. Furthermore there exists for given $u \in \Omega$, $s \in [\tau, T)$, a neighborhood $N_{u,s}$ of the origin in $\mathbb{R}^n$ with the property that, if $c \in N_{u,s}$, the solution of $\dot{x} = F(x, u(x, t), t)$, $x(s) = c$, is defined on $[s, T]$ and $x(T) = 0$. It is evident that

$$N_{u,s} := \{x(s, \tau, b) | b \in N_u\} \tag{3.1}$$

represents such a neighborhood!

3.2. Statement of the main results.

THEOREM 3.1 (Main Theorem). *For the control process in* $\mathbb{R}^n$

$$\dot{x} = F(x, u, t), \quad x(\tau) = b, \quad x(T) = 0$$

there exists a unique optimal feedback control $u_* \in \Omega$ *which minimizes the integral*

$$J(\tau, b, u) = \int_\tau^T G(x, u, t)\, dt$$

for all initial states b in a neighborhood of the origin in $\mathbb{R}^n$. *This feedback control is the unique solution of the functional equation*

$$F_u(x, u_*(x, t), t) J_x(t, x, u_*) + G_u(x, u_*(x, t), t) = 0 \tag{*}$$

for $t \in [\tau, T)$ *and small* $|x|$. *Furthermore*

$$u_*(x, t) = D_*(t)x + h_*(x, t)$$

and

$$J(\tau, b, u_*) = b^T K_*(\tau) b + j_*(\tau, b),$$

where the matrix functions $D_*(t)$ *and* $K_*(t)$ *are defined on* $[\tau, T)$ *and depend only*

on the truncated problem.

The *truncated problem* is the case that $f(x, u, t) = 0$ and $g(x, u, t) = 0$. R. W. Brockett has proved in [2] that under our hypothesis an optimal control exists. One can easily show that his results can be written in the following form:

$$u_*(x, t) = D_*(t)x \tag{3.2}$$

where

$$D_*(t) = -R^{-1}(t)B^T(t)K_*(t). \tag{3.3}$$

Here $K_*(t)$ satisfies the Riccati equation on $[\tau, T)$:

$$\dot{K}(t) + Q(t) + K(t)A(t) + A^T(t)K(t) - K(t)B(t)R^{-1}(t)B^T(t)K(t) = 0.$$

If $W_*(t)$ satisfies the *dual Riccati equation*

$$\dot{W}(t) + B(t)R^{-1}(t)B^T(t) - W(t)A^T(t) - A(t)W(t) - W(t)Q(t)W(t) = 0,$$

$$W(T) = 0$$

on $[\tau, T]$, then we have $K_*^{-1}(t) = W_*(t)$ for $t \in [\tau, T)$. Finally

$$J(\tau, b, u_*) = b^T K_*(\tau) b.$$

3.3. Construction of the optimal feedback control.

LEMMA 3.1. *For each feedback control $u \in \Omega$, $u(x, t) = D(t)x + h(x, t)$, we have the property*

$$J(\tau, b, u) = b^T \hat{K}(\tau) b + j(\tau, b)$$

for $b \in N_u$. The matrix function $\hat{K}(\tau)$ depends only on the truncated problem. Furthermore the functional equation

$$F(x, u(x, t), t)^T J_x(t, x, u) + J_t(t, x, u) + G(x, u(x, t), t) = 0$$

holds for $t \in [\tau, T)$ and $x \in N_{u,t}$.

Proof. Just like the proof of Lemma 2.1 we shall show that the solution of the differential equation $\dot{x} = F(x, u(x, t), t)$ is of the form $x(t) = \Phi(t)\Phi^{-1}(\tau)b + O(|b|^2)$, uniformly for $t \in [\tau, T]$. We can write again $\dot{x} = A_*(t)x + v(x, t)$, $x(\tau) = b$ and it is known that the solution exists on $[\tau, T]$ for $b \in N_u$ and furthermore $x(T) = 0$. The function $v(x, t)$ contains the higher order terms in x. Hence $|v(x(t), t)| \leqq \Theta(x)|x(t)|$ for $t \in [\tau, T]$, where the function Θ has the property $\lim_{x \to 0} \Theta(x) = 0$. With the variation of constants formula we find

$$x(t) = \Phi(t)\Phi^{-1}(\tau)b + \int_\tau^t \Phi(t)\Phi^{-1}(\xi)v(x(\xi), \xi)\, d\xi.$$

Note that $\Phi(T) = 0$! The continuity of Φ on $[\tau, T]$ has the result that $\Phi(t)\Phi^{-1}(s)$ is bounded for all $t \in [\tau, T]$ and $s \in [\tau, t]$; say $|\Phi(t)\Phi^{-1}(s)| \leqq M$ where M is a positive number (this is not trivial; note that there are no troubles when $t \to T$ and $s \to T$). Hence

$$|x(t)| \leqq M|b| + \int_\tau^t M\Theta(x(\xi))|x(\xi)|\, d\xi.$$

Choose $\delta > 0$ such that $|\Theta(x)| \leqq 1$ for $|b| < \delta$. Then

$$|x(t)| \leqq M|b| + \int_\tau^t M|x(\xi)|\, d\xi.$$

From the Gronwall inequality (see [7, § 1.7]) we get

$$(**) \qquad |x(t)| \leqq M|b| \exp(M(t-\tau))$$

for $|b| < \delta$. Furthermore there exists a positive number K such that $|v(x(t), t)| \leqq K|x(t)|^2$ for $t \in [\tau, T]$. So

$$|x(t) - \Phi(t)\Phi^{-1}(\tau)b| \leqq M \int_\tau^t |v(x(\xi), \xi)|\, d\xi \leqq MK \int_\tau^t |x(\xi)|^2\, d\xi.$$

Substitution of (**) into the latter equation gives the result

$$x(t) - \Phi(t)\Phi^{-1}(\tau)b = O(|b|^2),$$

uniformly for $t \in [\tau, T]$.

The remainder of the proof is analogous to the proof of Lemma 2.1. Note that $\hat{K}(t)$ may have a singularity in $t = T$. □

LEMMA 3.2. *There exists a unique solution $W_*(t)$ on $[\tau, T]$ to the matrix differential equation (dual Riccati equation)*

$$\dot{W}(t) + B(t)R^{-1}(t)B^T(t) - W(t)A^T(t) - A(t)W(t) - W(t)Q(t)W(t) = 0,$$

$$W(T) = 0.$$

The property $W_*(t) > 0$ holds on $[\tau, T)$. If $K_*(t) := W_*^{-1}(t)$ on $[\tau, T)$ then $K_*(t)$ satisfies the Riccati equation

$$\dot{K}(t) + Q(t) + K(t)A(t) + A^T(t)K(t) - K(t)B(t)R^{-1}(t)B^T(t)K(t) = 0$$

on $[\tau, T)$.

Proof. This lemma is a consequence of the analysis of R. W. Brockett in [2, § 3.22]. □

LEMMA 3.3. *Suppose there exists a feedback control $u_* \in \Omega$, $u_*(x, t) = D_*(t)x + h_*(x, t)$, which satisfies the functional equation*

$$(*) \qquad F_u(x, u_*(x, t), t)J_x(t, x, u_*) + G_u(x, u_*(x, t), t) = 0$$

for $t \in [\tau, T)$ and $x \in N_{u_*, t}$. Then u_* is the unique optimal *feedback control. Furthermore*

$$D_*(t) = -R^{-1}(t)B^T(t)K_*(t)$$

and

$$J(\tau, b, u_*) = b^T K_*(\tau)b + j_*(\tau, b),$$

where $K_*(t)$ is defined in Lemma 3.2. The function $j_*(\tau, b)$ contains the higher order terms in b.

Proof. The method to proof that u_* represents the unique optimal feedback control is analogous to the method followed in Lemma 2.4. Now we can choose $|u_*(x_*(t), t)| \leqq \varepsilon$ and $|u_1(x_1(t), t)| \leqq \varepsilon$ because we have assumed that $u_*(x_*(t), t)$

and $u_1(x_1(t), t)$ are bounded functions on $[\tau, T]$. By Lemma 2.2 we have

$$u_*(x, t) = -\tfrac{1}{2}R^{-1}(t)B^T(t)J_x(t, x, u_*) + O(|x|^2)$$

and in Lemma 3.1 we have

$$J_x(t, x, u_*) = 2\hat{K}(t)x + O(|x|^2)$$

for $t \in [\tau, T]$ and $x \in N_{u_*,t}$. So

$$u_*(x, t) = -R^{-1}(t)B^T(t)\hat{K}(t)x + O(|x|^2). \tag{3.4}$$

In the truncated case the corresponding formula is:

$$u_*(x, t) = -R^{-1}(t)B^T(t)\hat{K}(t)x.$$

Comparing this result with (3.2) and (3.3) it follows that $\hat{K}(t) = K_*(t)$ on $[\tau, T)$, where $K_*(t)$ is defined in Lemma 3.2.

$$u_*(x, t) = -R^{-1}(t)B^T(t)K_*(t)x + O(|x|^2)$$

and

$$J(\tau, b, u_*) = b^T K_*(\tau)b + O(|b|^3). \quad \square$$

Before proving the main theorem we consider again the Hamiltonian system in $\mathbb{R}^{2n}$:

$$\begin{aligned} \dot{x} &= F(x, u_*(x, p, t), t), \\ \dot{p} &= -\{F_x(x, u_*(x, p, t), t)p + G_x(x, u_*(x, p, t), t)\} \end{aligned} \tag{3.5}$$

with the boundary values

$$x(\tau) = b, \qquad x(T) = 0.$$

Here $u_*(x, p, t)$ is defined in Lemma 2.2.

LEMMA 3.4. *For small* $|b|$ *system* (3.5) *has a solution* $(x_*(t), p_*(t))$ *with the property*

$$x_*(t) = \tfrac{1}{2}W_*(t)p_*(t) + O(|p_*(t)|^2)$$

for $t \in [\tau, T]$. *Furthermore* $p_*(t)$ *is a bounded function on* $[\tau, T]$.

Proof. The Hamiltonian system has the form

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} A(t) & -\tfrac{1}{2}B(t)R^{-1}(t)B^T(t) \\ -2Q(t) & -A^T(t) \end{pmatrix} \begin{pmatrix} x \\ p \end{pmatrix} + h(x, p, t).$$

It can easily be verified that the linear system (i.e. the case that $h(x, p, t) = 0$) has for each $b \in \mathbb{R}^n$ a solution of the form $x_*(t) = \frac{1}{2}W_*(t)p_*(t)$. Analogous to the proof of Lemma 2.5 we shall use the implicit function theorem to prove that the nonlinear system has a solution of the desired form. We need again a property which we shall derive from the solvability of the linear system. So consider again the linear Hamiltonian system as a final value problem. The solution can be

written as

$$x(t, x_T, p_T) = \Theta_{11}(t, T)x_T + \Theta_{12}(t, T)p_T,$$
$$p(t, x_T, p_T) = \Theta_{21}(t, T)x_T + \Theta_{22}(t, T)p_T.$$

We have seen that for each $b \in \mathbb{R}^n$ there exists a solution on $[\tau, T]$ with $x(\tau) = b$ and $x(T) = 0$. So

$$\forall b \in \mathbb{R}^n \exists p_T \in \mathbb{R}^n : \Theta_{12}(\tau, T)p_T = b.$$

Hence the matrix $\Theta_{12}(\tau, T)$ is regular. Now consider the nonlinear system as a final value problem. The solution has the form

$$x(t, x_t, p_T) = \Theta_{11}(t, T)x_T + \Theta_{12}(t, T)p_T + O\left(\left|\begin{pmatrix} x_T \\ p_T \end{pmatrix}\right|^2\right),$$

$$p(t, x_T, p_T) = \Theta_{21}(t, T)x_T + \Theta_{22}(t, T)p_T + O\left(\left|\begin{pmatrix} x_T \\ p_T \end{pmatrix}\right|^2\right).$$

The question is: does there exist for arbitrary $b \in \mathbb{R}^n$, $|b|$ small, a vector $p_T \in \mathbb{R}^n$ such that $x(\tau, 0, p_T) = b$? Again, the implicit function theorem can help us. Define

$$F(b, p_T) := x(\tau, 0, p_T) - b.$$

Then $F(0, 0) = 0$ and $F_{p_T}(0, 0) = \Theta_{12}(\tau, T)$. So $F_{p_T}(0, 0)$ is regular, and there exists a neighborhood $\mathcal{S}$ of the origin in $\mathbb{R}^n$ and a function $p_T : \mathcal{S} \to \mathbb{R}^n$ such that

(i) $$\tilde{p}_T(0) = 0,$$

(ii) $$F(b, \tilde{p}_T(b)) = 0 \quad \text{for } b \in \mathcal{S}.$$

Hence $x(\tau, 0, \tilde{p}_T(b)) = b$ for $b \in \mathcal{S}$. Thus the Hamiltonian system (3.5) has a solution on $[\tau, T]$ for small $|b|$. From the considerations of the linear system we have

$$x_*(t) = \tfrac{1}{2}W_*(t)p_*(t) + O(|p_*(t)|^2)$$

for $t \in [\tau, T]$. The boundedness of $p_*(t)$ on $[\tau, T]$ is a consequence of the continuity of the right-hand side of (3.5) on $[\tau, T]$. □

Remark. It follows that

$$p_*(t) = 2K_*(t)x_*(t) + O(|x_*(t)|^2)$$

for $t \in [\tau, T)$.

Proof of the Main Theorem. It is sufficient to establish the existence of a feedback control $u_* \in \Omega$ which satisfied the functional equation (*) for $t \in [\tau, T)$ and small $|x|$. Define

$$u_*(x, t) := u_*(x, p_*(x, t), t)$$

where $p_*(x, t)$ represents the solution of (3.5) and $u_*(x, p, t)$ such as defined in Lemma 2.2. Hence

$$u_*(x, t) = -\tfrac{1}{2}R^{-1}(t)B^T(t)p_*(x, t) + O(|x|^2)$$
$$= -R^{-1}(t)B^T(t)K_*(t)x + O(x^2)$$

for $t \in [\tau, T]$. In Lemma 3.4 we have seen that the solution of $\dot{x} = F(x, u_*(x, t), t)$, $x(\tau) = b$ exists on $[\tau, T]$ for small $|b|$ and furthermore $x(T) = 0$. Because $p_*(t)$ is bounded on $[\tau, T]$ it follows that $u_*(x_*(t), t)$ is bounded on $[\tau, T]$. Hence we can conclude that $u_* \in \Omega$. An analogous argument as in the previous section shows us that u_* satisfies the functional equation (*). □

3.4. A method for calculating $u_*(x, t)$. In § 2 we used the fact that the optimal feedback control $u_*(x, t)$ is a solution of the following two equations:

$$F(x, u_*(x, t), t)^T J_x(t, x, u_*) + J_t(t, x, u_*) + G(x, u_*(x, t), t) = 0,$$
$$F_u(x, u_*(x, t), t) J_x(t, x, u_*) + G_u(x, u_*(x, t), t) = 0.$$

It turned out to be possible to calculate $u_*(x, t)$ from these equations using the boundary value $J(T, x, u_*) = L(x)$ to solve the partial differential equation. This method fails here. It is true that the optimal feedback control is again a solution of the two functional equations but we cannot solve the partial differential equation because the only information we have about J is that $J(T, 0, u_*) = 0$ and this is not sufficient. This is a reason for us to follow a different method here. Consider the following free end-point problem

$$\dot{p} = \tilde{F}(p, y, t), \qquad p(\tau) = c,$$
$$\min \int_\tau^T \tilde{G}(p, y, t)\, dt.$$

Note that p plays the role of state vector and y plays the role of control vector. The functions $\tilde{F}$ and $\tilde{G}$ are defined as follows:

$$\tilde{F}(p, y, t) := -\{F_x(y, u_*(y, p, t), t)p + G_x(y, u_*(y, p, t), t)\}$$
$$\tilde{G}(p, y, t) := -[F_x(y, u_*(y, p, t), t)p + G_x(y, u_*(y, p, t), t)]^T x$$
$$-\{F(y, u_*(y, p, t), t)^T p + G(y, u_*(y, p, t), t)\}.$$

Here $u_*(x, p, t)$ is defined in Lemma 2.2. We shall call this control system the *dual system*. It is easy to verify that

$$\tilde{F}(p, y, t) = -A^T(t)p - 2Q(t)y + \tilde{f}(p, y, t)$$

and

$$\tilde{G}(p, y, t) = \tfrac{1}{4} p^T B(t) R^{-1}(t) B^T(t) p + y^T Q(t) y + \tilde{g}(p, y, t).$$

Here the functions $\tilde{f}$ and $\tilde{g}$ contain the higher order terms in y and p. It is clear that the dual system can be solved by the method described in § 2, provided that $Q(t) > 0$ on $[\tau, T]$. However, what is the connection with the original system? The two systems have one important common property; namely they both generate the same Hamiltonian system:

$$\dot{x} = F(x, u_*(x, p, t), t),$$
$$\dot{p} = -\{F_x(x, u_*(x, p, t), t)p + G_x(x, u_*(x, p, t), t)\}.$$

The boundary values however are different. In the original case we have $x(\tau)=b$, $x(T)=0$ and in the dual case $p(\tau)=c$, $x(T)=0$. Namely, if $y_*(p,x,t)$ here plays the role of $u_*(x,p,t)$ in Lemma 2.2 then it is easy to verify that $y_*(p,x,t)=x$ and furthermore $-\{\tilde{F}_p(p,y_*(p,x,t),t)x+\tilde{G}_p(p,y_*(p,x,t),t)\}=F(x,u_*(x,p,t),t)$. This argument enables us to construct the solution of the original system from the solution of the dual system. If $y_*(p,t)$ denotes the optimal feedback control with respect to the dual problem then it follows that $x_*(p,t)=y_*(p,t)$ is the solution of the Hamiltonian system. From this we can calculate $p_*(x,t)$ by the regular transformation $p_*(x,t)=2K_*(t)x_*(t)+O(|x_*(t)|^2)$ (see Lemma 3.4). Finally we can calculate the optimal feedback control with respect to the original system by $u_*(x,t)=u_*(x,p_*(x,t),t)$. In the case that $Q(t)$ is not positive definite but only positive semi definite, it does not seem to be possible to introduce a dual system with the properties sketched above.

Example.

$$\dot{x}=x^3+u, \quad x(0)=x_0, \quad x(T)=0,$$

$$\min \int_0^T (x^2+u^2)\,dt.$$

Here $A(t)=0$, $B(t)=1$, $Q(t)=1$ and $R(t)=1$. Furthermore $f(x,u,t)=x^3$ and $g(x,u,t)=0$. The linear system $\dot{x}=u$ is controllable and the condition $Q>0$ holds. Hence we can use the method described above. The equation $F_u(x,u,t)p+G_u(x,u,t)=0$ gives $u_*(x,p,t)=-\frac{1}{2}p$, so the dual system has the following form:

$$\dot{p}=-2y-3y^2p, \qquad p(0)=p_0,$$

$$\min \int_0^T (\tfrac{1}{4}p^2+y^2+2y^3p)\,dt.$$

The method of § 2 gives the result

$$y_*(p,t)=\tfrac{1}{2}p\tanh(T-t)-\tfrac{1}{8}p^3\tanh^4(T-t)+\cdots.$$

Hence

$$x_*(p,t)=\tfrac{1}{2}p\tanh(T-t)-\tfrac{1}{8}p^3\tanh^4(T-t)+\cdots$$

and it follows that

$$p_*(x,t)=2x\coth(T-t)+2x^3+\cdots.$$

Finally we find

$$u_*(x,t)=-\tfrac{1}{2}p_*(x,t)=-x\coth(T-t)-x^3+\cdots.$$

REFERENCES

[1] D. L. LUKES, *Optimal regulation of nonlinear dynamical systems*, this Journal, 7 (1969), pp. 75–100.

[2] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[3] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[4] M. ATHANS AND P. L. FALB, *Optimal Control*, McGraw-Hill, New York, 1966.
[5] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[6] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
[7] F. BRAUER AND J. A. NOHEL, *The Qualitative Theory of Ordinary Differential Equations*, W. A. Benjamin, New York, 1969.